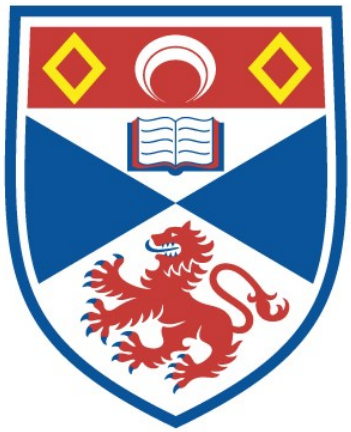
Automatic inference of latent emotion from spontaneous facial micro-expressions

Liangfei Zhang

A thesis submitted for the degree of PhD
at the
University of St Andrews



2023

Identifier to use to cite or link to this thesis:
DOI: https://doi.org/10.17630/sta/649

Candidate's declaration

I, Liangfei Zhang, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 29,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in August 2019.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 01.11.2023 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 01.11.2023 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Liangfei Zhang, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 01.11.2023 Signature of candidate

Date 01.11.2023 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Liangfei Zhang, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 01.11.2023 Signature of candidate

Abstract

Emotional states exert a profound influence on individuals' overall well-being, impacting them both physically and psychologically. Accurate recognition and comprehension of human emotions represent a crucial area of scientific exploration. Facial expressions, vocal cues, body language, and physiological responses provide valuable insights into an individual's emotional state, with facial expressions being universally recognised as dependable indicators of emotions. This thesis centres around three vital research aspects concerning the automated inference of latent emotions from spontaneous facial micro-expressions, seeking to enhance and refine our understanding of this complex domain.

Firstly, the research aims to detect and analyse activated Action Units (AUs) during the occurrence of micro-expressions. AUs correspond to facial muscle movements. Although previous studies have established links between AUs and conventional facial expressions, no such connections have been explored for micro-expressions. Therefore, this thesis develops computer vision techniques to automatically detect activated AUs in micro-expressions, bridging a gap in existing studies.

Secondly, the study explores the evolution of micro-expression recognition techniques, ranging from early handcrafted feature-based approaches to modern deep-learning methods. These approaches have significantly contributed to the field of automatic emotion recognition. However, existing methods primarily focus on capturing local spatial relationships, neglecting global relationships between different facial regions. To address this limitation, a novel third-generation architecture is proposed. This architecture can concurrently capture both short and long-range spatiotemporal relationships in micro-expression data, aiming to enhance the accuracy of automatic emotion recognition and improve our understanding of micro-expressions.

Lastly, the thesis investigates the integration of multimodal signals to enhance emotion recognition accuracy. Depth information complements conventional RGB data by providing enhanced spatial features for analysis, while the integration of physiological signals with facial micro-expressions improves emotion discrimination. By incorporating multimodal data, the objective is to enhance machines' understanding of latent emotions and improve latent emotion recognition accuracy in spontaneous micro-expression analysis.

Acknowledgements

I am profoundly grateful for the unwavering support and encouragement I received from my parents, Mr Xiaodong Zhang and Mrs Yanxia Zhang, throughout my academic journey. Their love, understanding, and constant belief in my abilities have been my driving force, and I dedicate this thesis to them.

I extend my heartfelt appreciation to my supervisor, Dr Oggie Arandelovic, for his exceptional guidance, mentorship, and insightful feedback. His expertise and dedication have been instrumental in shaping the direction of my research and refining my ideas. The symbiotic relationship between his guidance and my efforts has culminated in a research direction that I am genuinely proud of. As I reflect on this journey, I am reminded of how his mentorship has not only shaped my academic trajectory but has also imparted lasting lessons that will undoubtedly influence my future endeavours.

I am also indebted to Professor Xiaopeng Hong for his valuable collaboration, which has enriched the scope and depth of my work. His constructive insights and willingness to share his expertise have significantly contributed to the quality of my research.

To my dear friends, who have stood by me through the challenges and triumphs of my PhD journey, thank you for making my life colourful and joyful. The moments of laughter we shared provided much-needed balance amidst the rigours of research.

Lastly, I am grateful to all those who have contributed to my academic growth, directly or indirectly. Your support, whether through discussions, feedback, or encouragement, has been instrumental in producing this work.

Funding

This work was supported by the China Scholarship Council - University of St Andrews Scholarships (PhD) [201908060250].

CONTENTS

LIST OF FIGURES

List of Tables

ACRONYMS

3DFER 3D facial expression recognition

3DHOG 3D histograms of oriented gradients

4DME a spontaneous 4D micro-expression dataset with multimodalities

AI artificial intelligence

ASM active shape model

AU facial action unit

BERT bidirectional encoder representations from transformers

BN Bayesian network

CAS Chinese academy of sciences

CAS(ME)^2 Chinese academy of sciences macro-expressions and micro-expressions

CAS(ME)^3 a third generation facial spontaneous micro-expression database from Chinese academy of sciences

CASME Chinese academy of sciences micro-expressions

CASME II Chinese academy of sciences micro-expressions II

CBP centralized binary pattern

CDE composite database evaluation

CNN convolutional neural network

CNS central nervous system

ECG heart rate/fingertip pulse – electrocardiogram

EDA electrodermal activity

EEG electroencephalogram

EMFACS emotion facial action coding system

EMG electromyogram

ERT ensemble of regression tree

FACS facial action coding system

FDM facial dynamics map

FER facial expression recognition

GCN graph convolutional network

GSR galvanic skin response

HIGO histograms of image gradient orientation

HMM hidden Markov model

HOOF histograms of oriented optical flow

iGPT image generative pre-training

LBP local binary pattern

LBP-SIP local binary pattern with six intersection points

LBP-TOP local binary pattern-three orthogonal planes

LOSO leave one subject out

LP Label Powerset

LSTM long short term memory

MDMO Main Directional Mean Optical flow feature

MEGC micro-expressions grand challenge

MER micro-expression recognition

MEVIEW facial micro-expressions “in the wild”

MLP multilayer perceptron

MMER multimodal emotion recognition

MMEW micro-and-macro expression warehouse

MSM multi-head self-attention mechanism

NLP natural language processing

PNS peripheral nervous system

PPG pulse photoplethysmography

RAkELd random k-labelsets

RIFE real-time intermediate flow estimation

RNN recurrent neural network

ROI region of interest

RSP respiration

SAMM spontaneous actions and micro-movements

SDE sole database evaluation

SLSTT short and long range relation based spatio-temporal transformer

SMIC spontaneous micro-expression corpus

SVM support-vector machines

UAR unweighted average recall

UF1 unweighted F1-score

ViT vision transformer

CHAPTER ONE

INTRODUCTION

This chapter outlines the main premises of the thesis, focusing on the current standing of emotion recognition in the realm of affective computing, while also elucidating the associated merits and obstacles of automatic facial expression recognition. Additionally, the chapter presents clear definitions of *micro-expression*, which serve as the foundational concept for the subsequent research. The objective of the thesis is to contribute significantly to the field by proposing innovative approaches to enhance the analysis of spontaneous micro-expressions. This improvement aims to enhance the capabilities of automated systems in recognising latent emotions. Finally, the chapter concludes with an overview of the thesis chapters, providing readers with a structured roadmap of the content and organisation of the research.

1.1 Background

Emotional states have the potential to exert a significant influence on the physiological and psychological well-being of individuals. In contrast to mood, which refers to sustained emotional dispositions, emotions are more immediate and tied to specific stimuli or situations. Emotions involve a complex interplay between our thoughts, feelings, and physiological responses throughout the body, including changes in the brain, heart, skin, blood flow, muscles, facial expressions, and voice. These physiological changes vary depending on the specific emotion experienced. Positive emotions can enhance human health and work productivity, while negative emotions can have detrimental effects on physical and mental health. In particular, prolonged exposure to negative emotions may serve as a contributing factor to the development of depression, which can result in tragic outcomes [136]. Because of the intricate interplay between physiology and psychology in emotional responses, accurate and timely recognition of human emotions is an ongoing area of scientific inquiry across various interdisciplinary fields. Facial expressions are deemed the most reliable and universally accepted way of recognising emotions, while vocal

cues such as pitch, loudness, and speech rate, as well as body language such as gesture and posture, can also provide valuable information about a person's emotional state, intentions, and attitudes. In addition, physiological responses are objective measures of emotional arousal and can be useful in identifying emotions in situations where self-reported measures may not be reliable. These approaches provide a comprehensive and well-rounded method of recognising and understanding emotions in diverse contexts.

Within interpersonal communication, facial expressions represent a crucial element for conveying emotions. Though some disagreement on this remains, a notable number of psychologists believe that despite different cultural environments and the individuals' use of different languages, the expression of their emotions is rather universal [40]. Ekman defined six primary emotions as anger, happiness, sadness, surprise, disgust, and fear [37]. These emotions have been referred to as "basic emotions" and are accompanied by a distinct set of facial expressions that reflect the unique psychological activity associated with each emotion. Proficient recognition of facial expressions is an essential component of effective communication and can facilitate understanding of an individual's emotional state and mental well-being.

Figure 1.1: Micro-expressions (top) and macro-expressions (bottom) are qualitatively identical, but the former are involuntary and much shorter in expression than the latter.

Different from "conventional" facial expressions, which are more precisely technically termed facial *macro-expressions* that can be consciously controlled, *micro-expressions* are activated unconsciously through brief contractions of facial muscles that are inhibited by psychological factors, as depicted in Figure 1.1. Micro-expressions were initially discovered by Haggard and Isaacs in 1966 while analysing motion picture films of psychotherapy sessions for nonverbal cues between patients and therapists [58]. Ekman and Friesen subsequently incorporated micro-expression recognition into their deception studies, and popularised through the TV show *"Lie To Me"*. Although micro-expressions were reported in the 1960s, the first report published in a peer-reviewed, scientific article validating their existence was by Porter and Ten Brinke in 2008 [144]. Similarly, Matsumoto et al. published the first peer-reviewed scientific article on

individual micro-expression recognition skills in 2000 [120]. Micro-expressions offer valuable insight into an individual's emotional state, even when an attempt is made to conceal it. Therefore, by studying micro-expressions, one can gain a deeper understanding of the intricacies of human emotions. The utilisation of micro-expression holds significant potential in comprehending genuine emotional states of individuals and offering crucial cues for detecting deception. Micro-expression analysis enables experts to identify even the most subtle changes in an individual's facial expressions, potentially indicating their dishonesty. The reliability of micro-expressions also makes them a valuable tool in various emotion-related tasks, including communication negotiation [122], teaching evaluation [198], safe driving [30], health care especially mental health monitoring [54] and social security [68].

Facial expression recognition (FER) is a multi-disciplinary research field encompassing physiology, psychology, image processing, machine vision and pattern recognition. It has gained significant attention in the areas of pattern recognition and artificial intelligence in recent years, as has micro-expression recognition (MER). As Ekman and Friesen highlighted, micro-expressions are rapid facial muscle movements that last for a fraction of a second, often undetected by untrained observers, making professional training necessary for their manual analysis [39]. The human-based processes of training as well as recognition itself are time-demanding, yet the recognition accuracy is still not satisfactory for most practical purposes. Therefore, many researchers have endeavoured to develop computer vision techniques for the automatic analysis of micro-expressions.

1.2 Motivation and research gaps

Facial *actions* encompass a broad range of facial movements and extend beyond the scope of facial expressions alone. These actions can be elicited by a multitude of factors, including external stimuli that affect the face. The activation of facial muscles underlies these actions, which serve diverse functions beyond the mere display of emotions. For instance, facial muscles can be engaged in communication, exemplified by actions like winking, which convey specific messages or signals. Additionally, facial actions can also serve practical purposes, such as relieving an itch or addressing a physical discomfort. Thus, the range of facial actions extends beyond emotional expression and encompasses a variety of communicative and adaptive functions. Facial action coding system (FACS) was developed by Ekman and Friesen through observing and utilising biofeedback to map out how different facial muscle movements correspond to different expressions [41]. Based on the anatomical characteristics of the human face, FACS divided the face into several independent and interconnected motion units known as facial action unit (AU). Each AU refers to a specific muscular complex activated during a kind of facial movement.

After that, Ekman et al. further proposed the correspondence table between basic emotions and AU [42], which has become the standard for most psychological researchers to manually identify emotions from facial expressions. However, no such corresponding relations between AUs and emotions reflected from micro-expressions have been proposed by experts. Also, there is a gap in automated AU detection for micro-expressions. Therefore, I propose using computer vision methods to detect activated AUs in micro-expression, and hope that it could provide feedback to psychologists and facilitate further research in this area.

The seminal work of Pfister et al. and the release of the database of micro-expression movie clips [140], effected a marked empowerment of computer scientists in the realm of MER. The first generation of solutions built upon the well-established computer vision tradition and introduced a series of handcrafted features, 3D histograms of oriented gradients (3DHOG) [143] as the simplest extension of the 'traditional' HOG features, subsequently succeeded by more nuanced extensions such as local binary pattern-three orthogonal planes (LBP-TOP) [140], histograms of image gradient orientation (HIGO) [104] histograms of oriented optical flow (HOOF) [113], and their variations. The next generation shifted focus towards convolutional neural network (CNN) based deep learning methods [135, 85, 100, 199, 200]. Early work by and large uses convolutional kernels to extract spatial information from micro-expression video frames. This kind of pixel level operators can be considered as capturing "*short-range*", local spatial relationships. "*Long-range*", global relationships between different spatial regions have also been proposed and studied, notably by means of graph convolutional network (GCN) based architectures [115, 13, 203, 90, 97]. These methods typically use the activation of AUs as nodes to construct graphs and combine the relationships between different AU engagements with image features to enhance the discriminatory power in the context of MER. However, though these approaches consider global spatial relations so as to assist learning, they can only learn after the extraction of local features, i.e. they are unable to learn both kinds of relations jointly. Therefore, this thesis also aims to develop a third-generation architecture that can simultaneously learn both short and long-range relationships from micro-expression data and improve the accuracy of automatic emotion recognition.

Enhancing emotion recognition accuracy entails exploring avenues beyond just improving the machine learning model, considering richer data types can also help achieve better performance in automatic expression recognition. Human experience of the world is often multimodal, referring to how something happens or is experienced through multiple modalities, and a research question is characterised as multimodal when it contains multiple modalities. Incorporating multimodal signals can enable artificial intelligence (AI) to learn about the real world better. Depth information is one of the most popular complements to conventional RGB image/video

data, which could enhance the machine knowledge of spatial features. However, relying solely on readily visible physical signals, such as facial expression, speech, gesture, or posture, is not guaranteed as people can control them to hide their real emotions, especially during social communication. In contrast, physiological signals, which are in response to the central nervous system (CNS) and peripheral nervous system (PNS) of the human body, can provide reliable information about emotions according to Cannon's theory [14]. One significant advantage of using physiological signals is that they are largely involuntarily activated and, therefore, difficult to control. Researchers have attempted to establish relationships between emotional changes and various types of physiological signals. This thesis also explores the benefits of incorporating multimodal data for improving emotion recognition accuracy, specifically for FER and MER.

1.3 Hypotheses

Based on the underlying motivation and identified research gaps within the field, this section delves into the formulation of hypotheses critical to advancing our understanding of recognising latent emotions through spontaneous micro-expressions. These hypotheses serve as the foundation for guiding our investigation and shaping the trajectory of this research endeavour. Through a thorough exploration and testing of these hypotheses, we aim to pave the way for significant strides in the domain of latent emotion recognition.

Hypothesis 1: Computer vision methods can effectively detect and analyse facial muscle movements, such as AU, facilitating a better understanding of the relationship between micro-expressions and emotional states.

Hypothesis 2: Advanced architectures that can learn both short and long-range spatio-temporal relationships from micro-expression data can significantly improve the accuracy of automatic latent emotion recognition.

Hypothesis 3: Incorporating multimodal data, such as physiological signals and depth information, alongside spontaneous micro-expressions can enhance the accuracy and robustness of emotion recognition systems, providing a more comprehensive understanding of emotions.

These general hypotheses suggest that leveraging computer vision techniques, advanced deep learning architectures, and multimodal data integration can contribute to advancing the field of automatic latent emotion recognition from spontaneous micro-expression by improving the detection and analysis of facial actions, enhancing the modelling of spatio-temporal relationships, and incorporating complementary information from different modalities. I also define several detailed research hypotheses from these general ones.

Firstly, the proposal to use computer vision methods for detecting activated AU in micro-expressions addresses the gap in automated AU detection for micro-expressions. This contribution aims to provide feedback to psychologists and facilitate further research in understanding the relationship between micro-expression and latent emotional states through facial muscle activation.

Hypothesis 1.1: Segmenting facial key subregions based on activated facial muscles enhances the accuracy and effectiveness of detecting micro AU and analysing facial micro-expressions.

Hypothesis 1.2: Transferring a complex multi-label classification problem into smaller ones based on segmented regions simplifies the task of micro-expression AU detection.

Hypothesis 1.3: The use of an AU-independent cross-validation method provides a reliable and robust evaluation metric for assessing the performance of AU detection approaches in micro-expression analysis.

Secondly, the development of a third-generation architecture that can simultaneously learn both short and long-range relationships from micro-expression data represents an advancement in the field of automatic emotion recognition. By incorporating global spatial relationships through graph-based architectures, the accuracy of emotion recognition in micro-expressions can be improved.

Hypothesis 2.1: Novel approaches that utilise transformer architectures for video-based MER achieve comparable or superior performance compared to existing deep learning methods that employ other architectures.

Hypothesis 2.2: The use of alternative input representations, such as long-term optical flow matrices instead of original colour images, enhances the accuracy and robustness of MER.

Hypothesis 2.3: Integrating temporal information and spatial relations in feature extraction improves the discriminative power of the extracted features for latent emotion recognition.

Hypothesis 2.4: The inclusion of temporal aggregation mechanisms that connect spatio-temporal features extracted from multiple frames contributes to the overall effectiveness of video-based MER systems.

Lastly, the exploration and integration of multimodal data, such as incorporating physiological signals, in the context of latent emotion recognition in micro-expressions, have the potential to significantly enhance the accuracy and robustness of emotion recognition systems. This

comprehensive approach, leveraging multiple modalities, including visual cues and physiological signals, can provide a deeper and more nuanced understanding of emotions, leading to improved latent emotion recognition capabilities.

Hypothesis 3.1: Combining micro-expression and physiological signals in a multimodal learning framework improves the performance of latent emotion recognition.

Hypothesis 3.2: Separable and mixable network flow effectively extracts features from various physiological signals.

Hypothesis 3.3: The standardised normal distribution weighted feature fusion method better reconstructs informative maps from different frames of micro-expression video.

Hypothesis 3.4: The external feature guided attention module achieves multimodal learning for both micro-expression (colour and depth information) and latent emotion recognition (micro-expression and physiological signals).

1.4 Organisation of this thesis

A concise overview of each chapter in this thesis is provided below, offering a brief description and highlighting the key focus of each section.

Chapter 1: *Introduction*

This chapter provides an introduction to this thesis, elucidates the central claims and motivations of the entire work, and outlines the key contributions.

Chapter 2: *Context Survey*

This chapter presents a comprehensive and up-to-date review of micro-expression analysis and multimodal emotion recognition (MMER), including a detailed critical analysis of previous approaches with both hand-crafted and deep learning designs. Moreover, this chapter introduces the micro-expression and 3D facial expression databases, and summarises the differences and highlights of those databases.

Chapter 3: *Facial Action Unit Detection from Micro-Expression*

This chapter describes a handcrafted approach for detecting the activated AU from micro-expressions. This chapter introduces this novel task and designs an evaluation metric to experiment with the proposed approach's performance. Intensive experiments are conducted on two publicly available micro-expression databases with AU labels. The

results indicate the effectiveness of the approach and demonstrate the ability to consider AUs as the intermediate variable between micro-expressions and emotions.

Chapter 4: *Short and Long Range Relation Based Spatio-Temporal Transformer for Micro-Expression Recognition*

This chapter presents a novel deep learning transformer framework, named short and long range relation based spatio-temporal transformer (SLSTT), for video-based micro-expression recognition. The details of each block in the framework are described in the subsections and a comparison with others' work shows that the proposed entirely transformer-centred architecture outperforms the previous CNN-based ones.

Chapter 5: *Multimodal Authentic Emotion Recognition from Micro-expression and Physiological Signals*

This chapter details my work on multimodal micro-expression recognition, which involved integrating RGB frames, depth information, and physiological signals to enhance emotion recognition. I developed guided attention modules that incorporated depth information with RGB inputs, and physiological signals to achieve multimodal fusion in the framework. Results from a series of experiments showed that including physiological signals significantly improved the recognition of genuine emotions within micro-expressions.

Chapter 6: *Conclusions and Future Work*

This chapter presents a comprehensive synthesis of the principal contributions emanating from this thesis and proffers insights into potential avenues for future research in the domains of latent emotion recognition and micro-expression analysis. The contributions presented herein not only substantiate the underlying hypotheses but also pave the way for further investigations in these areas.

CHAPTER TWO

CONTEXT SURVEY

Facial expression recognition (FER) has emerged as a prominent research area in the field of computer vision, encompassing several subtasks aimed at enhancing the accuracy and authenticity of affective computing. Among these, micro-expression recognition (MER) has received considerable attention and seen rapid development over the past decade. A multitude of recognition algorithms and databases have been published, significantly contributing to the field. This chapter presents an extensive review, starting from the basic emotion models and progressing to the literature pertaining to FER and its various subtasks, with a specific focus on MER. The chapter also highlights the advantages and limitations of each algorithm and database, discusses the relative strengths and weaknesses of the various approaches, and provides an overview of the related databases.

2.1 Basic emotions and facial action units

In 1872, Darwin's book "The Expressions of the Emotions in Man and Animals" was published, which led to the recognition and study of emotions. Following Darwin's work, researchers have made significant advancements in the study of facial expressions. To recognise emotions, they must be defined. While a definition of basic emotions was proposed decades ago, there is no universally upheld consensus. Psychologists typically model emotions in two ways: by dividing them into discrete categories or by using multiple dimensions. In this section, I will present and discuss various models of emotions and the associated research.

2.1.1 Discrete emotion models

According to basic emotion theory, humans have a limited number of basic emotions (e.g., fear, anger, joy, sadness) that are biologically and psychologically ingrained [72]. These emotions

have a set of associated behaviours and evolved to handle fundamental life tasks, such as survival. Basic emotions can be combined to form complex emotions. These emotions have innate neural substrates and universal behavioural phenotypes. The differences between some of these emotions may have developed later for social functions rather than survival. Ultimately, emotions provide a way for humans to explain and understand their experiences and behaviours.

The exact number of basic emotions is a subject of debate among researchers, and various proposals have been put forward. In a special issue of Emotion Review [168], several research psychologists outlined the latest thinking about each theoretical model of basic emotions. For instance, Ekman and Cordaro initially suggested seven basic emotions: fear, anger, joy, sadness, contempt, disgust, and surprise[38]. The different lists of basic emotions are largely similar, with some exceptions and disagreements over terminology (refer to Table 2.1). To prevent any confusion, items have been grouped across the lists that appear to represent the same emotional state, despite potentially having different names. All four lists consist of one positive emotion (labelled happiness, enjoyment, or play) and three negative emotions: sadness (labelled grief by Panksepp and Watt [132]), fear, and anger; interest/seeking is included in Izard's, Levenson's and Panksepp and Watt's lists [74, 98, 132], but is not recognised as a basic emotion by Ekman and Cordaro, who consider it a "cognitive state of focused attention." Similarly, Panksepp and Watt's model is the only one that excludes disgust, as they believe it evolved to regulate physiological needs like hunger or physical pain [132].

Table 2.1: Similarities and discrepancies among the clear-cut basic emotions included in each of the four models.

IZARD [74]	**PANKSEPP & WATT** [132]	**LEVENSON** [98]	**EKMAN & CORDARO** [38]
Happiness	Play	Enjoyment	Happiness
Sadness	Panic/Grief	Sadness	Sadness
Fear	Fear	Fear	Fear
Anger	Rage	Anger	Anger
Disgust		Disgust	Disgust
Interest	Seeking	Interest	
Contempt			Contempt
	Lust	Love	
	Care	Relief	Surprise

Some researchers argue that four basic emotions – fear, anger, joy, and sadness – are adequate to account for human emotional experiences [76, 51, 177, 230]. As Izard argued, people require the category label of fear to explain flight for safety, anger to explain frustration when blocked from achieving a goal, joy (or an equivalent) to express pride in accomplishments, and sadness to express experience of a significant loss [73]. Some other models include more complex emotions than basic ones. Plutchik suggested eight primary emotions: anger, fear, sadness,

disgust, surprise, anticipation, trust, and joy, which he arranged in a colour wheel as shown in Figure 2.1. The stronger emotions are in the centre, and the weaker ones are at the outer edge. These basic emotions can be mixed to form complex emotions, just like colours [142]. Parrott's approach is another notable theory, where he identified over 100 emotions and organised them in a three-layer tree-structured list, see Table 2.2. The first layer consists of six primary emotions that branch out into different forms of feeling, and other layers refine the granularity of the previous layer, making abstract emotions more concrete [134]. Though more emotions are included in these models, discrete emotion models only use word descriptions for emotions, it is still challenging to analyse complex emotions, such as mixed emotions, which may be difficult to express precisely in words and require quantitative analysis.

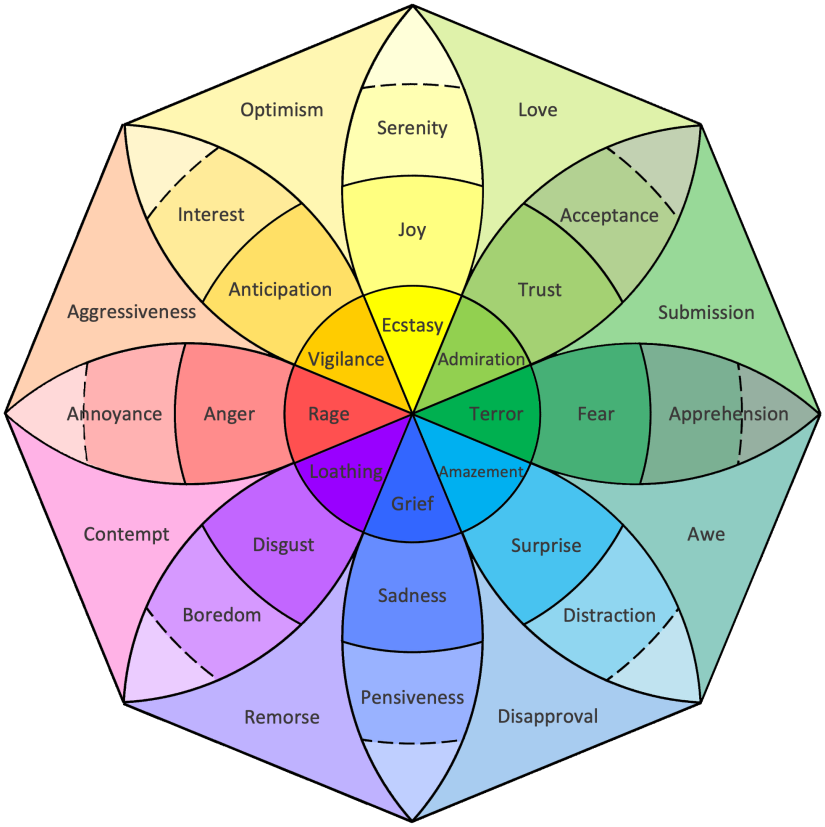


Figure 2.1: Comprehensive visual representation: Plutchik's emotion wheel mapping the spectrum of human emotional states.

2.1.2 Multi-dimensional emotion space models

Although using discrete labels such as 'fear' and 'joy' is the most straightforward way to represent an emotion, there are some disadvantages to label-based representations. One of the main issues is that labels are not cross-lingual, as emotions do not have exact translations in different languages;

Table 2.2: Parrot's emotion framework.

Primary emotion	Secondary emotion	Tertiary emotion
Love	Affection	Adoration, Fondness, Liking, Attractiveness, Caring, Tenderness, Compassion, Sentimentality
	Lust/Sexual desire	Desire, Passion, Infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, Bliss, Gaiety, Glee, Jolliness, Joviality, Joy, Delight, Enjoyment, Gladness, Happiness, Jubilation, Elation, Satisfaction, Ecstasy, Euphoria
	Zest	Enthusiasm, Zeal, Excitement, Thrill, Exhilaration
	Contentment	Pleasure
	Pride	Triumph
	Optimism	Eagerness, Hope
	Enthrallment	Enthrallment, Rapture
	Relief	Relief
Surprise	Surprise	Amazement, Astonishment
Anger	Irritability	Aggravation, Agitation, Annoyance, Grouchy, Grumpy, Crosspatch
	Exasperation	Frustration
	Rage	Anger, Outrage, Fury, Wrath, Hostility, Ferocity, Bitter, Hatred, Scorn, Spite, Vengefulness, Dislike Resentment
	Disgust	Revulsion, Contempt, Loathing
	Envy	Jealousy
	Torment	Torment
Sadness	Suffering	Agony, Anguish, Hurt
	Sadness	Depression, Despair, Gloom, Glumness, Unhappy, Grief, Sorrow, Woe, Misery, Melancholy'
	Disappointment	Dismay, Displeasure
	Shame	Guilt, Regret, Remorse
	Neglect	Alienation, Defeatism, Dejection, Embarrassment, Homesickness, Humiliation, Insecurity, Insult, Isolation, Loneliness, Rejection
	Sympathy	Pity Sympathy
Fear	Horror	Alarm, Shock, Fear, Fright, Horror, Terror, Panic, Hysteria, Mortification
	Nervousness	Anxiety, Suspense, Uneasiness, Apprehension (Fear), Worry, Distress, Dread

for example, "disgust" does not have an exact translation in Polish [148]. Additionally, emotions with the same label may have different intensities, such as describing happiness as being a little bit happy or very happy. To address these limitations, psychologists often represent emotions or

feelings in an *n*-dimensional space, typically two or three-dimensional. The most well-known space, originating from cognitive theory, is the 2D valence-arousal or pleasure-arousal space [92]. The valence dimension indicates the positivity or negativity of emotion, ranging from unpleasant feelings to pleasant feelings, such as a sense of happiness. The arousal dimension indicates the level of excitement that the emotion represents, ranging from sleepiness or boredom to wild excitement. While the 2D emotion space is useful in distinguishing between positive and negative emotions, it may not be sufficient in distinguishing between similar emotions. For instance, both fear and anger fall under the category of negative valence and high arousal. To address this issue, Mehrabian expanded the emotion model from 2D to 3D, as shown in Figure 2.2. The additional axis in the 3D model is called dominance and ranges from submissive to dominant, representing the degree of control that a person has over a particular emotion [121]. With the inclusion of this dimension, it becomes easier to differentiate between anger and fear, as anger is located in the dominant axis while fear is in the submissive axis.

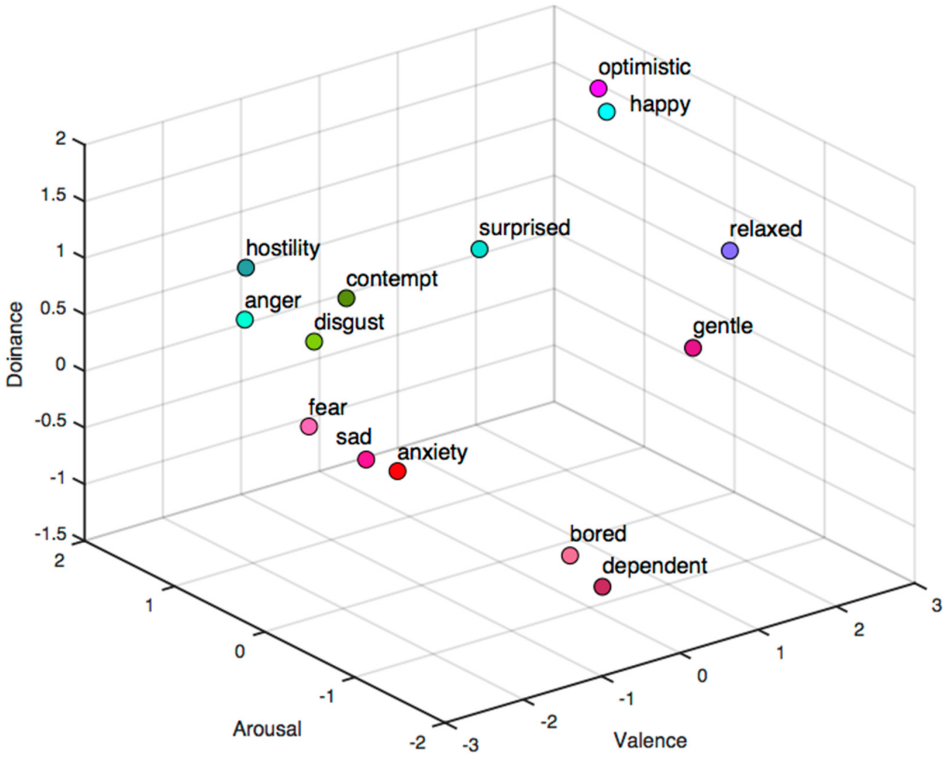


Figure 2.2: VAD (Valence-Arousal-Dominance) emotional state model.

2.1.3 Facial action coding system

Researchers have yet to reach a consensus on the definition of emotions, but they generally agree that emotions can significantly affect facial expressions [75]. Recent research has identified

Table 2.3: Main codes in facial action coding system (FACS).

AU	Description	Muscular basis
1	Inner brow raiser	Frontalis, pars medialis
2	Outer brow raiser	Frontalis, pars medialis
4	Brow lowerer	Corrugator supercilii, depressor supercilii
5	Upper lid raiser	Levator palpebrae superioris
6	Cheek raiser	Orbicularis oculi, pars orbitalis
7	Lid tightener	Orbicularis oculi, pars palpebralis
9	Nose wrinkler	Levator labii superioris alaquae nasi
10	Upper lip raiser	Levator labii superioris
11	Nasolabial deepener	Zygomaticus minor
12	Lip corner puller	Zygomaticus major
13	Cheek Puffer	Levator anguli oris (a.k.a. Caninus)
14	Dimpler	Buccinator
15	Lip corner depressor	Depressor anguli oris (a.k.a. Triangularis)
16	Lower lip depressor	depressor labii inferioris
17	Chin raiser	Mentalis
18	Lip pucker	Incisivii labii superioris and Incisivii labii inferioris
20	Lip stretcher	Risorius with platysma
22	Lip funneler	Orbicularis oris
23	Lip tightener	Orbicularis oris
24	Lip pressor	Orbicularis oris
25	Lips part	Depressor labii inferioris, or relaxation of Mentalis, or Orbicularis oris
26	Jaw drop	Masseter, relaxed Temporalis and internal Pterygoid
27	Mouth stretch	Pterygoids, Digastric
28	Lip suck	Orbicularis oris
41	Lid droop	Relaxation of Levator palpebrae superioris
42	Slit	Orbicularis oculi
43	Eyes closed	Relaxation of Levator Palpebrae superioris; Orbicularis oculi, pars palpebralis
44	Squint	Orbicularis oculi, pars palpebralis
45	Blink	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis
46	Wink	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis

two types of facially expressed emotions, approach or withdrawal, based on their relationship to cognitive processes. Tomkins proposed that the subjective experience of emotions results from feedback from facial muscle changes. Studies have explored how an individual's subjective experience of emotions influences the performance of their muscular movements [166]. To enable facial movement research in various fields, Ekman and Friesen developed the FACS,

which offers a standardised language for measuring and describing facial behaviour [41], main codes are shown in Table 2.3. FACS uses facial action units (AUs) to represent fundamental actions of individual or group muscles. Although FACS itself does not include emotion-specific descriptors, it is commonly used to interpret nonverbal communicative signals, such as facial expressions related to emotion or other human states.

To clarify, FACS is an index of facial movements that does not provide any biomechanical information about the degree of muscle activation. The AUs are identified by a number, shorthand name, and anatomical basis and are rated on a 5-point intensity scale. The FACS modifiers include letters A-E appended to the AU number to indicate intensity levels from minimal to maximum. Other modifiers used in FACS codes for emotional expressions include "R" for actions that occur on the right side of the face and "L" for actions on the left side. An action that is unilateral but has no specific side is indicated with a "U", while an action that is bilateral but has a stronger side is indicated with an "A" for asymmetric [41]. To make emotion-based inferences from single or combinations of AUs, researchers often use related resources, such as emotion facial action coding system (EMFACS), the FACS Investigators' Guide, and the FACS interpretive database [42]. Table 2.4 provides an example of such emotion-based inferences using AUs.

Table 2.4: The examples of emotion-related facial actions.

Emotion	Action units
Happiness	$6+12$
Sadness	$1+4+15$
Surprise	$1+2+5B+26$
Fear	$1+2+4+5+7+20+26$
Anger	$4+5+7+23$
Disgust	$9+15+17$
Contempt	$R12A+R14A$

2.2 Micro-expression recognition approaches

The automatic recognition of human facial expressions, known as FER, has been a topic of interest among researchers seeking to gain a deeper understanding of human emotions. FER systems can be classified into two main categories: static image-based and dynamic sequence-based methods. Static image-based methods encode spatial information from a single image, while dynamic sequence-based methods consider the temporal relationships between frames in an input facial expression sequence. Traditionally, handcrafted features or shallow learning techniques such as local binary pattern (LBP), non-negative matrix factorisation, and sparse

learning have been used for FER []. However, with the advent of facial expression recognition competitions such as FER2013 [49] and EmotiW [33], as well as advancements in processing power and network architectures, research has shifted towards deep learning methods, which have achieved superior recognition accuracy. In recent years, various research institutions have also provided a large number of datasets, and more and more deep-learning methods have been proposed to address challenging scenarios in this area. One significant challenge is the recognition of micro-expressions, which refers to the recognition of emotions expressed in a sequence of faces known to be brief and subtle. In recent years, computer vision technology has been increasingly utilised for automatic MER, which has improved the feasibility of applications involving micro-expressions. In this section, I provide a broad overview of the various methods employed in the realm of MER, ranging from classical manually engineered features to newly emerging deep learning-based approaches.

2.2.1 The first generation: hand-crafted features

2.2.1.1 3D histograms of oriented gradients

Polikovsky et al. proposed the use of a 3D gradient feature to describe local spatio-temporal dynamics of the face [143]. Following the segmentation of a face into 12 regions according to the FACS [40], each region corresponding to an independent facial muscle complex, and the appearance normalisation of individual regions, Polikovsky et al. obtained 12 separate spatio-temporal blocks. The magnitudes of gradient projections along each of the three canonical directions are then used to construct histograms across different regions, which are used as features. The authors assumed that each frame of the micro-expression image sequence involves only one AU, which represents one specific activated facial muscle complex in FACS, and this unit can be used as an annotation of the image. The k-means algorithm is used for clustering in the gradient histogram feature space in all frames of micro-expression image sequences, and the number of clusters is set to the number of AUs that have appeared in all micro-expression samples. The AU corresponding to the greatest number of features is regarded as the real label of each cluster.

The feature extraction method of this work is relatively simple and is an extension of the plane gradient histogram. The model construction adopts a more complicated process, which can be regarded as k-mean cluster of the vectors for different facial cubes. It is robust to the correctness of the labels and insensitive to a small number of false annotations.

The main limitation of this work lies in the aforementioned assumption that only a single AU is active in each frame, which is overly restrictive in practice.

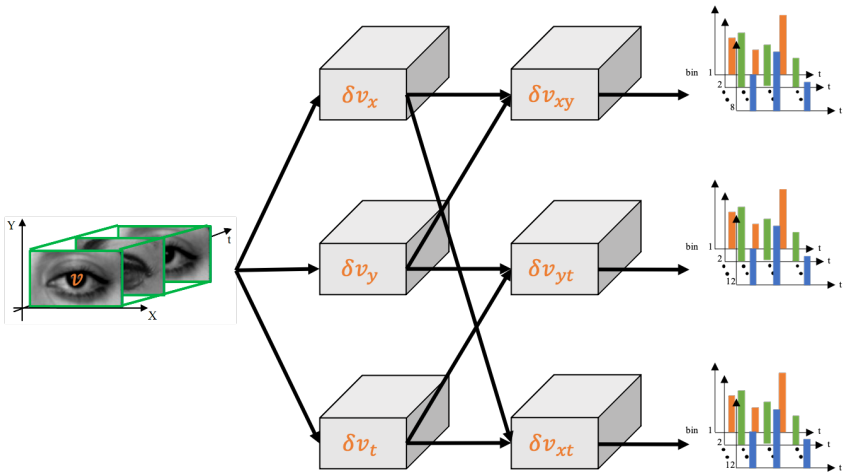


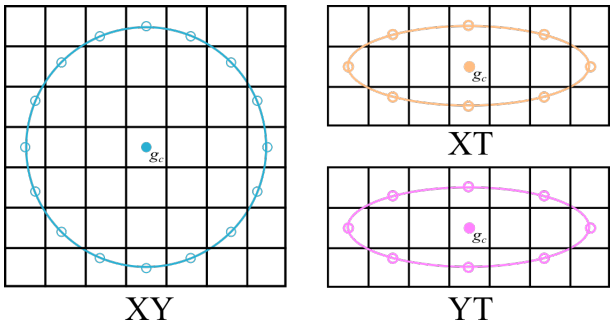
Figure 2.3: Conceptual summary of the descriptor extraction process for a facial cube using 3D histograms of oriented gradients (3DHOG).

2.2.1.2 Local binary pattern-three orthogonal planes

A LBP is a descriptor originally proposed to describe the local appearance of an image. The key idea behind it is that the *relative* brightness of neighbouring pixels can be used to describe local appearance in a geometrically and photometrically robust manner [127, 43, 82]. The basic LBP feature extractor relies on two free parameters, call them R and P. Uniformly sampling P points on the circumference of a circle with the radius R centred at a pixel, and taking their brightness relative to the centre pixel (brighter than, or not – one bit of information) allows the neighbourhood to be characterised by a P-bit number.

In recognition of micro-expressions, in order to encode the spatio-temporal co-occurrence pattern, local binary pattern-three orthogonal planes (LBP-TOP) [217] is used to extract the LBP features separately for the XY, XT, and YT planes in image sequences. Neighbourhood sampling is now performed over a circle in the purely spatial plane, and over ellipses in the spatio-temporal planes.

Pfister et al. made one of the earliest attempts to recognise micro-expressions automatically. Their method, in which LBP-TOP is used for the feature extraction, has been highly influential in the field and much follow-up work drew inspiration from it [140]. Pfister et al. first use a 68-point active shape model (ASM) [27] to locate the key points of the face. Based on the key points obtained, the deformation relationship between the first facial frame of each sequence and the model facial frame is calculated using the local weighted mean [50]. A geometric transformation is then applied to each frame of the sequence so as to normalise small pose variations and coarse expression changes. In order to account for differences in the number of frames between different



$$LBP(M,R) = \sum_{m=0}^{M-1} s(g_m - g_c)2^m$$

Figure 2.4: Detailed sampling for LBP-TOP with $RX = RY = R = 3$, $RT = R = 1$, $PXY = M = 16$, and $PXT = PYT = M = 8$.

input sequences, temporal interpolation model is used to temporally interpolate between frames, thus normalising sequence length to a specific count. LBP-TOP features are extracted from these normalised sequences. Finally, support-vector machines (SVM), random forest, and multiple kernel learning methods are used for classification. Wang et al. expressed the micro-expression sequence and its LBP features by a tensor and performed sparse tensor canonical correlation analysis on the tensor to learn the relationship between the micro-expression sequence and its LBP features [187]. The simple nearest neighbour algorithm is used for classification. In experiments, the authors demonstrate the superiority of their approach over the original LBP-TOP method.

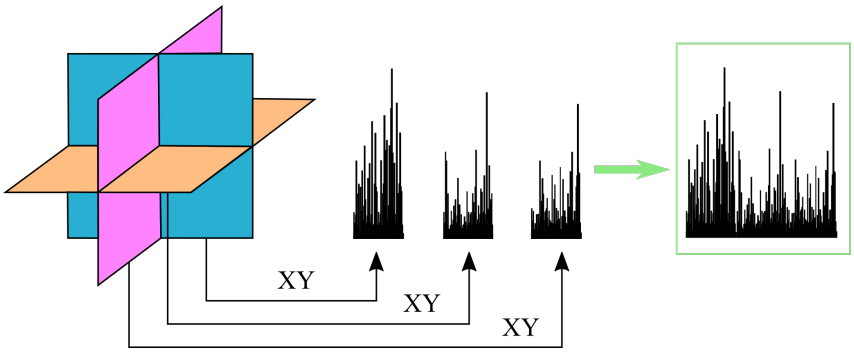
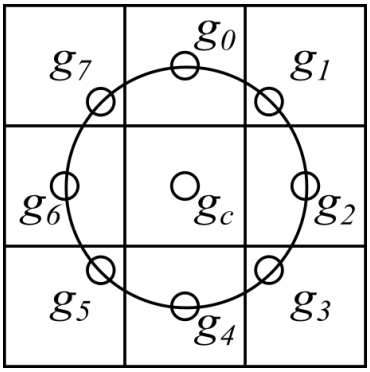


Figure 2.5: Conceptual illustration of the Histogram of Concatenated LBP-TOP feature.

Local binary pattern with six intersection points (LBP-SIP) [192] extends LBP features for MER in a different manner. The main improvement of the work of Wang et al. is to reduce the feature dimension to improve feature extraction. Compared with LBP-TOP, it reduces information redundancy, thus providing a more compact representation. Experimental evidence suggests

that its extraction is nearly three times faster than that of LBP-TOP. Specifically, in the same experimental environment, the average LBP-TOP extraction time of the Chinese academy of sciences micro-expressions II (CASME II) database is 18.289s, and the LBP-SIP extraction time is 15.888s. Furthermore, in the context of the use of the descriptors for recognition, the LBP-TOP based MER takes 0.584s per sequence, in contrast to only 0.208s for LBP-SIP based.

Centralized binary pattern (CBP) [202] descriptor is another variation on the conceptual theme set out by LBP. In broad terms, it is computed in a similar way to LBP. However, unlike in the case of LBP, CBP compares the central pixel of an area with a pair of neighbours, see Figure 2.6. Therefore, the corresponding binary code length is about half of that of LBP, with a lower dimensionality of the corresponding histogram. Indeed, the key advantage of CBP compared to LBP is that it produces lower dimensionality features. Hence, Guo et al. employ the CBP-TOP operator in place of LBP-TOP, with an extreme learning machine for classification, and experimentally demonstrate that performance improvement is indeed effected by their approach [56].



$$CBP(M,R) = \sum_{m=0}^{(M/2)-1} s\left(g_m - g_{m+(M/2)}\right) 2^m + s\left(g_c - \frac{1}{M+1}\left(\sum_{m=0}^{M-1} g_m + g_c\right)\right) 2^{M/2}$$

Figure 2.6: Detailed sampling for CBP operator with $RX = RY = R = 1$, $PXY = M = 8$.

In addition to standard texture features, some researchers have also considered the use of colour on micro-movement extraction (colour has indeed been shown to be important in face analysis more generally [1]). If the usual RGB space that the original face image data is represented in, is adopted for the extraction of the aforementioned local appearance features (such as the commonly used LBP-TOP), the three channels result in redundant information, failing to effect improvement over greyscale. Hence, Wang et al. considered this problem and instead proposed the use of tensor independent colour space [184]. In another work [186], the researchers tried to use CIELab and CIELuv colour spaces, which have already demonstrated success in applications

needing human skin detection [213]. Their experiments showed that the transformation of colour space can effect an improvement in recognition.

2.2.1.3 Histograms of oriented optical flow

One of the influential works which does not follow the common theme of using LBP-like local features is that of Liu et al. which uses a different local measure, namely optical flow. The authors extract the main motion direction in the video sequence and calculate the average optical flow characteristics in the partial facial blocks [113]. Hence, they introduce Main Directional Mean Optical flow feature (MDMO). Firstly, the face key point of each frame is located by using the discriminative response map fitting model [3]. Then the optical flow field of each frame relative to the succeeding frame is used to find an affine transformation matrix which corrects for pose change. The transformation matrix makes the difference of facial landmarks in each frame from the first frame minimal. The authors then calculate the average of the most similar motion vectors of the optical flow field in each region as the motion characteristic of the region. Specifically, they calculate the histograms of oriented optical flow (HOOF) feature [18] in each region and quantise all optical flow direction vectors to eight intervals to obtain a histogram of the aforementioned directions. The resulting histogram features are finally fed into a support vector machine, trained to classify micro-expressions.

Following in spirit but unlike the work of Liu et al., Xu et al. used the optical flow field as the key low-level feature to describe the pattern of micro-expression movement using the facial dynamics map (FDM) [204]; see Figure 2.7. The FDM better reflects intricate local motion patterns characteristic of micro-expressions, and has the appeal of being beneficial in interpretability by virtue of its useful visualisation. Nevertheless, the uniform and indeed major disadvantage of HOOF methods lies in their high computational cost, which makes them unsuitable for real-time, large-scale MER.

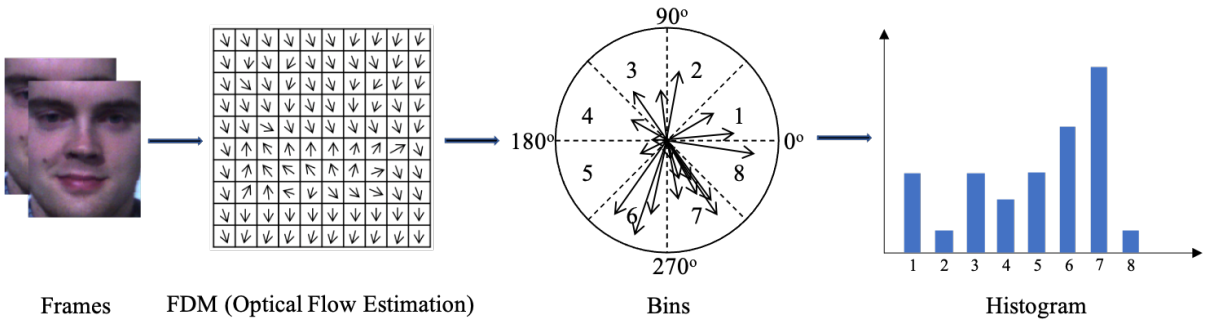


Figure 2.7: One-dimensional histogram of FDM from optical flow estimation.

2.2.2 The second generation: convolutional neural networks

Deep learning in the realm of micro-expression analysis started around 2016, and the annual number of publications shows an exponential increase in the following years. A vast number of deep learning frameworks have been proposed, with their own strategy of choosing and detecting facial region of interest (ROI), input data, network structure or temporal recurrent strategy. I overview these next.

Kim et al. use deep learning and introduce a feature representation based on expression states – convolutional neural network (CNN) are employed for encoding different expression states (start, start to apex, apex, apex to end, and end). Several objective functions are optimised during spatial learning to improve expression class separability [87]. The encoded features are then processed by a long short term memory (LSTM) network to learn features related to time scales. Interestingly, their approach failed to demonstrate an improvement over more old-fashioned, hand-crafted feature based methods, merely performing on par with them. While these results need to be taken with a degree of caution due to the limited scale of empirical testing and low data diversity (a single dataset, CASME II, discussed shortly, was used), they suggest that the opportunity for innovation in the sphere of deep learning in the context of micro-expressions is wide open.

Peng et al. also adopt a deep learning paradigm, while making use of ideas previously shown to be successful in the realm of conventional methods, by using a sequence of optical flow data as input [137]. To overcome the limitation imposed by the availability of training data, their dual time scale convolutional neural network comprises a shallow neural network for MER and only four layers for the convolutional and pooling stages. On a dataset formed by merging Chinese academy of sciences micro-expressions (CASME) and CASME II, using four different micro-expression classes – namely negative, positive, surprise, and other – their approach achieved higher accuracy than the competing methods: STCLQP [65], MDMO [113] and FDM [204].

Khor et al. proposed an enhanced long-term recursive convolutional network for MER, which uses the architecture [34] to characterise small facial changes [85]. The model includes a deep spatial feature extractor and a time extractor. These two variants of the network are enriching the spatial dimension by input channel superposition and the time dimension by depth feature superposition. Experimental evidence suggests that spatial and time modules play different roles within this framework, and that they are highly interdependent in effecting accurate performance. The experiments were performed with the usual evaluation metric, also with the appealing modification that training and test were performed on datasets with different provenances, namely, while training was done on CASME II, testing was performed on spontaneous actions

and micro-movements (SAMM), vice versa.

Xia et al. proposed a spatio-temporal recurrent convolutional network that captures spatio-temporal changes in micro-expression sequences. The approach employs a CNN with recurrent connections to automatically learn the visual features of micro-expressions and uses an end-to-end optimisation strategy. Additionally, a temporal data augmentation strategy and a balanced loss function were used to overcome the issues of limited and unbalanced training samples [200]. Gupta introduced a new method based on temporal and spatial characteristics. The approach aims to address the problems of incomplete feature encoding and insufficient training data by using AUs, landmarks, gaze, and appearance features of all video frames to encode subtle expression changes while preserving mostly relevant micro-expression information [57]. These proposed methods have shown potential in improving the accuracy of micro-expression recognition.

2.2.3 Closing remarks

To summarise this section, in the realm of conventional computer vision approaches to micro-expression recognition and analysis, there is a broad similarity between different approaches described in the literature, all of them being based on appearance based local (in time or space) features. In general, simple spatial LBP-TOP features (and similar variations) perform better than spatio-temporal 3DHOG and HOOF, when high-resolution images are used. However, when image resolution is low, the reverse is observed. This observation is consistent with what one might expect from theory. Namely, the performance of LBP-TOP features is adversely affected by the reduction in resolution due to their reliance on local spatial information. The loss of spatial details hampers their effectiveness. On the other hand, HOOF and 3DHOG heavily rely on temporal variations, making them less susceptible to changes in image resolution. While they are not entirely unaffected by such changes, the inter-frame information they capture remains relatively more robust.

Contrasting conventional computer vision approaches are emergent deep learning methods. Though a number of different micro-expression recognition algorithms based on deep learning have now been described in the literature, the performance of this umbrella of methods is yet to demonstrate its value in this field. Finally, for completeness, I include a detailed summary of a comprehensive list of different conventional and deep learning approaches in Table A.1, including many minor variations on the themes directly surveyed in this section and which do not offer sufficient novelty to warrant being discussed in detail.

2.3 Multimodal emotion recognition

Another challenging issue that can better uncover people's true emotions is multimodal emotion recognition (MMER). MMER is a field of research that combines information from multiple sources to recognise and classify human emotions. The need for MMER arises from the fact that human emotions are often conveyed through multiple modalities, and a single modality may not be sufficient to accurately recognise them. Recent advancements in technology and data collection have enabled researchers to explore new approaches to recognising emotions, leading to the integration of FER with other modalities such as voice, text, depth data, and physiological signals. Depth data provides additional information about facial expressions, while physiological signals, such as heart rate and skin conductance, can provide insight into a person's emotional state. This integration has led to improved performance in emotion recognition tasks, making MMER an increasingly important area of research. In this section, I provide an overview of the different modalities used in emotion recognition related to FER, their respective advantages and limitations, and the challenges and opportunities associated with their integration.

2.3.1 Facial expression recognition with depth information

In the field of FER, 2D texture data have been the most widely used data representation. However, with the emergence of new technologies, alternative methods such as 3D models of the face or combining texture and depth information as multimodal data have become more prevalent. Consequently, a new subtask of FER called 3D facial expression recognition (3DFER) has emerged. This part introduces the conventional feature and facial model-based methods for 3DFER, as well as the leading deep learning approaches.

2.3.1.1 Conventional feature and facial modal based methods

Conventional approaches for static 3DFER can be divided into two categories: feature-based and facial model-based methods. Feature-based methods extract facial surface geometric information, such as curvature, distances between landmarks, and local shape, from the input data. These features are then fed into various classifiers for emotion recognition, such as SVM [157, 4, 79], hidden Markov model (HMM) [94, 149, 150], random forest [36] or neural networks [70, 80]. However, feature-based methods require correctly located landmarks for feature extraction, which was a difficult task until recently when it became automated. Additionally, the performance of feature-based methods relies on the discriminative power of the facial features adopted. An example of the framework is shown in Figure 2.8.

Facial model-based methods for 3DFER typically use a generic face model created using

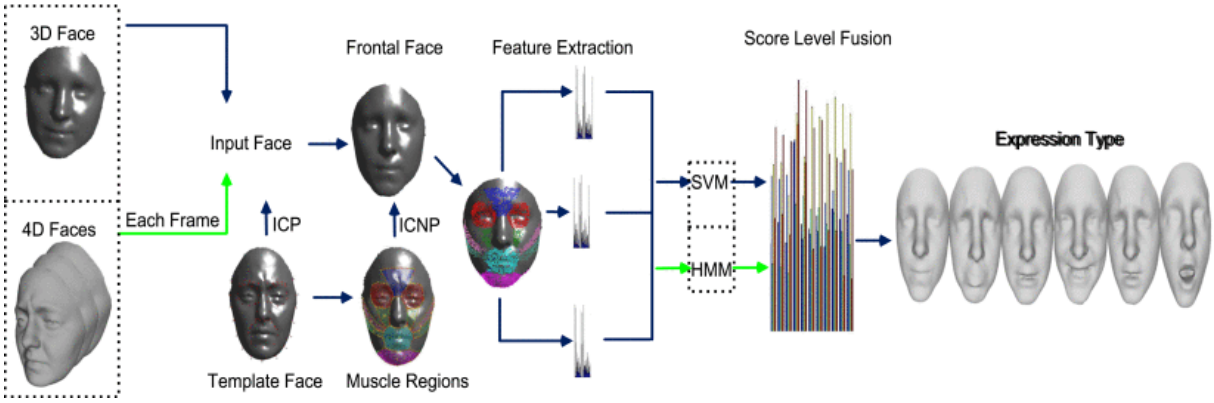


Figure 2.8: An example framework of conventional approach. For the input face scan (3D/4D frame) with some kind of expression, the entire face is segmented into 11 pre-defined areas, and different features (i.e. coordinate, normal, and shape index) extracted from each muscle region are fed into SVM (for 3D data) or HMM (for 4D data) for prediction, where the weights of different facial regions for score level fusion are learned offline on the training data [228].

the neutral expression to determine emotions by measuring the feature vector formed by the coefficient of shape deformation. This approach needs to bring into correspondence the tracking model to 3D face scans by means of a registration step. Some recent examples of model-based approaches include Zhao et al.'s, which combined a Bayesian belief net and statistical facial feature models and achieved an average recognition rate of over 82% on the BU-3DFE database [225]. Zhen et al. presented a novel approach to 3DFER problem based on the muscular movement model, which combines the advantages of both feature-based and model-based methods. The model forms 11 muscle regions, each of which is described by a certain number of geometric features to capture shape characteristics and uses a genetic algorithm to learn the weights of the sections. SVM and HMM classifiers are used for expression prediction in static and dynamic 3DFER datasets, BU-3DFE and BU-4DFE respectively [228].

Aside from Zhen et al.'s work, early research in 3D motion-based facial expression analysis includes Yin et al.'s, which recognised the six basic expressions using motion vectors for classification without explicitly modelling temporal dynamics [211]. A deformable model was utilised for tracking the changes between frames and from which the motion vectors could be found and the BU-4DFE database was used for experiments. Another approach used ASM to represent pairs of 2D and 3D images in order to track the movements of landmarks and identify deformations corresponding to specific AUs [169, 170]. Also, a small 3D database was created to analyse expression dynamics, where feature points were tracked to capture 3D mesh deformation during expression [17]. Dimensionality reduction embedded video sequences into a low-dimensional manifold, used to build a probabilistic model containing temporal information. In Sun and Yin's work, the deformable model [211] was adapted to track changes in each frame

to extract geometric features [161]. Linear discriminant analysis was applied for dimensionality reduction, and 2-dimensional HMMs was used to model spatial and temporal relationships between features in analysing facial expression dynamics using the BU-4DFE database.

2.3.1.2 Deep learning approaches

Deep learning techniques, first theorised in the 1980s, have been successfully applied in various fields. Deep neural networks have also been used to classify facial images into emotion categories, achieving higher accuracy than traditional methods. Automatic deep FER involves three steps: pre-processing, deep feature learning, and classification. Pre-processing includes image cropping, rotation correction, data augmentation, and spatial normalisation. Then, a deep learning technique like CNN or recurrent neural network (RNN) is applied to perform feature extraction and classification in an end-to-end manner. Alternatively, the neural network can be used only for feature extraction, and then independent classifiers, such as SVM, can be applied to the extracted representations.

Researchers have also used deep learning-based methods for 3DFER, such as the approach of Li et al., who proposed a new deep CNN model for subject-independent multimodal 2D+3D FER. This is the first work of introducing deep CNN to 3DFER and deep learning-based feature level fusion for multimodal 2D+3D FER [99], see Figure 2.9 for the framework. Other researchers, such as Oyedotun et al., proposed a CNN model that learns discriminative features from both RGB and depth map latent representation [128], while Yang and Yin use CNNs and landmark clues, with the sole use of 3D geometrical facial models [209].

Apart from the texture based CNN, Chen et al. directly used 3D facial point clouds based on a fast and light manifold CNN model. The model adopts a human vision inspired pooling structure and a multi-scale encoding strategy to enhance geometry representation, which highlights shape characteristics of expressions and runs efficiently, achieving state-of-art performance on BU-3DFE [23]. Jan et al. designed a novel system for 3DFER based on accurate facial parts extraction and deep feature fusion, achieving better performance than using the entire face [80]. Zhu et al. introduced a discriminative attention-based CNN, to capture more comprehensive expression-related representations [235]. Recently, some researchers have started using 4D data [102, 6, 7]. For example, Li et al. proposed a dynamic geometrical image network. Geometrical images were generated by estimating the differential quantities from the given 3D facial meshes. A score-level fusion was then performed on the probability scores of different geometrical images for emotion recognition.

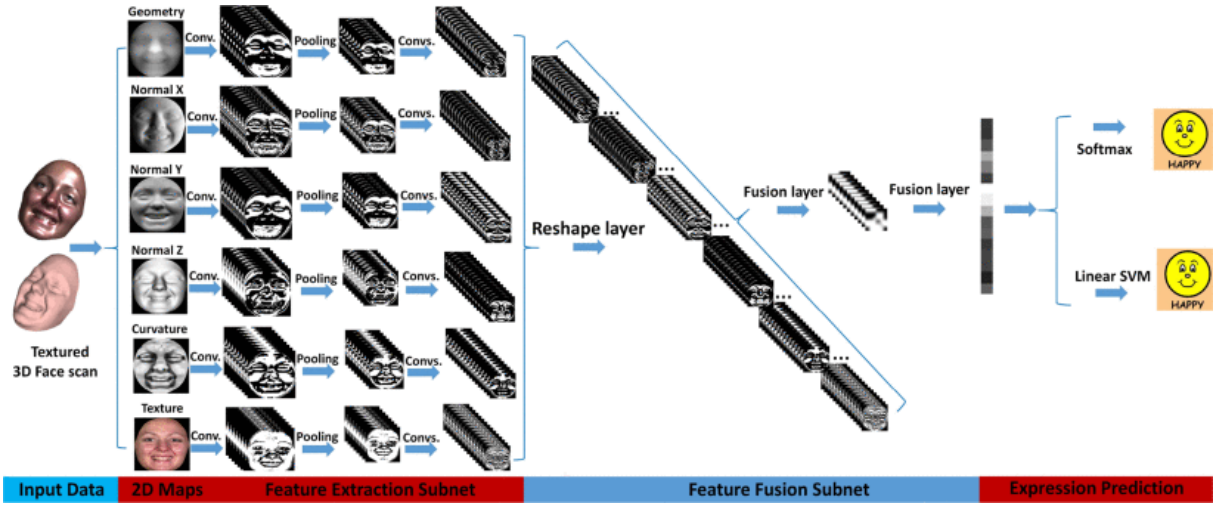
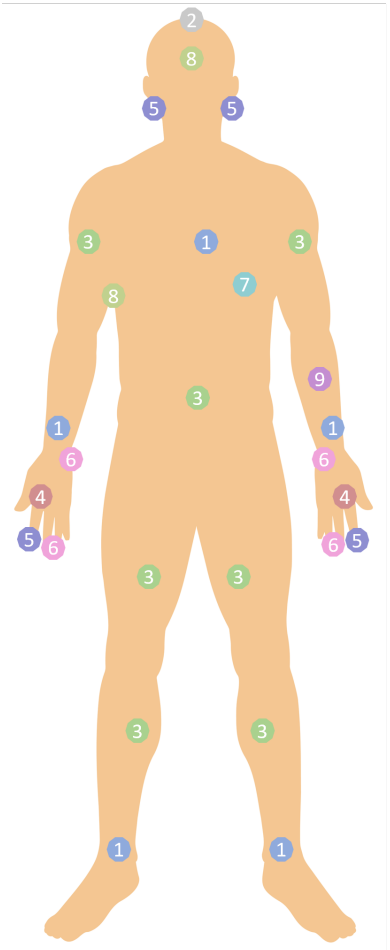


Figure 2.9: An example framework of 2D+3D FER. Each textured 3D face scan is represented as six types of 2D facial geometric and photometric attribute maps (i.e. 3D coordinates based geometry map, normal vectors based normal maps, principle curvatures based curvature map, and texture map). These attribute maps are jointly fed into the feature extraction subnet of DF-CNN with sharing parameters, generating hundreds of multi-channel feature maps. All these feature maps are then fed into the feature fusion subnet. Finally, the softmax-loss layer is followed for network training [99].

2.3.2 Multimodal emotion recognition with physiological signals

Emotions are complex experiences that involve not only outward physical expressions but also internal feelings, thoughts, and other processes that may not be consciously perceived by the individual. For example, people might smile in a formal social occasion even if he is in a negative emotional state. The other category is using the internal signals—the physiological signals, which include the electroencephalogram (EEG), temperature, heart rate/fingertip pulse – electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP), etc, see Figure 2.10. Some of these physiological processes can be naturally recognised by others, such as sensing someone's clammy hands or a racing heart. The relationship between bodily sensations and external expressions is a topic of ongoing research and historical controversy. While some early theorists, such as James, emphasised the role of bodily changes in emotional experiences [78], others, like Cannon and Schachter, argued that physiological responses alone were not sufficient to discriminate emotions [14, 151]. Recent studies have explored the possibility of using pattern recognition techniques to classify emotions based on physiological signals, such as facial electromyogram signals or autonomic nervous system responses, with varying degrees of accuracy.

ECG, is a non-invasive method of measuring the electrical activity of the heart. This activity is displayed as a waveform on a computer screen or chart recorder, which can help identify the



1 Electrocardiogram (ECG): Electrical activity of the heart measured using electrodes placed on the chest or wrists.

2 Electroencephalogram (EEG): Electrical activity of the brain measured using electrodes attached to the scalp.

3 Electromyogram (EMG): Electrical activity of muscles measured using electrodes placed on or near the muscles of interest.

4 Electrodermal Activity (EDA): Skin's electrical conductivity indicating changes in sweat gland activity, measured using electrodes attached to the fingers or palms.

5 Blood Oxygen Saturation (SpO2): Level of oxygen in the blood measured using optical sensors usually placed on the fingertip or earlobe.

6 Photoplethysmography (PPG): Changes in blood volume within tissues, typically measured using optical sensors placed on the skin, such as the wrist or fingertip.

7 Respiration (RSP): Measurement of the inhalation and exhalation patterns during breathing using sensors placed on the chest or abdomen.

8 Body Temperature: Temperature of the body measured using temperature sensors placed on the skin or within wearable devices.

9 Blood Pressure: Pressure exerted by circulating blood on the walls of blood vessels measured using cuff-based sensors placed around the arm.

Figure 2.10: A simplified diagram illustrating the bio-signals to be measured from the human body using wearable systems and their corresponding sensing points.

normality or abnormality of a heartbeat. Only a limited number of recent studies are related to ECG-based emotional expression recognition [12, 81], while many studies have focused on recognition of emotional expressions using EEG signals due to robust sensing of emotions in the brain [123, 124, 131]. However, the high-dimensionality of EEG signals makes it hard to identify the most effective features for emotional expression recognition. Therefore, some techniques have been proposed to detect emotions by fusing several physiological signals. The techniques of feature fusion can be divided into early and late fusion. Early fusion, also known as feature-level fusion, involves combining extracted features from the signals into a single set before sending them to the classifier. Late fusion, also called decision-level fusion, involves taking the final result by voting on the results produced by several classifiers. The functioning and distinction between early and late fusion for multimodal signals are visually depicted in Figure 2.11.

The most straightforward approach in early fusion is to concatenate the feature vectors from all

modalities, also known as plain early fusion. In a study by Verma and Tiwary, plain early fusion was employed to fuse energy-based features extracted from 32-channel EEG signals. That work achieves a recognition rate of 81.45% for thirteen emotion classifications using SVM [174]. An alternative to feature selection is to encode the dependencies between features. This can be done by using probabilistic inference models like HMM and Bayesian network (BN). For example, a BN was built to fuse features from both EEG and ECG signals in recognising emotions [153].

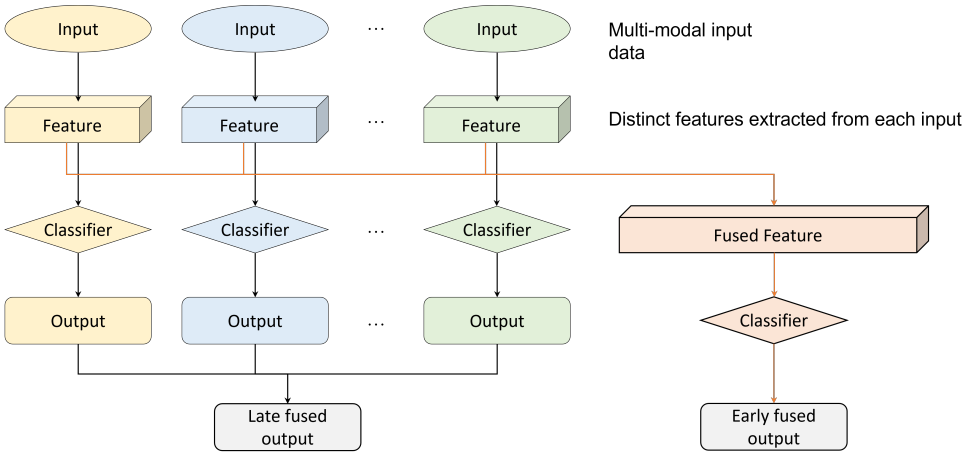


Figure 2.11: An illustrative representation of the process of early and late fusion in handling multimodal signals.

Late fusion involves the use of multiple classifiers that can be trained independently, and the final decision is made by combining the outputs of each classifier. The correspondence between the channels is only identified during the integration step. Since the input signals can be recognised independently, there is no need to put them together simultaneously. A framework for emotion recognition based on the weighted fusion of basic classifiers was proposed [182]. The researchers developed three SVMs using power spectrum, high fractal dimension, and Lempel-Ziv complexity features, respectively. The results of these classifiers were combined using weighted fusion based on each classifier's confidence estimation for each class.

2.4 Databases

Consideration of data used to assess different solutions put forward in the literature is of major importance in every sub-field of modern computer vision. Arguably, considering the relative youth of the field, this consideration is particularly important in the realm of micro-expression recognition. Standardisation of data is crucial in facilitating fair comparison of methods, and its breadth and quality are vital to understanding how well different methods work, their limitations, and what direction new research should follow.

Some of micro-expression related databases include USF-HD [154], Polikovsky's Database [143], York Deception Detection Test (YorkDDT) [195], CASME [205], spontaneous micro-expression corpus (SMIC) [103], CASME II [206], SAMM [29], Chinese academy of sciences macro-expressions and micro-expressions (CAS(ME)^2) [146], facial micro-expressions "in the wild" (MEVIEW) [69], micro-and-macro expression warehouse (MMEW) [10], a third generation facial spontaneous micro-expression database from Chinese academy of sciences (CAS(ME)^3) [101], and a spontaneous 4D micro-expression dataset with multimodalities (4DME) [105]. The nature and purpose of these datasets vary substantially, in some cases subtly, in others less so. In particular, the first three databases are older and proprietary, and contain video sequences with non-spontaneous micro-expression exhibitions. The USF-HD is used to evaluate methods which aim to distinguish between macro-expressions and micro-expressions. Different yet, Polikovsky's database was collected for assessing keyframe detection in the context of micro-expressions, whereas the YorkDDT is specifically aimed at lie detection.

For the acquisition of data for non-spontaneous databases, participants are required to watch the video or image data of the micro-expressions and try to imitate them. Therefore this data should be used with due caution and not assumed to represent the strict ground truth. Therefore, only open-source spontaneous micro-expression databases will be discussed here. These exhibit significant differences between them, and their particularities are important to appreciate so that the findings in the current literature can be interpreted properly and future experiments designed appropriately.

2.4.1 Open-source spontaneous micro-expression databases

Recall that the duration of a micro-expression is usually only 1/25 to 1/5 of a second. In contrast, the frame rate of a regular camera is 25 frames per second. Therefore, if conventional imaging equipment is used, only a small number of frames capturing a micro-expression is obtained, which makes any subsequent analysis difficult. Nevertheless, considering the ubiquity of such standardised imaging equipment, some datasets such as SMIC-VIS and SMIC-NIR, do contain sequences with precisely this frame rate. On the other hand, in order to facilitate more accurate and nuanced micro-expression analysis, most micro-expression datasets in widespread use in the existing academic literature use high-speed cameras for image acquisition. For example, SMIC uses a 100 fps camera and CASME uses a 60 fps one, in order to gather more temporally fine-grained information. The highest frame rate in the existing micro-expression database is at the rate of 200 frames per second. This section provides an overview of each open spontaneous micro-expression database and their specific characteristics.

2.4.1.1 CASME

CASME [205] dataset contains 195 sequences of spontaneously exhibited micro-expressions. The database is divided into two parts, referred to as Part A and Part B. The resolution of images in Part A is 640×480 pixels, and they were acquired indoors, with two obliquely positioned LED lights used to illuminate faces. Part B images have the resolution of 1280×720 pixels and were acquired under natural light. Micro-expressions in CASME are categorised as expressing one of the following: amusement, sadness, disgust, surprise, contempt, fear, repression, or tension; see Figure 2.12. Considering that some emotions are more difficult to excite than others in a laboratory setting, the number of examples across the aforementioned classes is unevenly distributed.

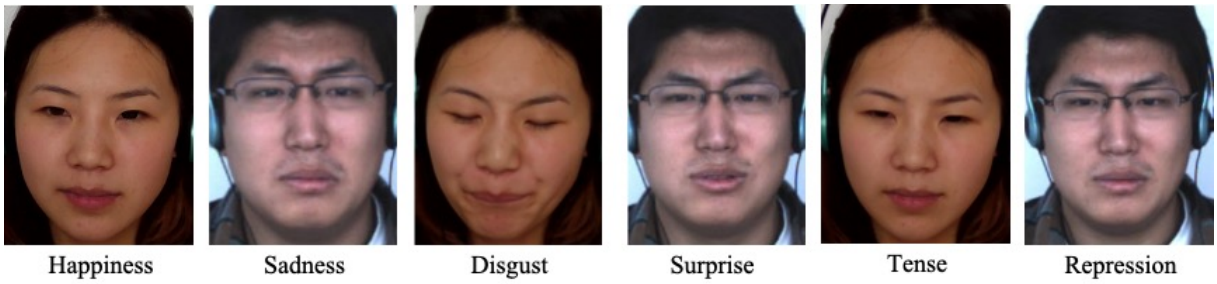


Figure 2.12: Example apex frames from sequences in the CASME database [205].

2.4.1.2 SMIC

SMIC [103], contains videos of 20 participants, exhibiting 164 spontaneously produced micro-expressions. What most prominently distinguishes SMIC from other micro-expression datasets is the inclusion of multiple imaging modalities. The first part of the dataset contains videos acquired in the visible spectrum using a 100-fps high-speed (HS) camera. The second part also contains videos acquired in the visible spectrum (VIS) but at a lower frame rate of 25 fps. Lastly, videos in the near-infrared (NIR) spectrum are included (n.b., only 10 out of 16 individuals in the database). Hence, sometimes reference is made not to SMIC as a whole but to its constituents; see Figure 2.13.

2.4.1.3 CASME II

CASME II [206] dataset is a large collection of spontaneously produced micro-expressions, containing 247 video sequences of 26 Asian participants with an average age of approximately 22 years. The data was captured under uniform illumination, without a strobe. In contrast to CASME, the emotional category labels in CASME II are much broader – namely, happiness,

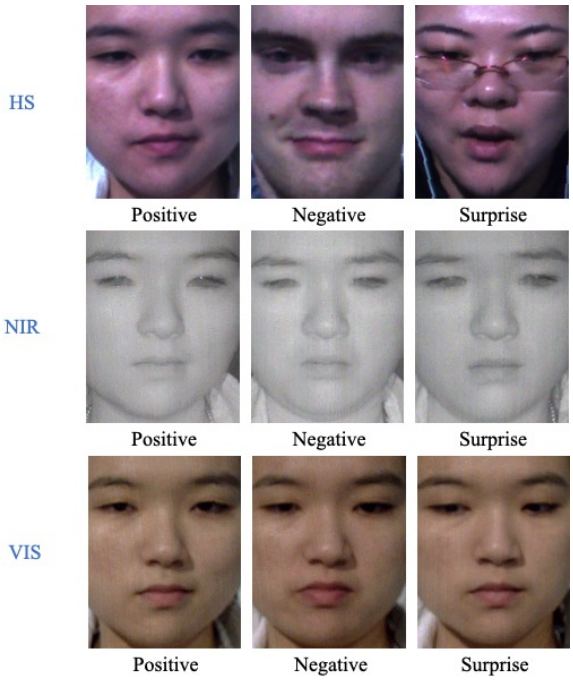


Figure 2.13: Example frames from sequences in the three subsets of SMIC, namely SMIC-HS, SMIC-VIS and SMIC-NIR respectively [103].

sadness, disgust, surprise, and 'others' – thus making the trade-off between class representation and balance, and emotional nuance, in the opposite direction.

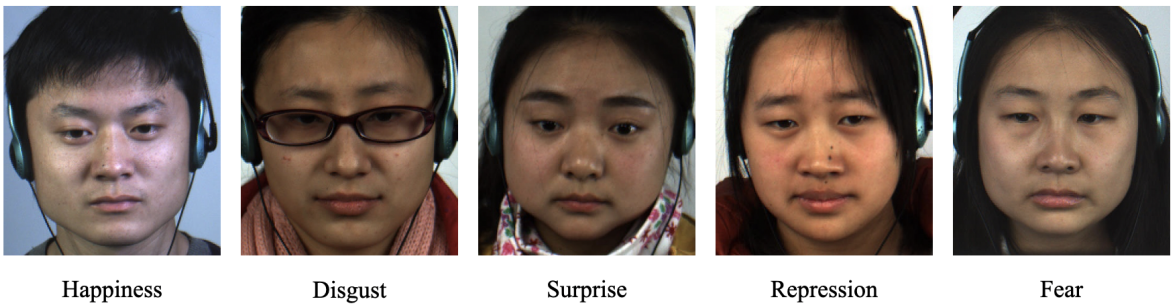


Figure 2.14: Example apex frames from sequences in the (CASME II) database [206].

2.4.1.4 SAMM

The Spontaneous Actions and Micro-Movement (SAMM) [29] dataset is the newest addition to the choice of micro-expression related databases freely available to researchers. It contains 159 micro-expressions, spontaneously produced in response to visual stimulus, of 32 gender balanced participants with an average age of approximately 33 years. Being the most recently acquired dataset, in addition to the standard categorised imagery, SAMM contains a series of annotations

which have emerged as being of potential use from previous research. In particular, associated with each video sequence are the indexes of the frame when the relevant micro-expression starts and ends, and the index of the so-called vertex frame (frame when the greatest temporal change in appearance is observed). In addition to being categorised as expressing contempt, disgust, fear, anger, sadness, happiness, or surprise, each video sequence in the dataset also contains a list of FACS AUs engaged during the expression.

Figure 2.15: Example images from the SAMM dataset [29].

2.4.1.5 CAS(ME)^2

Like several other corpora described previously, the CAS(ME)^2 [146] database is also heterogeneous in nature. The first part of this corpus, referred to as Part A, contains 87 long videos, which contain both macro-expressions and micro-expressions. The second part of CAS(ME)^2, Part B, contains 303 separate short videos, each lasting only for the duration that an expression (be it a macro-expression, or a micro-expression) is exhibited. The numbers of macro-expression and micro-expression samples are 250 and 53 respectively. In all cases, in comparison with most other datasets, the expressions are rather coarsely classified as positive, negative, surprised, or 'other'.

2.4.1.6 MEVIEW

The MEVIEW dataset [69] is a unique micro-expression database in that it is the only one that was collected from the Internet, making it a more realistic and uncontrolled dataset than

those collected under laboratory conditions. The dataset includes a range of video content, such as poker game videos and TV interviews that were downloaded from YouTube. The poker game videos are particularly interesting as they involve players who may try to hide or fake their true emotions, leading to the occurrence of micro-expressions. MEVIEW consists of 31 videos featuring 13 different individuals, with an average length of 3 seconds per video. The videos were labelled using the FACS encoding, which provides a standardised way to describe and quantify facial expressions. The dataset also contains additional information such as the gender and age of the individuals featured in the videos. While MEVIEW is relatively small compared to some of the other micro-expression databases, it provides a unique and valuable resource for researchers interested in studying micro-expressions in real-world scenarios. Its uncontrolled nature presents new challenges and opportunities for researchers to develop and test novel approaches to micro-expression analysis.

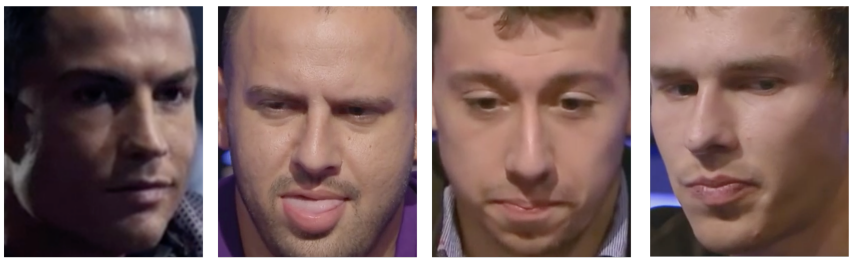


Figure 2.16: Example faces with micro-expression from the MEVIEW database [69]

2.4.1.7 MMEW

MMEW [10] contains both macro- and micro-expressions sampled from the same subjects. This new characteristic can inspire promising future research to explore the relationship between macro- and micro-expressions of the same subject. The samples in MMEW have a larger image resolution (1920×1080 pixels), providing greater visual detail. Furthermore, MMEW has a larger face size in image sequences of 400×400 pixels, which could affect the accuracy of expression recognition algorithms. Moreover, MMEW has more elaborate emotion classes. The emotion classes in MMEW include Happiness (36), Anger (8), Surprise (89), Disgust (72), Fear (16), Sadness (13), and Others (66). In addition to micro-expressions, MMEW also provides 900 macro-expression samples with the same class category (Happiness, Anger, Surprise, Disgust, Fear, Sadness), acted out by the same group of participants. These may be helpful for further cross-modal research (e.g. from macro- to micro-expressions), as well as for exploring differences and similarities between macro- and micro-expressions of the same emotion. An example of macro- and micro-expression samples from MMEW is shown in Figure 2.17.

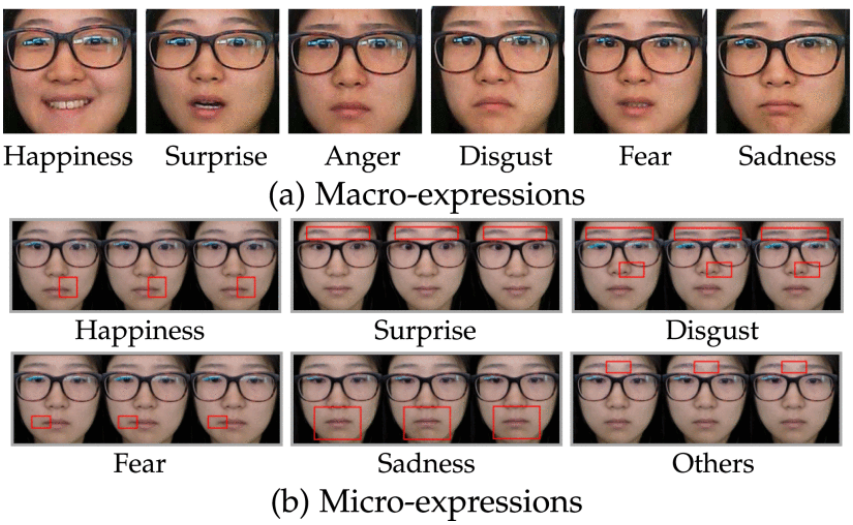


Figure 2.17: Six macro-expressions and six micro-expressions, sampled from the same person, in MMEW database [10]

2.4.1.8 CAS(ME)^3

CAS(ME)^3 [101] is a third-generation multimodal spontaneous micro-expression database that goes beyond just RGB images and includes depth information, physiological, and voice signals. The database was developed to address the challenges of micro-expression elicitation, collection, and annotation. CAS(ME)^3 is composed of three parts. Part A and Part B contain 1,300 labelled long videos with 943 micro-expressions and 3,143 macro-expressions labelled by professional coders, and 1,508 unlabelled long videos of 216 subjects recorded in the same environment and labelled by the same labellers. This helps validate micro-expression analysis methods on a larger scale without database bias.

Additionally, CAS(ME)^3 uniquely introduces multimodality to micro-expression analysis in Part C, which contains 166 micro-expressions and 347 macro-expressions from 31 subjects. Part C employs a third-generation of micro-expression eliciting paradigm, mock crime (refer to Section 2.4.2.4 for details), which offers higher ecological validity. Ecological validity refers to the extent to which research findings can be generalised and applied to real-world situations. In the context of micro-expression analysis, the higher ecological validity of Part C in CAS(ME)^3 signifies that the micro-expressions and macro-expressions captured during the mock crime paradigm more closely resemble the spontaneous expressions that occur in real-life scenarios. This paradigmatic shift allows researchers to study and develop analysis methods that align with real-world contexts, leading to more reliable and practical outcomes. In comparison to the high ecological validity database MEVIEW, Part C doubled the number of micro-expression samples and also labelled macro-expressions, making it possible to enrich multimodal expression analysis

with physiological signals such as heart rate and voice signals. Overall, CAS(ME)^3 has about eight million frames, including 1,109 micro-expressions and 3,490 macro-expressions.

2.4.1.9 4DME

The 4DME dataset [105] is a valuable resource for researchers who want to investigate the benefits of 4D data and multimodal data fusion in micro-expression recognition. It comprises 267 micro-expression samples and 123 macro-expression samples from 41 subjects. The videos have a resolution of 1200×1600 for 4D data and 640×480 for RGB, greyscale, and depth data. They were recorded at a frame rate of 60 fps for 4D and greyscale data and 30 fps for RGB and depth data. The participants in the dataset were 56 in number, aged between 22 and 57, with diverse cultural backgrounds. The micro-expression and macro-expression samples are labelled with 22 categories of AU labels and five categories of emotion labels. What sets 4DME apart from others is that multi-emotion labelling with a maximum of two emotions was allowed. Clips with complex AU combinations were labelled as 'Others,' as were clips containing only 'dependence' AUs. Clips containing key AUs for both 'Positive' and 'Negative' were assigned to 'Others' as these two emotions are conflicting. Researchers can use this database to investigate whether 4D data can improve micro-expression recognition performance and how the fusion of various data sources could facilitate the task of micro-expression recognition.

2.4.1.10 Conclusion and comparative analysis

Table 2.5: A recap of spontaneous micro-expression databases

Database	Micro-expressions	Participants	FPS	Ethnicities	Average Age	Resolution	Facial Resolution
CASME [205]	195	35	60	1	22.03	640×480 1280×720	150×190
SMIC [103]	164 71 71	20 10 10	100 25 25	3	26.7	640×480	190×230
CASME II [206]	247	35	200	1	22.03	640×480	280
SAMM [29]	159	32	200	13	33.24	2040×1088	400×400
CAS(ME)^2 [146]	57	22	30	1	22.59	640×480	N/A
MEVIEW [69]	40	16	25	N/A	N/A	1280×720	N/A
MMEW [10]	300	36	90	1	22.35	1920×1080	400×400
CAS(ME)^3 [101]	943 N/A 166	100 116 31	30	1	22.74	1280×720	N/A
4DME [105]	267	65	60 30	7	27.8	1600×1200 640×480	150×150

At present, the amount of micro-expression databases and the number of micro-expression samples contained in each database is minimal. Therefore, combining different databases in

Table 2.6: Labelled emotion classes included in spontaneous micro-expression databases

Database	Number	Emotions
CASME [205]	8	Happiness, Sadness, Disgust, Surprise, Contempt, Fear, Repression, Tense
SMIC [103]	3	Positive, Negative, Surprise
CASME II [206]	5	Happiness, Disgust, Surprise, Repression, Others
SAMM [29]	7	Disgust, Fear, Anger, Sadness, Happiness, Surprise, Others
CAS(ME)^2 [146]	4	Positive, Negative, Surprise, Others
MEVIEW [69]	7	Contempt, Disgust, Fear, Anger, Happiness, Surprise, Unclear
MMEW [10]	7	Disgust, Fear, Anger, Sadness, Happiness, Surprise, Others
CAS(ME)^3 [101]	7 4	Happiness, Disgust, Fear, Anger, Sadness, Surprise, Othes Positive, Negative, Surprise, and Others
4DME [105]	5	Positive, Negative, Surprise, Repression, Others

an experiment may be an approach method for the training of MER models at present. When the currently available spontaneous micro-expression databases are considered, each of them can be seen to offer some kind of advantage over the others; nevertheless, the amount of data in any of them not meet the requirements of the traditional deep learning algorithms. SAMM and CASME II have the highest frame rate of 200fps, and SAMM has the highest resolution. The SMIC database contains both high-speed camera samples as well as samples suitable as training data for a model used in a typical non-high-speed camera environment. CAS(ME)^2 contains not only the FACS information and emotion labels associated with individual micro-expressions, but also can be used to distinguish between macro- and micro-expressions. MMEW contains both macro- and micro-expressions with the same participants, and includes Anger, Fear, and Sadness classes not found in CASME II. On the other hand, MEVIEW is the only database that includes "in the wild" samples, while CAS(ME)^3 and 4DME provide depth information. Additionally, CAS(ME)^3 is the only database that offers multimodal data with voice and physiological signals. A detailed comparison of each database is provided in Table 2.5 and 2.6.

2.4.2 Data collection and methods for systematic micro-expression evocation

One difficulty in the process of collecting micro-expression video sequence corpora lies in the difficulty of inciting micro-expressions in a reliable and uniform manner. A common approach adopted in the published literature consists of presenting participants with emotional content (usually short clips or movies) which is expected to rouse their emotions, while a the same time asking them to disguise their emotions and maintain a neutral facial expression. A typical data acquisition setup is diagrammatically shown in Figure 2.18.

When the aforementioned data collection protocol is considered with some care, it is straightfor-

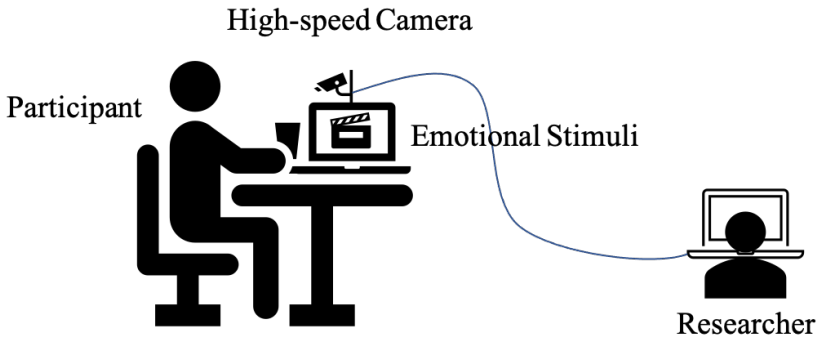


Figure 2.18: For the purposes of data collection, participants watch an emotional video while their faces are imaged by a high-speed camera.

ward to see that a number of practical problems present themselves. Firstly, in some instances, the assumption that the content presented to the participants will elicit sufficient emotion may be invalidated. Thus, no meaningful micro-expression may be present in a video sequence of a person's face (e.g., in SMIC out of 20 individuals who participated in the recording sessions, only 16 exhibited sufficiently well expressed micro-expressions). This problem can be partially ameliorated by ensuring the stimuli are strong enough, though this must be done with due consideration of possible ethical issues. On the complementary side, so to speak, considering that micro-expressions are involuntarily expressed, it is important to suppress as much as possible any conscious confound. In other words, there must exist sufficient incentives to encourage participants to conceal their true feelings.

2.4.2.1 CAS series data acquisition protocol

During data collection of CASME [205], CASME II [206] and CAS(ME)^2 [146], participants were asked to watch different emotional videos while maintaining a neutral facial expression. As explained before, the intention is to incite involuntary, micro-expressions, rather than have them acted, which results in data which is not realistic. During the collection process, the participants were required to remain expressionless and not move their bodies, thus removing any need for body or head pose normalisation. Lastly, as a means of encouraging participants to conceal their emotions, they were offered the potential of a monetary award. Specifically, the award was paid out if a participant successfully managed to hide their emotion from the researcher supervising the process (the researcher was unaware of the video content).

2.4.2.2 SMIC data acquisition protocol

Much like in the case of three Chinese academy of sciences (CAS) databases, in the process of data collecting for SMIC [103], the participants were shown emotional videos and asked to attempt to conceal their reactions, and a researcher, unaware of the video content watched, was asked to guess the participants' emotions. Unlike for CAS series databases when participants were incentivised by a reward for success (in hiding their emotions), now participants were *disincentivised* by a 'punishment' – namely, unsuccessful participants had to fill in a lengthy questionnaire.

2.4.2.3 SAMM data acquisition protocol

Highlighting the point I made previously – the need to understand well the nuanced differences between different micro-expression datasets – the data acquisition protocol employed in collecting the SAMM dataset is different still from all of the described thus far. Firstly, all participants were asked to fill out a questionnaire before the actual imaging session. The aim of this was to allow the researchers to personalise emotional stimuli (e.g. exploiting specific individual's fears, likes and dislikes, etc.). Additionally, and again in contrast with e.g. CAS databases (2nd generation), in order to make the participants more relaxed and less affected by their knowledge that they are partaking in an experiment, the participants were filmed without any supervision of or oversight by the researchers.

2.4.2.4 CAS(ME)^3 data acquisition protocol

The latest addition to the CAS series is CAS(ME)^3, which differs significantly from the previous three databases. CAS(ME)^3 was constructed with depth information. Parts A and B were collected using a second-generation elicitation paradigm, the same as the previous CAS databases. Part A and Part B consist of 100 and 116 subjects, respectively, each asked to watch 13 emotionally stimulating videos and attempt to maintain a neutral expression. Part C, on the other hand, employed a mock crime scenario to elicit spontaneous micro-expressions. Participants were asked to steal a small amount of money from an envelope and were subsequently questioned about the theft. The scenario was designed to create a stressful situation that would elicit spontaneous micro-expressions associated with guilt or deception. The videos collected from Part C were used to evaluate the performance of algorithms in recognising spontaneous micro-expressions in a real-world scenario. The use of the mock crime scenario adds an element of realism to the dataset and makes it more applicable to real-world situations.

2.4.2.5 The three generations of micro-expression elicitation methods

The study of micro-expressions has gained significance in various fields such as psychology, criminology, and human-computer interaction. However, the ecological validity of micro-expression databases is critical in determining their suitability for analysing micro-expressions in complex real-world situations, which depends on the elicitation paradigm used. To address this issue, researchers have classified published databases into three generations based on the micro-expression elicitation methods, each with increased ecological validity.

The first generation involved posing fleeting facial expressions by actors who observed standard expressions, resulting in databases such as USF-HD and Polikovsky's database. These posed micro-expressions lacked ecological validity, limiting their usefulness. The second generation involved collecting micro-expressions using emotional stimuli through the neutralisation paradigm, as seen in databases such as CASME, SMIC, CASME II, SAMM, and MMEW. While these were more spontaneous, they were still collected in lab settings, and their ecological validity was limited.

The third generation, which includes paradigms such as mock crime, dictator games, and prisoner's dilemma, is designed to elicit micro-expressions with high ecological validity. MEVIEW, an in-the-wild database by Husak et al., contains video clips from real scenarios, but they have too many uncontrollable factors. Therefore, micro-expression samples still need to be collected in well-controlled laboratory scenarios, such as the subset collected by the mock crime paradigm in CAS(ME)^3. This subset has improved ecological validity and eliminates uncontrollable factors, making it more suitable for micro-expression elicitation and research related to its application in lie detection.

Overall, the three generations of micro-expression elicitation methods reflect the evolution of micro-expression research toward improving ecological validity. The first and second generations laid the foundation for this research, while the third generation is advancing it further by collecting micro-expression samples from more ecological situations. This evolution is crucial for improving the accuracy and robustness of micro-expression analysis, allowing researchers to better understand and analyse real-life scenarios where micro-expressions are prevalent.

CHAPTER THREE

FACIAL ACTION UNIT DETECTION FROM MICRO-EXPRESSION

Micro-expressions describe unconscious facial movements which reflect a person's psychological state even when there is an attempt to conceal it. Often used in psychological and forensic applications, their manual recognition requires professional training and is time-consuming. Therefore, achieving automatic recognition by means of computer vision would confer enormous benefits. AU is a coding of facial muscular complexes which can be independently activated. Each AU represents a specific facial action. In this chapter, I propose a method for the challenging task that is the detection of activated AUs when a micro-expression occurs, which is crucial in the inference of emotion from a video capturing a micro-expression. This specific problem is made all the more difficult in the light of limited amounts of data available and the subtlety of micro-movements. I propose a segmentation method for key facial sub-regions based on the location of AUs and facial landmarks, which extracts 11 facial key regions from each sequence of micro-expression images. AUs are assigned to different local areas for multi-label classification. Considering that there is little prior work on the specific task of *detection* of AU activation in the existing literature on micro-expression analysis, for the evaluation of the proposed method I design an AU independent cross-validation method and adopt unweighted average recall (UAR), unweighted F1-score (UF1), and their average as the scoring criteria. Evaluated using the established standards in the field and compared with previous work, my approach is shown to exhibit state-of-the-art performance.

3.1 Motivation

Facial expressions can reflect human emotions. Due to different cultural environments, individuals use different languages to communicate, but their emotions are expressed by the same facial expressions [40]. In addition to the regular macro-expressions which take place on larger time scales, small and speedy movements that are inadvertently exhibited for short periods of time, better reveal emotions which individuals attempt to conceal. Ekman and Friesen first reported on a case of these particular expressions. In a recording of a conversation between a psychiatrist and a patient with depression, there are occasional frames with very painful expressions of a patient otherwise displaying a happy appearance in the video. Researchers call that kind of fast, unconscious, spontaneous facial movements such people produce when they experience intense emotions, micro-expression. Micro-expressions usually happen within less than 0.5 seconds. If the occurrence of micro-expressions is detected and the emotional meaning represented by them is recognised, the real mental activities of individuals could be accurately identified.

Facial *actions* are distinguishable from facial expressions. AUs correspond to muscular complexes that are activated during facial movements. Moreover, external stimuli, such as a gust of wind blowing across the face, can elicit the activation of AUs. Facial expressions are observable as facial movements that arise from diverse cognitive processes, including emotional responses. Micro-expressions, for instance, frequently arise when an individual endeavours to suppress or conceal emotions [58]. The precise identification of micro-expressions allows us to comprehend authentic emotions, establishing a crucial foundation for discerning individuals' subjective experiences in domains such as public safety and psychotherapy. Analysing the AUs embedded within micro-expressions represents the most intuitive approach to decipher the emotions conveyed by micro-expressions, utilised in manual micro-expression recognition. Consequently, facial AUs can also be considered an intermediary variable in automatic recognition, bridging micro-expressions and emotions.

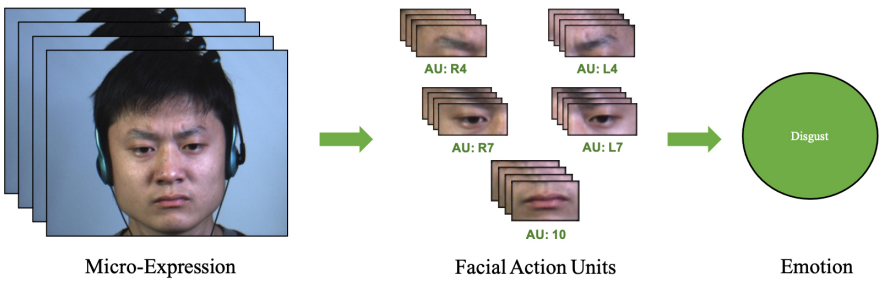


Figure 3.1: Example of AUs detected from a micro-expression and recognised to an emotion.

Due to the small range of movement and the short duration of facial movements when micro-

expressions happen, individuals need professional training to recognise micro-expressions manually. The human based processes of training, as well as recognition itself, are demanding, yet the recognition accuracy is still not satisfactory for most practical purposes. Many researchers have tried using and developing new computer vision techniques to recognise micro-expressions automatically. This automatic approach to the identification of micro-expressions has unique advantages, which significantly improves the feasibility of micro-expression applications. No matter how fast the facial movement is, as long as the camera records it, the computer can obtain the corresponding information and process it. In addition, once an efficient and stable model is trained, it can process large volumes of micro-expression data at low cost, which far exceeds the efficiency of manual recognition of micro-expressions by professionals. Thus, some research uses high-speed cameras for the collection of micro-expressions. Recently, following the publication of open-source micro-expression databases, the amount of work related to micro-expressions increased every year. Research thus far all but invariably uses 3DHOG [143], LBP-TOP [140], HOOF [113] and their variations or deep learning methods as features. However, most previous work focuses on emotion recognition directly. Even though some considered AUs as supplementary features [203, 97, 115], there is tiny work on micro-expression AU detection task specifically.

The importance of proposing separate methods for micro-expression AU detection is paramount in the field. In previous works, researchers have extensively utilised local information for facial AU detection, primarily focusing on macro-expressions [219, 162, 232, 218]. These approaches often leveraged well-defined regions and composition rules to recover facial expressions through sparse coding or patch selection methods. With the advent of deep learning techniques, there has been a surge in AU detection studies for macro-expressions [221, 220], showcasing the power in capturing nonlinear representations. However, the transition to micro-expression AU detection demands specialised methodologies that consider the distinct challenges posed by the subtle and limited quantity of micro-expressions. Specifically, unlike macro-expressions, which can often be adequately captured and analysed using single images, micro-expressions are characterised by subtle movements that may only become evident when observed dynamically over time. Therefore, employing video samples for micro-expression AU detection is essential to capture the temporally sensitive changes in facial expressions.

Therefore, in this chapter, I focus on the AU detection task for micro-expression analysis and demonstrate my proposed framework can achieve the effect of state-the-art in the task, even without deep learning methods. In addition, since my method does not use deep learning, it does not require a lot of time for training and running. It can almost meet the requirement of real-time detection. Taking CASME II as an example, it only takes about *1s* to complete the AU detection

test of all samples using my framework.

The main contributions of this chapter are as follows:

1. I proposed a novel facial key subregion segmentation method based on the facial muscle of AU activated and a novel framework to detect multi-labelled facial micro-expression AUs by transferring a big multi-label classification to several small ones based on the segmented regions.

2. I design an AU independent 5-fold cross-validation method for Facial AU detection in micro-expression and conduct intensive experiments on two publicly micro-expression databases with AU labels. The results represent the effectiveness of my approach.

3.2 Methodology

Recall that my main aim in the present work is the identification of AUs activated during a facial micro-expression. Thus, the proposed method can be broadly seen as comprising the following stages: facial sub-region segmentation, facial sub-region feature extraction, and multi-label classification. These are summarised in Figure 3.2 and explained in detail hereafter.

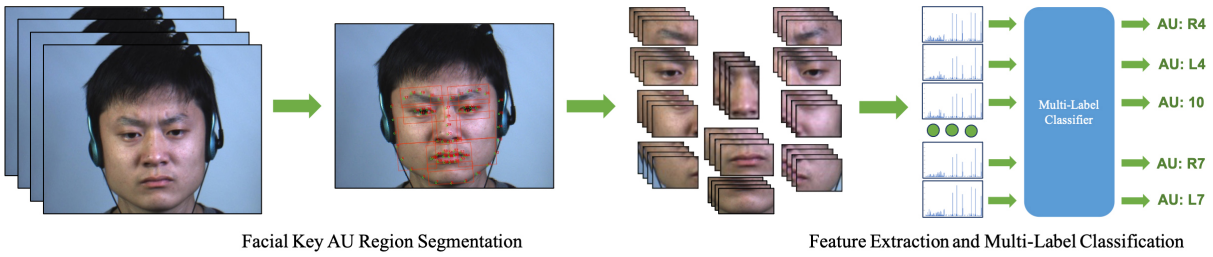


Figure 3.2: Proposed framework of facial action unit detection for micro-expression.

3.2.1 Local facial region segmentation

FACS [40] is currently recognised as the universal standard for encoding facial actions, associating each facial action with an AU. It serves as the foundation for labelling facial movements in micro-expression datasets. Thus, the primary objective of this research is to detect all AUs activated when a micro-expression is displayed using a sequence of images (note that the majority of micro-expressions involved in emotion inference activate multiple AUs). As AUs are canonical and elementary primitives used to describe facial movements, it follows from the anatomical structure of the face that a specific AU is spatially localised, corresponding to a specific sub-region of the face. For instance, AU1, AU2, and AU4 specifically describe

movement in the eyebrow area. Consequently, segmenting the face into multiple sub-regions and identifying the AUs present in each sub-region is less complex than simultaneously identifying all AUs across the entire face. To mirror the spatial layout of AUs, a segmentation method is proposed, dividing the facial region into 11 sub-regions: Left and Right Brow, Left and Right Eye, Left and Right Cheek, Left and Right Nasolabial Area, Nose, Mouth, and Chin. The local facial region segmentation method is based on facial landmark detection – a crucial step in many face recognition and analysis algorithms. The task involves the localisation of salient areas of a face, such as the eyebrows, eyes, nose, mouth, or face contour, from a given image of a face.

I pursued the standard 68 key-point positioning strategy [] to ascertain the precise landmarks on the face. The results of landmark recognition are summarised in Figure 3.3. In order to accomplish the localisation of these key points, the ensemble of regression tree (ERT) algorithm [83] was employed. ERT represents a regression tree method that exploits gradient boosting to enhance the learning process. It employs cascading regression factors alongside multiple GBDTs (Gradient Boosted Decision Trees), wherein the leaf nodes act as repositories for the residuals. Throughout the regression process, when the input data is situated within a specific node, the corresponding residual is incorporated into the input to refine the regression. Ultimately, the amalgamation of all the residuals allows for the determination of the final position of the facial landmarks.

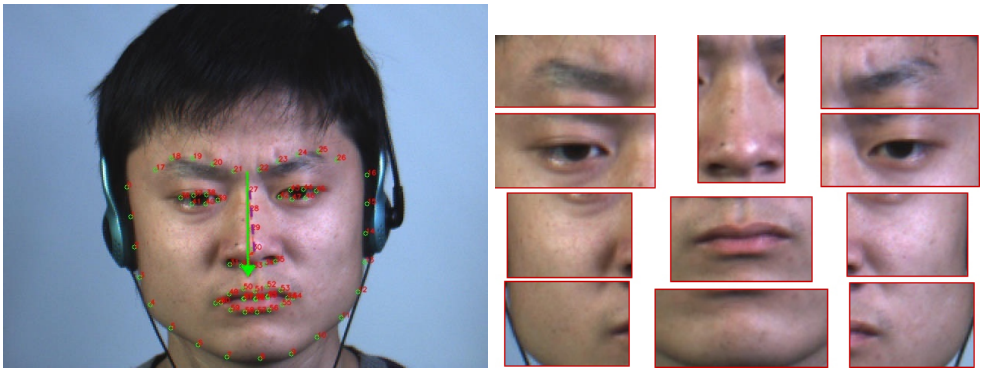


Figure 3.3: An example of detected landmarks and segmentation areas in a micro-expression image from CASME II database.

Due to the brief duration and limited range of muscular movement exhibited during each micro-expression, the subject's pose undergoes minimal changes. As a result, it is unnecessary to detect all facial key points in every frame of a micro-expression image sequence. Instead, I employ the central frame of the image sequence as the reference image for identifying facial landmarks, and subsequently extend these results to each frame within the micro-expression sequence. Additionally, I propose a segmentation method for facial sub-regions based on the 68 facial landmarks present in each micro-expression sequence.

As depicted in Figure 3.3, points 27–30 facilitate the establishment of the nose's central line, which spans the facial region captured in the image. This line serves as the vertical reference during the segmentation process for the entire face area. In order to ensure that all facial sub-regions are in a vertical alignment, each sub-region image undergoes rotation based on this vertical line. The precise delineation of the 11 chosen face sub-regions is illustrated in Figure 3.3. The regions encompassing the eyebrows, eyes, nose, and mouth are determined based on their corresponding landmarks. Likewise, the cheek and nasolabial areas are ascertained by considering the upper and lower contour points of the face, in addition to the landmarks situated in the upper lip region (e.g. points 0, 4, and 50). The chin region, on the other hand, is defined by the lower lip point 57 and the lowest point 8 along the facial contour.

3.2.2 Sub-regional feature extraction and multi-label classification

The LBP-TOP feature extraction method serves as a prominent approach in the realm of micro-expression recognition research and often serves as a baseline model for novel investigations in this field. LBP-TOP features capture the interplay between the appearance of a pixel and its surrounding neighbourhood. To encode the spatio-temporal co-occurrence model, feature extraction is performed on three different planes: the XY plane, the XT plane, and the YT plane within the image sequence. The method outlined by Hong et al. was employed to determine the radii, denoted as RX, RY, and RT, along the three space-time axes (X, Y, T). Subsequently, uniform sampling of points is conducted on each plane using ellipses determined by the corresponding axes in the respective space-time plane. The purpose of this sampling is to calculate the local binary mode on each plane. Following this step, the histogram of data within each facial sub-region is utilised to extract unified features pertaining to each facial AU.

In the proposed method, LBP-TOP feature extraction is performed on the image sequence of each key facial sub-region, rather than the entire face. This approach enables a focused analysis of the meaningful key facial components involved in micro-expressions. Additionally, irrelevant facial information unrelated to emotion and AUs can be disregarded, thereby promoting the specificity of the features utilised in the learning process.

Upon extracting the features from each facial sub-region, the subsequent step involves conducting multi-label classification utilising micro-expressions labelled with multiple AUs. Traditional supervised learning predominantly concentrates on single-label learning scenarios. Nevertheless, real-life target samples often possess greater complexity, exhibiting multiple semantics and encompassing multiple labels. This is particularly evident in micro-expression AU detection tasks, where the majority of expressions involve the activation of more than one facial AU, thereby making multi-label learning the natural and suitable choice.

The strategy of multi-label classification is to transform the problem structurally, to make the extracted features more readily usable by existing single-label learning algorithms. Firstly, I apply the Label Powerset (LP) algorithm to transform a multi-label learning problem into a multi-class (single-label) classification task. This is achieved by learning one single-label classifier $h : X \rightarrow P(L)$, where L is a set of disjoint labels, $P(L)$ is the powerset of L, containing all possible label subsets. The label set predicted by LP is already in the training set, and it cannot be generalised to the unseen label set. In order to overcome this limitation of LP, the LP classifier used by random k-labelsets (RAkELd) [171] only trains a subset of length k in Y output dataset and then integrates a large number of LP classifiers to predict. In general, this type of method considers the relationship between the class labels, but for datasets with many class labels and a large amount of data, the computational complexity of problem transformation is an obvious limitation. However, micro-expression datasets are not big enough and facial sub-region segmentation process reduced the number of labels of each sample. So, this limitation of these methods has little effect in the present context.

The main purpose of employing the LP algorithm is to convert the multi-label classification problem into a single-label one. Every combination of different labels is henceforth considered a class in itself. This algorithm will generate more classes when there are more labels. Therefore, if it is applied to the entire face image since the number of all micro-expression AUs appearing on the entire face is large, the result of learning when they are all used as labels in one multi-label classification is very poor. However, after I segment the face according to the range of AUs, the number of AUs that may appear in each salient area is much lower across than the whole facial region. Hence, segmentation is crucial in preventing an excessive increase in the computational cost of the LP algorithm, resulting in far better performance. RAkELd algorithm is a variant of the LP algorithm. It converts an LP from multiple labels into multiple LPs of length k to predict the results jointly. This method can effectively reduce complexity when there are too many types of labels in the LP algorithm. Finally, I adopt the Gaussian Naïve Bayes algorithm on the extracted sub-regional features to learn a model of multi-labelled AUs activated during micro-expressions.

3.3 Experimental assessment

It is important to emphasise that all of the aforementioned corpora were acquired in relatively controlled conditions for the specific purpose of micro-expression analysis. In particular, the data acquisition process involved the participants watching emotional videos while attempting to hide the facial expression of the aroused emotions. Thus, these datasets are more standardised and easier to process compared to datasets collected from unconstrained real-world scenarios.

Considering that the duration of a micro-expression is usually under 0.5s, in order to capture a greater number of image frames of images during the occurrence of micro-expressions, data is usually acquired using high-speed cameras. SAMM and CASME II contain data with the highest frame rate of 200 frames per second. To evaluate the effectiveness of my proposed method, I initially apply it to the CASME II dataset, followed by its application to the combined dataset of CASME II and SAMM.

3.3.1 Data preparation

For my experiments, I chose to adopt the use of the CASME II database [206], which is widely used in the field owing to its size and high video frame rate. Its collection method is spontaneously induced, which is representative of real-world conditions. Professional psychologists have marked all AUs in each micro-expression image sequence. A total of 19 AUs were included in CASME II, namely AU 1, 2, 4–7, 9, 10, 12, 14–18, 20, 24–26 and 38. In my experiment, the 11 facial sub-regions are the smallest modules. Therefore, these 19 AUs are separated into each sub-region according to the area where they appear. However, because some AUs may appear in both left and right half of the face, such as AU1 (Inner Brow Raiser), they are included in both left and right facial areas. In addition, the original labels in CASME II also include some single-side AUs, such as L1, L2 and R4. Therefore, I divide AUs which are activated on both sides into two parts. For example, the initial both-side label *AU1* is relabelled as *L1&R1*, and the initial single-side label *L1* remains. Finally, in my experiment, a total of 26 AUs were included. The specific AUs included in each facial sub-region are shown in Table 3.1.

Table 3.1: AUs in each local key facial sub-regions

Facial Sub-region	**AUs**
Left Brow	L1, L2, L4
Right Brow	R1, R2, R4
Left Eye	L5, L7
Right Eye	R5, R7
Left Cheek	L6
Right Cheek	R6
Nose	9, 38
Mouth	10, 12, 15, 16, 18, 20, 24, 25
Chin	17, 26
Left Nasolabial Area	L14
Right Nasolabial Area	R14

The types of AUs labelled in the SAMM micro-expression dataset are more abundant than those in CASME II. However, there are several rare AUs only activated in one or two samples of

micro-expression. After analysis of the AU labels in SAMM and CASME II, I find the relabelled AUs I previously described for CASME II are the most common ones in both datasets. In order to unify the evaluation criteria of the experiment, I only used the 26 AUs as I described and relabelled AUs of samples in SAMM. The other rare AU labels were deleted, and only the AUs in Table 3.1 were used for the experiment. Due to there are no AU16 and AU38 labelled in SAMM, the final AU number applied in SAMM is 24.

3.3.2 Metrics

Algorithm 1 AU independent 5-fold data splitting

```
procedure PREPAREDATA
    Load the samples and their AUs.
end procedure
procedure CREATEAUBINS
    for each unique AU do
        Create an empty bin for the AU
    end for
end procedure
procedure INITIALISEFOLDS
    Create 5 empty folds to store the samples
end procedure
for each sample, AUs in dataset do
    for each AU in AUs do
        Add the sample to the respective AU bin
    end for
end for
for each AU bin do
    Randomly shuffle the samples in the bin
    Calculate the number of samples per fold
    for each fold do
        Distribute the required number of samples from the bin to the fold
    end for
end for
```

The frequency of activation of AUs is different across facial expressions. Some AUs are more commonly activated than others, such as AU4 (Brow Lowerer), which is the most frequently engaged AU. AU26 (Jaw Drop), is activated less in micro-expressions than others, especially when participants are asked to suppress their expressions. Therefore, in the model training process, in order to make sure that all the AUs' features could be learned, I randomly separate data that each AU appears in each facial sub-region into 5 folds and each time 4 of them as a training set. Thus, I ensure that samples of all AUs are in my training corpus. The remaining

subset of the micro-expression data is used as a test dataset to evaluate the final algorithm results. The splitting algorithm is shown in Algorithm 1. In this way, an AU independent 5-fold cross-validation strategy is applied to evaluate the proposed method.

As emphasised already, there is virtually no AU detection work in the context of micro-expressions and no standard metrics which could be used for evaluation in this realm. Therefore, I adopt the evaluation approaches from other AU detection work, as well as the metrics used in as related as possible micro-expression analysis problems. Accuracy and F1-score are widely used criteria in both AU detection and micro-expression recognition. The distribution of the number of each AU in the micro-expression database is unbalanced, so I choose UF1, and UAR to show the performance of my approach and equalise the influence of each AU.

$$Precision_c = \frac{\sum_{i=1}^{F} TP_{i,c}}{\sum_{i=1}^{F} TP_{i,c} + \sum_{i=1}^{F} FP_{i,c}}, \tag{3.1}$$

$$Recall_c = \frac{\sum_{i=1}^{F} TP_{i,c}}{\sum_{i=1}^{F} TP_{i,c} + \sum_{i=1}^{F} FN_{i,c}}, \tag{3.2}$$

$$F1\text{-}score_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c}, \tag{3.3}$$

$$UF1 = \frac{\sum_{c=1}^{C} F1\text{-}score_c}{C}, \tag{3.4}$$

$$UAR = \frac{\sum_{c=1}^{C} \frac{\sum_{i=1}^{F} TP_{i,c}}{N_c}}{C}, \tag{3.5}$$

where $TP_{i,c}$, $FP_{i,c}$ and $FN_{i,c}$ are true positive, false positive and false negative for each class c (of C AUs, 26 in my experiments), when samples of fold i as the test set. $Precision_c$ and $Recall_c$ represent the fraction of AU_c is correctly identified and the number of correct detection of AU_c over the actual number of samples with AU_c active. F is the number of fold (5), and N is the total number of samples. After obtaining the average of UAR and $UF1$, this quantity is used as the final evaluation score, which is also the comparison criterion used in the EmotionNet Challenge [11] (a popular AU detection challenge "in the wild").

Furthermore, to facilitate a more comprehensive comparison with the work conducted by Li et al., and LBP-TOP [140], LBP-SIP [191], and 3DHOG [143] they employed as the benchmark methods, I also adopted their subject-independent 4-fold cross-validation approach on both the CASME II and SAMM datasets. In this evaluation setup, two folds are allocated for training, while the remaining folds are used for validation and testing, respectively. However, I retained the employment of multi-label learning for the 26 selected AUs, as all of these AUs were also

Table 3.2: Experimental scores on CASME II, SAMM and CASME II & SAMM with AU independent 5-fold cross-validation

	CASME II			SAMM			CASME II&SAMM		
AU	**Accuracy**	**F1**	**Score**	**Accuracy**	**F1**	**Score**	**Accuracy**	**F1**	**Score**
L1	0.7729	0.6563	0.7146	0.9466	0.7525	0.8495	0.6361	0.5314	0.5838
R1	0.7649	0.5970	0.6810	0.9084	0.6006	0.7545	0.6516	0.4380	0.5448
L2	0.7656	0.6020	0.6838	0.7863	0.5492	0.6677	0.6411	0.5166	0.5789
R2	0.8433	0.6170	0.7301	0.7634	0.4889	0.6261	0.7018	0.4356	0.5687
L4	0.6813	0.6720	0.6767	0.8626	0.6796	0.7711	0.7030	0.6555	0.6792
R4	0.5037	0.4866	0.4951	0.8168	0.4876	0.6522	0.5664	0.5374	0.5519
L5	0.9728	0.7597	0.8662	0.7109	0.4155	0.5632	0.7818	0.4815	0.6317
R5	0.9572	0.4891	0.7231	0.7344	0.4505	0.5924	0.7714	0.4764	0.6239
L6	0.5804	0.3758	0.4781	0.9844	0.4961	0.7402	0.4569	0.3177	0.3873
R6	0.6902	0.4084	0.5493	0.9531	0.4880	0.7206	0.6527	0.3949	0.5238
L7	0.6070	0.4974	0.5522	0.5156	0.4507	0.4832	0.8234	0.6891	0.7563
R7	0.5720	0.4448	0.5084	0.5156	0.4572	0.4864	0.8234	0.6688	0.7461
9	0.7490	0.4431	0.5960	0.7734	0.4678	0.6206	0.7363	0.4241	0.5802
10	0.7985	0.5580	0.6782	0.9044	0.4749	0.6897	0.8947	0.5967	0.7457
12	0.6844	0.5645	0.6245	0.6471	0.5810	0.6140	0.6291	0.5158	0.5724
L14	0.7255	0.4205	0.5730	0.7344	0.4234	0.5789	0.4465	0.3312	0.3888
R14	0.7569	0.4987	0.6278	0.7422	0.4539	0.5980	0.3760	0.2919	0.3339
15	0.8821	0.6634	0.7728	0.9926	0.9425	0.9676	0.5564	0.4227	0.4896
16	0.9848	0.8294	0.9071	-	-	-	0.9900	0.8308	0.9104
17	0.7137	0.5118	0.6128	0.7891	0.4410	0.6151	0.4909	0.4015	0.4462
18	0.9962	0.4990	0.7476	0.9779	0.7801	0.8790	0.9424	0.4852	0.7138
20	0.9924	0.4981	0.7452	0.9265	0.4809	0.7037	0.8997	0.4736	0.6867
24	0.9620	0.6331	0.7975	0.9706	0.8256	0.8981	0.9599	0.7039	0.8319
25	0.9924	0.4981	0.7452	0.9338	0.7176	0.8257	0.9599	0.5897	0.7748
26	0.9961	0.4990	0.7475	0.9141	0.5543	0.7342	0.9086	0.4761	0.6923
38	0.9922	0.4980	0.7451	-	-	-	0.9948	0.4987	0.7467
	UAR	**UF1**	**Score**	**UAR**	**UF1**	**Score**	**UAR**	**UF1**	**Score**
Final	**0.8053**	**0.5469**	**0.6761**	**0.8252**	**0.5608**	**0.6930**	**0.7306**	**0.5071**	**0.6188**

chosen by Li et al..

3.3.3 Results and discussion

The test results of the models trained on CASME II, SAMM and CASME II & SAMM by the proposed method are shown in Table 3.2. Firstly, observe that the proposed approach achieves excellent results across the different micro-expression databases, testifying to the value of my multi-label AU detection based approach. It is also important to note the model performed equally well across the entire set of AUs. This finding demonstrates that my method can effectively address the challenge posed by highly unbalanced multi-labelled data, which is crucial for real-world applicability.

Table 3.3: F1-scores on CASME II dataset, with subject independent 4-fold cross-validation

AU	LBP-TOP	LBP-SIP	3DHOG	SCA[106]	Mine
1	0.1057	0.2308	0.2771	0.2857	**0.4678**
2	**0.4985**	0.3892	0.2769	0.4532	0.4786
4	0.7324	0.7354	0.7012	**0.8877**	0.5706
7	0.0635	0.0888	0.0000	0.2473	**0.5160**
12	0.2386	0.2143	0.0526	0.4792	**0.5528**
14	0.2185	0.2979	0.0000	0.3327	**0.5070**
15	0.0000	0.4318	0.0000	0.3954	**0.4754**
17	0.1667	0.4287	0.1212	**0.5159**	0.4776
UF1	0.2530	0.3521	0.1786	0.4496	**0.5057**

Table 3.4: F1-scores on SAMM dataset, with subject independent 4-fold cross-validation

AU	LBP-TOP	LBP-SIP	3DHOG	SCA[106]	Mine
2	0.2652	0.2144	0.0000	0.3289	**0.4873**
4	0.1538	0.0556	0.1667	0.1297	**0.4692**
7	0.4603	0.0400	0.2330	**0.4876**	0.4072
12	0.2376	0.0000	0.0833	0.4218	**0.4541**
UF1	0.2792	0.0775	0.1208	0.3420	**0.4545**

The performance of my method evaluated by AU independent 5-fold cross-validation of three experiments is summarised in Table 3.2. The results of the experiments conducted on CASME II & SAMM show little deterioration compared with those obtained by using only CASME II or SAMM data. A possible cause of the slight performance drop may lie in the fact that the SAMM database is more ethnically diverse – CASME II contains data from only one ethnic group, whereas SAMM includes 13 different ethnicities. It is also worth noting that the data acquisition protocols utilised for the collection of the two datasets are different, making the AU detection task on their composite harder than when no such confounding is present.

As for the subject independent 4-fold cross-validation in Table 3.3 and Table 3.4, only F1-score is applied for a fair comparison. The advantages of my approach in addressing the problem of unbalanced data are clearly demonstrated by this comparison. The results show that the proposed method's F1-score corresponding to each individual AUs lies between 0.4 and 0.6. This is in contrast with other methods, which exhibit dependency on the frequency of AU activation. For example, AU4 is the most commonly activated AU in CASME II, so my competitors' detection of other AUs is much worse than that of AU4. Through the implementation of sub-region segmentation methodologies, the adopted approach distinctly highlights a notable advantage: a superior capacity to effectively address imbalanced data. Specifically, the heightened activation of a specific AU predominantly exerts a discernible influence on AUs only situated within

Table 3.5: Cross-dataset robustness experiments: training on one dataset and testing on another, and vice versa.

	CASME II as test			SAMM as test		
AU	**Accuracy**	**F1**	**Score**	**Accuracy**	**F1**	**Score**
L1	0.9766	0.4941	0.7353	0.8980	0.4731	0.6856
R1	0.9766	0.4941	0.7353	0.9137	0.4775	0.6956
L2	0.8984	0.4733	0.6858	0.9216	0.4796	0.7006
R2	0.9375	0.4839	0.7107	0.9451	0.4859	0.7155
L4	0.1172	0.1049	0.1110	0.5176	0.3411	0.4294
R4	0.8281	0.6269	0.7275	0.5137	0.3394	0.4266
L5	0.9219	0.4797	0.7008	0.9804	0.4950	0.7377
R5	0.9219	0.4797	0.7008	0.9804	0.4950	0.7377
L6	0.9844	0.4961	0.7402	0.9608	0.4900	0.7254
R6	0.9609	0.4900	0.7255	0.9647	0.4910	0.7279
L7	0.7266	0.4208	0.5737	0.7804	0.4711	0.6258
R7	0.7344	0.4234	0.5789	0.2118	0.2110	0.2114
9	0.9610	0.4900	0.7255	0.9647	0.4911	0.7279
10	0.1875	0.1746	0.1811	0.9294	0.4817	0.7056
12	0.7344	0.4234	0.5789	0.8471	0.4586	0.6528
L14	0.6328	0.4558	0.5443	0.9255	0.4807	0.7031
R14	0.9609	0.4900	0.7255	0.9216	0.4796	0.7006
15	0.9766	0.4940	0.7353	0.9333	0.4828	0.7080
17	0.9609	0.4900	0.7255	0.8980	0.4731	0.6856
18	0.9688	0.4920	0.7304	0.9961	0.4990	0.7475
20	0.9609	0.4900	0.7255	0.9922	0.4980	0.7451
24	0.9609	0.4900	0.7255	0.9765	0.4940	0.7353
25	0.9531	0.4880	0.7206	0.9922	0.4980	0.7451
26	0.9609	0.4900	0.7255	0.9961	0.4990	0.7475
	UAR	**UF1**	**Score**	**UAR**	**UF1**	**Score**
Final	**0.8418**	**0.4556**	**0.6487**	**0.8734**	**0.4619**	**0.6676**

the corresponding region. Meanwhile, AUs positioned outside this delineated region remains resilient to such effects, undergoing parallel classification processes. In summary, my method comprehensively exhibits state-of-the-art performance, outperforming the otherwise leading methods in the literature.

In order to demonstrate the robustness of our method, we conducted experiments involving training the model on SAMM and testing it on CASME II, and vice versa. Surprisingly, the results from these cross-dataset experiments closely resembled those of the original experiments where the model was trained and tested on a single database. The experiment results can be seen in Table 3.5. This consistency in results across different datasets underlines the resilience and adaptability of our method. It suggests that our model's performance is not overly dependent on

idiosyncrasies or biases specific to any particular dataset. This robustness enhances our confidence in the generalisability and effectiveness of our approach in various real-world applications, regardless of the dataset employed.

3.4 Conclusion

In this chapter, I presented a novel method for detecting AU in micro-expressions through facial sub-region segmentation and multi-label classification. The proposed approach was empirically evaluated on two widely used micro-expression databases, namely CASME II and SAMM, where it demonstrated state-of-the-art performance.

The facial sub-region segmentation method relies on facial landmarks and the distribution positions of AU. The features of key facial areas are extracted, and the micro-expression AUs is separated into 11 key facial sub-regions to perform the multi-label classification. The novelty lies in dividing the numerous multi-labels into multiple smaller multi-label classifications, which jointly determine the final outcome for each label. The primary focus is on achieving AU multi-label classification by refining the facial sub-regions based on the location of the AUs.

Through AU independent 5-fold cross-validation and a comprehensive comparison with the leading methods in the literature using subject-independent 4-fold cross-validation, my approach successfully addresses the challenge of unbalanced data in micro-expression AU detection. It surpasses its competitors and achieves state-of-the-art results. The proposed method also opens a range of avenues for future research and further improvement. Amongst these, one of the most obvious ones is the optimisation of feature extraction and multi-label classification algorithms.

CHAPTER FOUR

SHORT AND LONG RANGE RELATION BASED SPATIO-TEMPORAL TRANSFORMER FOR MICRO-EXPRESSION RECOGNITION

Micro-expressions, known for their spontaneity, provide valuable insights into a person's latent emotions, even when attempts are made to conceal them. However, recognising micro-expressions poses a significant challenge in affective computing due to their brief duration and subtle intensity. Early studies relied on handcrafted spatio-temporal features, showing some potential. Nevertheless, recent advancements in deep learning techniques, surpassing these earlier approaches, now vie for state-of-the-art performance. However, effectively capturing both local and global spatio-temporal patterns remains an ongoing challenge. In this chapter, I propose a novel spatio-temporal transformer architecture, representing the first purely transformer-based approach for micro-expression recognition, devoid of any convolutional network usage. The architecture consists of a spatial encoder that learns spatial patterns, a temporal aggregator for temporal dimension analysis, and a classification head. Through a comprehensive evaluation of three widely-used spontaneous micro-expression datasets, namely,

SMIC-HS, CASME II, and SAMM, I demonstrate that the proposed approach consistently outperforms the state-of-the-art. Furthermore, the framework achieves an unweighted F1-score greater than 0.9 on each of the aforementioned datasets, marking a significant milestone in the published literature on micro-expression recognition. The source code is available at https://github.com/Vision-Intelligence-and-Robots-Group/SLSTT.

4.1 Introduction

Facial expressions are vital for interpersonal communication and recognising them is a significant task in affective computing. There is some debate about the universality of facial expressions, but many psychologists believe that emotions are expressed universally, regardless of cultural backgrounds [40]. While facial macro-expressions are consciously controlled and can be used to deceive, facial micro-expressions are involuntary and occur briefly due to psychological inhibition. These minute, sudden, and transient expressions provide a non-verbal means of articulating latent emotions, unaffected by conscious efforts. Accurately recognising facial micro-expressions is important for understanding people's mental states and emotions in general communication.

The initial methodologies for MER heavily relied on established computer vision techniques, employing handcrafted features and their variations [217, 143, 104, 113]. As the field progressed, a paradigm shift occurred towards leveraging deep learning methodologies, particularly CNNs [135, 85, 100, 199, 200]. In the early stages, convolutional kernels were employed to extract spatial information at a pixel level. However, this kind of pixel level operator can be considered as capturing "*short-range*", local spatial relationships. "*Long-range*", global relationships between different spatial regions have also been proposed and studied by graph convolutional networks (GCNs) based architectures [115, 13, 203, 90, 97]. In these innovative architectures, AUs are represented as nodes within graphs, facilitating the integration of AU engagements and image features to augment discriminatory capabilities for MER. However, though these approaches consider global spatial relations so as to assist learning, they can only learn these after local features are extracted, i.e. they are unable to learn both kinds of relations jointly.

In order to capture automatically both short- and long-range relations at the same time, I apply multi-head self-attention mechanism (MSM) instead of a convolutional kernel as the cornerstone of my deep learning MER architecture. As shown in Figure 4.1, the relations between block 1 and N will hardly ever be learnt by CNN but have been considered at the beginning of MSM. MSM based networks are called *Transformer*. Short-range and long-range relationships

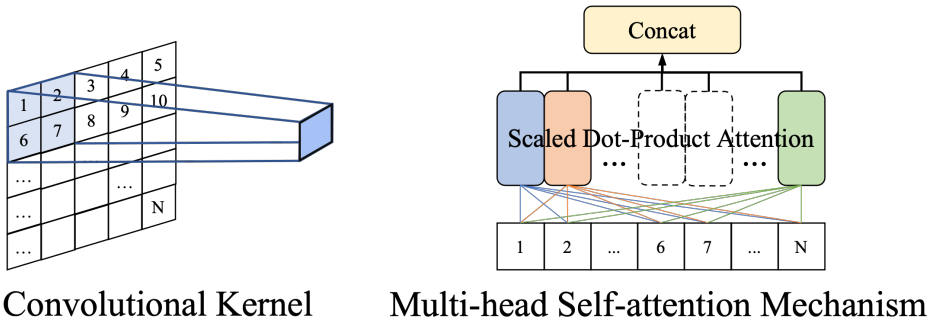


Figure 4.1: Comparison of the different spatial feature extraction methods of CNN and transformer.

between elements of a sequence can be learned in a parallelised manner because transformers utilise sequences in their entirety, as opposed to processing sequence elements sequentially like recurrent networks. Most recently, transformer networks came to the attention of the computer vision community. By dividing them into smaller constituent patches, two-dimensional images can be converted into one-dimensional sequences, translating the spatial relationships into the relationships between sequence elements (image patches). In this way, transformer networks can be simply applied to vision problems and on various tasks they have outperformed CNNs [84]. Examples include segmentation [210], image super-resolution [208], image recognition [35, 167], video understanding [160, 48] and object detection [15, 236].

Most MER research in the published literature is video-based, though there is a small but notable body of work on single-frame analysis [110, 46, 106]. This statistic reflects the consensus that for best performance both spatial and temporal information need to be considered. In particular, absolute and relative facial motions are extracted and analysed through spatial and temporal features respectively. Most handcrafted methods in existence use the same kind of operator to detect spatial and temporal information from different dimensions by considering the frames as 3D data. The resulting spatio-temporal features with uniform format are used together to implement video based MER. In deep learning based methods, spatial features are mainly extracted by means of a convolutional neural network. Some concatenate spatial features extracted from each frame and others use recurrent neural networks to derive temporal information. To integrate various spatio-temporal relations, my design makes use of long-term temporal information in spatial data (i.e. each frame of the video sample) prior to the spatial encoder, and a temporal aggregation block to fuse both short- and long-term temporal relationships afterwards.

In this work, I show how a transformer based deep learning architecture can be applied to MER in a manner which outperforms the current state of the art. The main contributions of the present

chapter are as follows:

1. I propose a novel spatio-temporal deep learning transformer framework for video based micro-expression recognition, which I name *short and long range relation based spatio-temporal transformer (SLSTT)*, the structure whereof is summarised in Figure 4.2. To the best of my knowledge, mine is the first deep learning MER work of this kind, in that it does not employ a CNN at any stage, but is rather entirely centred on a transformer architecture.

2. I use matrices of long-term optical flow, computed in a novel way particularly suited for MER, instead of the original colour images as the input to my network. The feature ultimately arrived at combines long-term temporal information and short- and long-range spatial relations, and is derived by a transformer encoder block.

3. I design a temporal aggregation block to connect spatio-temporal features of spatial relations extracted from each frame by multiple transformer encoder layers and achieve video based MER. The empirical performance and analysis of mean and LSTM aggregators is presented too.

I evaluate my approach on the three well-known and popular micro-expression databases, SMIC [103], CASME II [206] and SAMM [29], in both sole database evaluation (SDE) and composite database evaluation (CDE) settings and achieve state of the art results.

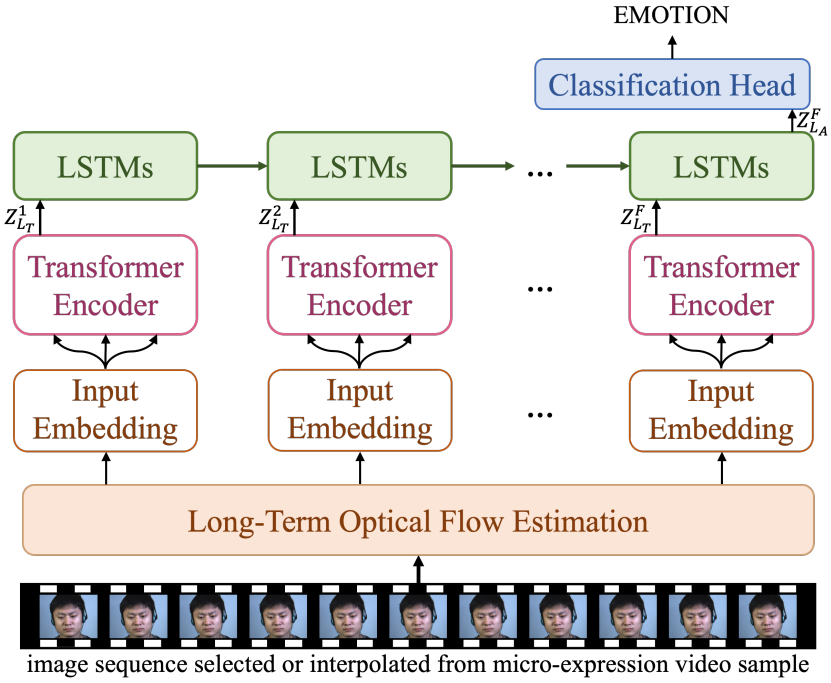


Figure 4.2: The framework of the proposed SLSTT.

4.2 Related work

4.2.1 Spatio-temporal feature extraction in micro-expression recognition

LBP quickly became the most popular operator for micro-expression analysis after Pfister et al. first applied it to MER [140]. This operator describes local appearance in an image. The key idea behind it is that the relative brightness of neighbouring pixels can be used to describe local appearance in a geometrically and photometrically robust manner. Its widespread use and favourable performance often make it the default baseline method when new data sets are published, or a new micro-expression related task proposed. As for deep learning approaches, CNN model can be thought of as a combination of two components: a feature extraction part and a classification part. The convolution and pooling layers perform spatial feature extraction.

Since one of the most characteristic aspects of micro-expressions is their sudden occurrence, temporal features cannot be ignored. While some methods in the literature do use only the single, apex frame instead of all frames in each micro-expression sample [138, 46, 110, 106], most employ all in the range between the onset frame and the offset, thus treating all temporal changes within this time period on the same footing. Some go further and employ temporal frame interpolation (as indeed I do herein) so as to increase the frame count [104, 113, 85, 185, 140].

A vast number of handcrafted feature based approaches treat raw video data as a 3D spatio-temporal volume, treating the temporal dimension as no different than the spatial ones. In other words, they apply the same kind of operator used to extract spatial features on pseudo-images formed by a cut through the 3D volume comprising one spatial dimension and the temporal dimension. For example, in LBP-TOP, LBP operators are applied on XT and YT planes to extract temporal features, and their histogram across the three dimensions forms the final representation. 3DHOG similarly treats videos as spatio-temporal cuboids with no distinction made between the three dimensions, but arguably with even greater uniformity than LBP-TOP in that the descriptor itself is inherently 3D based. Similar in this regard are optical flow based features, which too inherently combine local spatial and temporal elements – the use of optical strain [111], flow orientation [113] or its magnitude [110] are all variations on this theme.

As an alternative to the use of raw appearance imagery as input to a deep learning network, the use of pre-processed data in the form of optic flow matrices has been proposed by some authors [200, 114, 90]. In this manner, proximal temporal information is exploited directly. On the other hand, the learning of longer range temporal patterns has been approached in a variety of ways by different authors. Some extract temporal patterns simply by treating video sequences as 3-dimensional matrices [115, 111, 147], rather than 2-dimensional ones which naturally capture single images. Others employ structures such as the RNN or the LSTM [88, 85]. In addition to

the use of off-the-shelf recurrent deep learning strategies, recently there has been an emergence of methods which apply domain specific knowledge so as to make the learning particularly effective for micro-expression analysis [200].

4.2.2 Transformers in computer vision

For approximately a decade now, convolutional neural networks have established themselves as the backbone of most deep learning algorithms in computer vision. However, convolution always operates on fixed-size windows and is thus unable to extract distal relations. The idea of a transformer was first introduced in the context of natural language processing (NLP). It relies on a self-attention mechanism, learning the relationships between elements of a sequence. Transformers are able to capture 'long-term' dependence between sequence elements which is challenging for conventional recurrent models to encode. By dividing an image into sub-images and imposing a consistent ordering on them, a planar image can be converted into a sequence, so spatial dependencies can be learned in the same way as temporal features. For this reason, transformer based deep learning architectures have recently gained significant attention from the computer vision community and are starting to play an increasing role in a number of computer vision tasks.

A representative example in the context of object detection is DEtection TRansformer (DETR) [15] framework which uses transformer blocks first, for regression and classification, but the visual features are still extracted by a CNN based backbone. The image generative pre-training (iGPT) approach of Chen et al. attempts to exploit the strengths of transformers somewhat differently, pre-training bidirectional encoder representations from transformers (BERT) [32], originally proposed for language understanding, and thereafter fine-tuning the network with a small classification head. IGPT uses pixels instead language tokens within BERT, but suffers from significant information loss effected by a necessary image resolution reduction. In the context of classification, vision transformer (ViT) approach of Dosovitskiy et al. applies transformer encoding of image patches as a means of extracting visual features directly. It is the first pure vision transformer, and in its spirit and design, follows the original transformer [173] architecture faithfully. As such, it facilitates the application of scalable transformer architectures used in NLP effortlessly.

Following these successes, transformers have been applied to a variety of computer vision tasks, including those in the realm of affective computing [20, 194]. Notable examples include facial action unit detection [77] and facial image-based macro-expression recognition [119]. However, none of the existing approaches to micro-expression recognition adequately make use of both the spatial and temporal information due to the design difficulties posed by the challenges I

discussed in the previous sections.

4.3 Method details

In the present work, I propose a method that takes advantage both of the physiological understanding of micro-expressions and their characteristics, as well as of the transformer framework. The approach overcomes many of the weaknesses of the existing MER methods in the literature as discussed in the previous section. Importantly, my method is able to extract and thus benefit both from proximal (i.e. short-range) and distal (i.e. long-range) spatio-temporal features. Each element of the proposed framework is laid out in detail next, corresponds to each sub-section.

4.3.1 Long-term optical flow

Optical flow describes the apparent motion of brightness patterns between frames, caused by the relative movement of the content of a scene and the camera used to image it [141]. If the camera is static, optical flow can be used to infer both the direction and the magnitude of an imaged object's movement from the change in the appearance of pixels between frames [2].

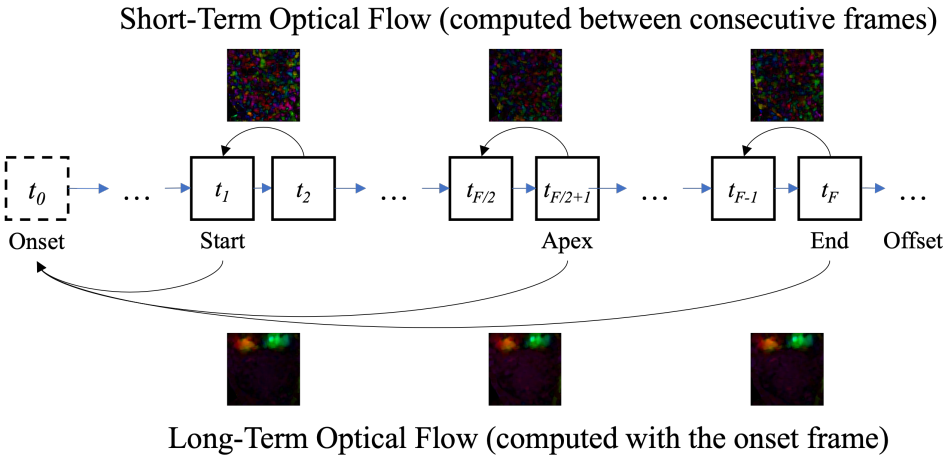


Figure 4.3: Different computing mechanism between short- and long-term optical flow.

Optical flow is inherently temporally local, i.e. save for practical considerations (numerical, efficiency, etc.) it is computed between consecutive frames of the sequence. This introduces a problem when micro-expression videos are considered, created by the already noted limited motion exhibited during the expressions. Therefore, herein I propose to calculate optical flow between each sample frame and the onset frame instead of consecutive frames, see Figure 4.3. To see the reasons behind this choice, consider Figure 4.4 which shows optical flow fields of

consecutive frames starting with the micro-expression onset frame. It can be readily observed that the fields are rather similar up to the apex frame, which can be attributed to the aforementioned brevity of the expression, with a similar trend thereafter but in the opposite direction. In contrast, my temporally non-local modified optical flow – long-term optical flow in a manner of speaking – exhibits a much more structured pattern, always being in the same direction, increasing in magnitude up to the apex frame and declining in magnitude thereafter. This results in much more stable and discriminative features associated with each micro-expression.

Let image intensity I be expressed as a function of space (x,y) and time t, i.e. as $I(x,y,t)$. I wish to relate the image intensity at the spatio-temporal locus $(x+\delta x, y+\delta y, t+\delta t)$ to that at (x,y,t), i.e. $I(x+\delta x, y+\delta y, t+\delta t)$ to $I(x,y,t)$.

Assuming that brightness of a point object is constant between frames and merely experiences image plane motion then using the Taylor series approximation leads to:

$$\begin{aligned} &I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) \\ &+ \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + (\textit{higher order terms}), \end{aligned} \tag{4.1}$$

and thus to:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \tag{4.2}$$

where:

$$u = \frac{\delta x}{\delta t}, v = \frac{\delta y}{\delta t}, \tag{4.3}$$

and $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ are the derivatives of the spatio-temporal image cuboid in the corresponding spatio-temporal directions. Written more succinctly as I_x, I_y and I_t, the optical flow equation can be re-written in the more common form as:

$$I_x u + I_y v = -I_t. \tag{4.4}$$

The equation cannot be solved directly since it is insufficiently constrained. Therefore, further assumptions and constraints are needed. The approach proposed by Lucas et al. remains one of the most widely used ones. It assumes that the displacement of the images is small and approximately constant within a neighbourhood of the point p under consideration. In this way, the optical flow equation can be assumed to hold for all pixels within a window centred at p. The

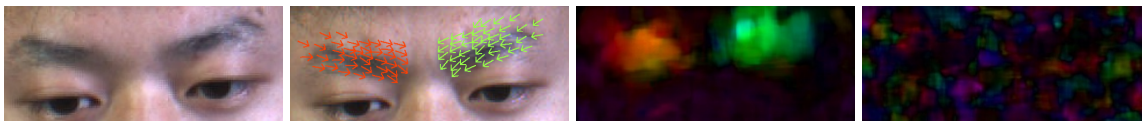
local image flow vector (u, v) must satisfy the following:

$$\begin{aligned} I_x(p_1)u + I_y(p_1)v &= -I_t(p_1) \\ I_x(p_2)u + I_y(p_2)v &= -I_t(p_2) \\ &\vdots \\ I_x(p_n)u + I_y(p_n)v &= -I_t(p_n), \end{aligned} \tag{4.5}$$

where p_1, p_2 and p_n are the pixels within the window. These equations can be written compactly in matrix form:

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_n) & I_y(p_n) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -I_t(p_1) \\ -I_t(p_2) \\ \vdots \\ -I_t(p_n) \end{bmatrix}, \tag{4.6}$$

which is, in general, an over-determined and inconsistent system. Thus, usually a least-squares fit is performed.



(a) Onset frame, magnified region of interest

(b) Apex frame with superimposed long-term optical flow samples

(c) Magnitude of long-term optical flow between onset & apex frames

(d) Magnitude of short-term optical flow between apex frame and its previous one

Figure 4.4: Illustration of optic flow computed between the onset and the apex frame, corresponding to the motion effected by the activation unit Brow Lowerer (AU4). Compare with the one computed between consecutive frames.

4.3.2 Spatial feature extraction

The key idea underlying the proposed method lies in the extraction of long-range spatial relations from each frame using a transformer encoder, with images as before being treated as sequences of constituent patches. More specifically, input frames are first represented as vector sequences with local spatial features of each image patch. The resulting sequences are then fed into the transformer encoder for long-term spatial feature extraction.

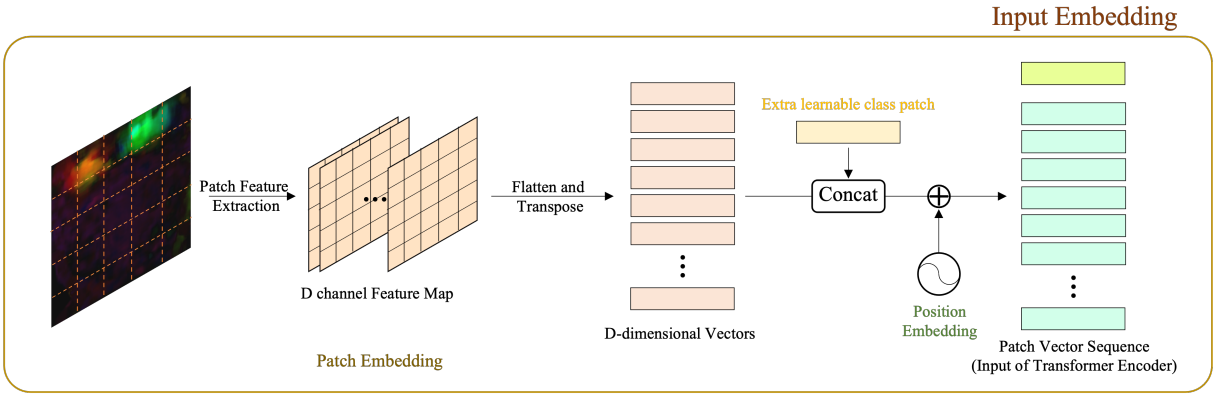


Figure 4.5: Long-term optical flow fields are as inputs of the Input Embedding blocks. After short-range spatial feature extraction, patch and position embedding, the resulting sequence of vectors is fed to standard transformer encoder layers.

4.3.2.1 Input embedding and short-range spatial relation learning

The standard transformer receives a 1D sequence as input. To handle 2D images, I represent each image as a sequence of rasterised 2D patches. Herein I do not use appearance images, that is the original video sequence frames, as input but rather the corresponding optical flow fields. An input embedding block is proposed as a means of representing input images as vector sequences for input to the transformer encoder.

The general input embedding mechanism considers the image $X \in \mathbb{R}^{H\times W\times C}$ as a sequence of non-overlapping $P\times P$ pixel patches, where H, W, and C are respectively the height, the width, and the channel count of the input. Different from the "separate and flat" linear patch embedding proposed by Dosovitskiy et al., I first extract local spatial features in patch regions with a patch-wise fully connected layer. Patches of image X are represented as $X_p \in \mathbb{R}^{N\times(P^2,C)}$. As shown in Figure 4.5, I extract the short-range spatial features from image X to feature map $X \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$, flatten and transpose them to N D-dimensional vectors, where $N=\frac{HW}{P^2}$ the resulting number of patches in each image. D-dimensional vectors are passed through all transformer encoder layers. The specific values of parameters used in my experiments are stated in Section 4.4.

After that, a learnable D-dimensional vector is concatenated with the sequence, as the class token ($Z_0[0]=x_{class}$), whose state as the output of the transformer encoder ($Z_{L_T}[0]$). The effective input sequence length for the transformer encoder is thus $N+1$. Then a position embedding is added to each vector in the sequence. The whole input embedding procedure can be described as

follows:

$$Z_0 = [X_{class}; X_p^1 E; X_p^2 E; \ldots; X_p^N E] + E_{pos},$$
$$E \in \mathbb{R}^{(P^2,C)\times D}, E_{pos} \in \mathbb{R}^{(N+1)\times D}, \tag{4.7}$$

where $Z_0 \in \mathbb{R}^{(N\times D)}$ is the input of the transformer encoder.

4.3.2.2 Long-range spatial relation learning by transformer encoder

After short-range spatial relations are extracted from the input long-term optical flow fields of each frame and embedded as vectors, they are passed to a transformer encoder for further long-range spatial feature extraction. My encoder contains L_T transformer layers; herein I use $L_T = 12$, adopting this value from the ViT-Base model of Dosovitskiy et al.'s (the pre-trained encoder I use in experiments). Each layer involves two blocks, a MSM and a Position-Wise fully connected Feed-Forward network (PWFF), as shown in Figure 4.6. Layer Normalisation (LN) is applied before each block and residual connections after each block [181, 5]. The output of the transformer layer can be written as follows:

$$Z'_l = MSM(LN(Z_{l-1})) + Z_{l-1}, l = 1 \ldots L_T, \tag{4.8}$$
$$Z_l = \text{PWFF}(LN(Z'_l)) + Z'_l, l = 1 \ldots L_T, \tag{4.9}$$

where Z_l is the output of layer l. The PWFF block contains two layers with the Gaussian Error Linear Unit (GELU) non-linear activation function. The feature embedding dimension thereby first increases from D to $4D$ and then reduces back to D, which equals 768 in my experiments.

Multi-head attention allows the model to focus simultaneously on information content from different parts of the sequences, so both long-range and short-range spatial relations can be learnt. An attention function is mapping a query and a set of key-value pairs to the output, a weighted sum of the values. The weights are computed using a compatibility function of the queries with the corresponding keys, and they are all vectors. The self-attention function is computed on a set of queries simultaneously. The queries, keys and values can be grouped together and represented as matrices Q, K and V, so the computation of the matrix of outputs can be written as:

$$Q = Z_{l-1} W_Q, \tag{4.10}$$
$$K = Z_{l-1} W_K, \tag{4.11}$$
$$V = Z_{l-1} W_V, \tag{4.12}$$

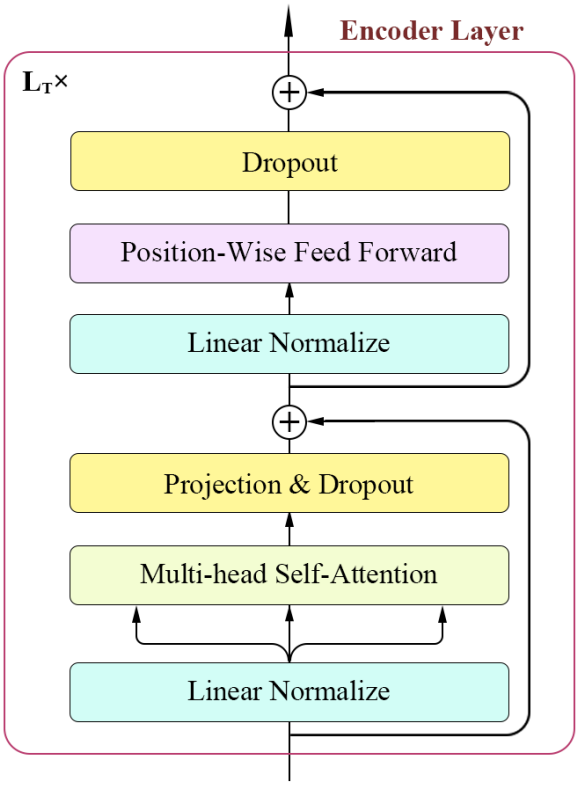


Figure 4.6: Detailed structure of a Transformer Encoder layer. The output of frame t processed by spatial encoder is $Z^t_{L_T}$.

$$SA(Z_l) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V, \tag{4.13}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D\times D_m}$ are learnable matrices and SA is the self-attention module. MSM can be seen as a type of self-attention with M heads in parallel operation and a projection of their concatenated outputs:

$$MSM(Z_l) = \text{Concat}(\{SA_h(Z_l), \forall h \in [1..M]\})W_O, \tag{4.14}$$

where $W_O \in \mathbb{R}^{M\cdot D_m\times D}$ is a re-projection matrix. D_m is typically set to $\frac{D}{M}$, so as to keep the number of parameters constant with changing M.

4.3.3 Temporal aggregation

After extracting both local and global spatial features associated with each frame using a transformer encoder, I introduce an aggregation block to extract temporal features before performing the ultimate classification. The aggregation function ensures that my transformer model can be trained and applied to the spatial feature sets of each frame, subsequently processing the temporal

relations between frames in each sample. Since facial movement during micro-expressions is almost imperceptible, all frames from a single video sample are rather similar one to another. Nevertheless, it is still possible to identify reliably a number of salient frames, such as the apex frame, that play a particularly important role in the analysis of a micro-expression. Therefore, I propose an LSTM architecture for temporal aggregation.

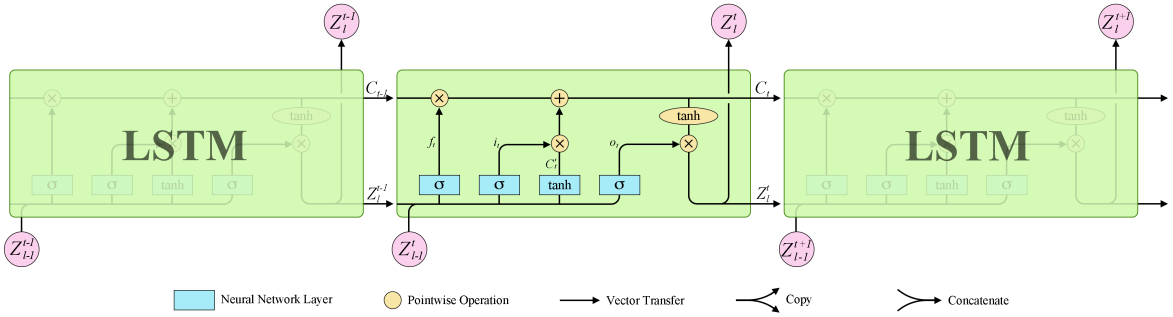


Figure 4.7: The repeating module in an LSTM aggregator layer.

LSTM [60] is a type of recurrent neural network with feedback connections, which overcomes two well-known problems associated with RNNs: the vanishing gradient problem, and the sensitivity to the variation of the temporal gap length between salient events in a processed sequence. The elements of the input are the sets of outputs from the transformer encoder for each frame. The inputs are not concatenated, and the input sequence length is thus dependent on the number of frames in each micro-expression video sample.

I used three LSTM layers in the aggregation block. The computation details of each layer are:

$$t = 1 \dots F, l = L_T + 1 \dots L_A,$$

$$f_t = \sigma(W_f \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_f), \tag{4.15}$$

$$i_t = \sigma(W_i \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_i), \tag{4.16}$$

$$o_t = \sigma(W_o \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_o), \tag{4.17}$$

$$C_t' = tanh(W_C \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_C), \tag{4.18}$$

$$C_t = f_t \times C_{t-1} + i_t \times C_t', \tag{4.19}$$

$$Z_l^t = o_t \times tanh(C_t), \tag{4.20}$$

where F is the number of chosen frames in each video sample, L_A is the total number of layers in both the transformer encoder and the LSTM aggregator. Z_l^t denotes the outputs of the layer l

after t frames have been processed. After all frames are processed in this manner, the result is a single feature set describing the entire micro-expression video sample. Finally, these features are fed into an multilayer perceptron (MLP) which is used for the ultimate MER classification. The details of how the previous output join the latter training are presented in Figure 4.7.

4.3.4 Network optimisation

Following the aggregation block, my network contains two fully connected layers which facilitate the final classification achieved using the SoftMax activation function. Cross Entropy loss is used as the objective function for training:

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{C} y_{ic}\log(p_{ic}), \tag{4.21}$$

where N is the number of the micro-expression video samples and C is the number of emotion classes. The value of y_{ic} is 1 when the true class of sample i is equal to c and 0 otherwise. Similarly, p_{ic} is the predicted probability that sample i belongs to class c.

When using gradient descent to optimise the objective function during network training, as the parameter set gets closer to its optimum, the learning rate should be reduced. Herein I achieve this using cosine annealing [116], i.e. using the cosine function to modulate the learning rate which initially decreases slowly, and then rather rapidly before stabilising again. This learning rate adjustment is particularly important in the context of the problem at hand, considering that the number of available micro-expression video samples is not large even in the largest corpora, readily learning to overfit if due care is not taken.

4.4 Empirical evaluation

In this section, I describe the empirical experiments used to evaluate the proposed method. I begin with a description of the data sets used, follow up with details on the data pre-processing performed, relevant implementation details, and evaluation metrics, and conclude with a report of the results and a discussion of the findings.

4.4.1 Data pre-processing

Following the best practices in the field, for my evaluation I adopt the use of three large data sets, namely SMIC-HS [103], CASME II [206], and SAMM [29], thus ensuring sufficient diversity of data, evaluation scale, and ready and fair comparison with other methods in the

literature. All video samples in these databases capture spontaneously exhibited, rather than acted micro-expressions, which is important for establishing the real-world applicability of findings.

4.4.1.1 Face cropping

As noted in the previous section, cropped face images are explicitly provided in both SMIC-HS and CASME II data sets, with the same registration method used in both; no cropped faces are provided as part of SAMM. In order to maintain data consistency across different databases, in my experiments I employ a different face extraction approach. In particular, I utilise ERT [83] algorithm implemented in DLib [89] to localise salient facial loci (68 of them) in a uniform manner regardless of which data set a specific video sample came from.

In the case of SMIC-HS and CASME II videos, the original authors' face extraction process consists of facial landmarks detection in the first frame of a micro-expression clip and then the detected face being registered to the model face using a local weighted mean transformation. Motivated by the short duration of micro-expressions, the faces in all remaining frames of the video sample are registered using the same matrix. However, in this work, I employ an alternative strategy. The primary reason lies in the need for sufficient and representative data diversity, which is particularly important in deep learning. In particular, the original face extraction method just described, often results in the close resemblance of samples which increases the risk of model overfitting. Therefore, herein I instead simply use a non-reflective 2D Euclidean transformation, i.e. one comprising only rotation and translation. By doing so, at the same time, I ensure the correct alignment of salient facial points and maintain information containing facial contour variability.

Furthermore, unlike the authors of SMIC-HS and CASME II, I do not perform facial landmark detection in the first frame of a micro-expression sample, but rather in the apex, thereby increasing the registration accuracy of the most informative parts of the video. As shown in Figure 4.8, points 27–30 can be used to determine the centre line of the nose that can be considered as the vertical symmetry line of the entire face area. Point 30 is set as the centre point, and the square size s (in pixels) is computed by adding the vertical distance from the centre point of the eyebrows (19) to the lowest point of the chin (8), $y_{apex[8]} - y_{apex[19]}$, to the height of chin, $y_{apex[8]} - y_{apex[57]}$, so that nearly the entire face is included in the cropped image:

$$s = (y_{apex[8]} - y_{apex[19]}) + (y_{apex[8]} - y_{apex[57]}). \tag{4.22}$$

Figure 4.8: The 68 facial landmarks used by my method, are shown for the location (green) and labelled number (red).

4.4.1.2 Temporal interpolation

Considering the short duration of micro-expressions, even when samples are acquired using high-speed cameras, in some instances only a small number (cc. 10) of frames is available. In an attempt to extract accurate temporal information, I also apply frame interpolation from raw videos, effectively synthetically augmenting data. In previous work, the Temporal Interpolation Model (TIM) relies on a path graph to characterise the structure of a sequence of frames, popularly used in several handcrafted feature based methods [100, 188, 104], whereas Liu et al. use simple linear interpolation. Herein I propose a novel approach to interpolation so that its result is smoother in terms of optical flow, it being the nexus of my entire MER methodology. Most existing optical flow based methods produce artifacts on motion boundaries by estimating bidirectional optical flows, scaling and reversing them to approximate intermediate flows. I adopt the real-time intermediate flow estimation (RIFE) method [231], which uses an end-to-end trainable neural network, IFNet, which speedily and directly estimates the intermediate flows.

Original RIFE interpolates one frame between two given consecutive frames, so I apply it recursively to interpolate multiple intermediate frames. Specifically, given any two consecutive input frames I_0, I_1, I apply RIFE once to get intermediate frame $\hat{I}_{0.5}$ at $t = 0.5$. I then apply RIFE to interpolate between I_0 and $\hat{I}_{0.5}$ to get $\hat{I}_{0.25}$, and so on. In my experiment, I prioritise interpolation in the temporal vicinity of the apex frame. The interpolated queue can be expressed as $\{\hat{I}_{a-0.5}, \hat{I}_{a+0.5}, \hat{I}_{a-1.5}, \hat{I}_{a+1.5}, \ldots, \hat{I}_{o+0.5} \text{ or } \hat{I}_{f-0.5}\}$, where a, o and f are frame indices of the apex, onset, and offset frames respectively. Recall that the apex frames are specified explicitly in CASME II and SAMM, and for SMIC-HS I choose the middle frame of each sample video as the apex. If the number of interpolation frames is lower than the reference count (the average

number of frames in this period across the database), I use the same method on the updated frame sequence iteratively to generate further intermediate frames.

4.4.2 Experimental settings

4.4.2.1 Implementation details

In the spatial feature extraction procedure, I employed base ViT blocks, with 12 Encoder layers, hidden size of 768, MLP size of 3072, and 12 heads. For initialisation, I use the official ViT-B/16 model [35] pre-trained on ImageNet [31]. I resize my input images to 384×384 pixels and split each image into patches with 16×16 pixels, so that the number of patches is 24×24. 768-dimensional vectors are passed through all transformer encoder layers. For temporal aggregation, I select 11 frames (apex, and five preceding and succeeding it) per sample as inputs for the mean aggregator and LSTM aggregator. I have tried other options with different numbers of frames, but it didn't work any better. I only use long-term optical flow in experiments, as motivated by the arguments discussed in Section 4.3.1. For learning parameters, the initial learning rate and weight decay are set to be 1e-3 and 1e-4, respectively. The momentum for Stochastic Gradient Decent (SGD) is set to 0.9, with the batch size 4 for all experiments. All the experiments were conducted with PyTorch.

4.4.2.2 Mean versus LSTM aggregator

I compare LSTM aggregator with an alternative which uses the simple mean operator for temporal aggregation. After each frame is processed by the spatial encoder, the corresponding output is used in the computation by the mean aggregation layer (layer $L_T + 1$):

$$Z^t_{L_T+1} = \frac{t-1}{t} Z^{t-1}_{L_T+1} + \frac{1}{t} Z^t_{L_T}, t = 1 \ldots F, \tag{4.23}$$

In a manner similar to that described previously in the context of the LSTM aggregator, outputs of each frame from my transformer encoder are taken as inputs to the temporal feature extraction module. Compared to the mean operator, LSTM has the advantage of larger expressive capability, resulting in different extracted relationships between different frames. Within the specific context of my work, this means that its ability to distinguish between emotions is also different, with LSTM expected to perform better.

4.4.2.3 Evaluation metrics

Following previous work and micro-expressions grand challenges (MEGCs), I conducted experiments on SMIC-HS, CASME II, and SAMM, evaluating the classification performance using the

corresponding original emotion classes, as well as the composite corpus formed using all three data sets and relabelled using three classes as proposed in MEGC 2019 [152]. All results are reported using leave one subject out (LOSO) cross-validation. Evaluation is repeated multiple times by holding out test samples of each subject group while the remaining samples are used for training. In this way, I best mimic real-world situations and in particular assess the robustness to variability in ethnicity, gender, emotional sensitivity, etc.

Sole database evaluation (SDE) In the first part of my empirical evaluation, experiments are conducted on three databases individually, using the corresponding original emotion labels, excepting the very rare (and thus underrepresented) classes in CASME II and SAMM. SMIC-HS uses 3 class labels whereas the other two sets both use 5. I use *accuracy* and *macro F1-score* to assess the recognition performance.:

$$Accuracy = \frac{\sum_{c=1}^{C}\sum_{i=1}^{S} TP_{i,c}}{N}, \tag{4.24}$$

$$Precision_c = \frac{\sum_{i=1}^{S} TP_{i,c}}{\sum_{i=1}^{S} TP_{i,c} + \sum_{i=1}^{S} FP_{i,c}}, \tag{4.25}$$

$$Recall_c = \frac{\sum_{i=1}^{S} TP_{i,c}}{\sum_{i=1}^{S} TP_{i,c} + \sum_{i=1}^{S} FN_{i,c}}, \tag{4.26}$$

$$F1\text{-}score_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c}, \tag{4.27}$$

$$macro\ F1\text{-}score = \frac{\sum_{c=1}^{C} F1\text{-}score_c}{C}, \tag{4.28}$$

where $TP_{i,c}$, $FP_{i,c}$ and $FN_{i,c}$ are true positive, false positive, and false negative rates for each class c (of C classes), with samples of subject i as test. S is the number of subjects in each database, and N is the total number of samples from all subjects.

Composite database evaluation (CDE) In the second part of my empirical evaluation, experiments are conducted on the composite database with 3 emotion classes (negative, positive, and surprise). The composite database, that is the database obtained by merging SMIC, CASME II, and SAMM contains a total of 68 subjects, 16 from SMIC, 24 from CASME II and 28 from SAMM. LOSO cross-validation is applied on each database separately and together on the composite database. UF1, also known as the *macro F1-score*, and UAR are used to assess performance. I have previously demonstrated the methodology for computing them in Section 3.3.2.

Table 4.1: SDE results comparison with LOSO on SMIC-HS (3 classes), CASME II (5 classes) and SAMM (5 classes). Best performances are shown in bold, second best by square brackets enclosure. (* Reported by Huang et al. [65], ** Reported by Khor et al. [86])

	SMIC-HS		CASME II		SAMM	
	Acc(%)	**F1**	**Acc(%)**	**F1**	**Acc(%)**	**F1**
Handcrafted						
LBP-TOP*	53.66	0.538	46.46	0.424	–	–
LBP-SIP*	44.51	0.449	46.56	0.448	–	–
STLBP-IP [64] (2015)	57.93	–	59.51	–	–	–
STCLQP [65] (2015)	64.02	0.638	58.39	0.584	–	–
Hierarchical STLBP-IP [238] (2018)	60.37	0.613	–	–	–	–
HIGO+Mag [104] (2018)	68.29	–	67.21	–	–	–
Deep Learning						
AlexNet**	59.76	0.601	62.96	0.668	52.94	0.426
DSSN [86] (2019)	63.41	0.646	70.78	0.730	57.35	0.464
AU-GACN [203] (2020)	–	–	49.20	0.273	48.90	0.310
MER-GCN [115] (2020)	–	–	42.71	–	–	–
Micro-attention [176] (2020)	49.40	0.496	65.90	0.539	48.50	0.402
Dynamic [159] (2020)	**76.06**	0.710	72.61	0.670	–	–
GEME [125] (2021)	64.63	0.616	[75.20]	[0.735]	55.88	0.454
SLSTT-Mean (Ours)	73.17	[0.719]	73.79	0.723	[66.42]	[0.547]
SLSTT-LSTM (Ours)	[75.00]	**0.740**	**75.81**	**0.753**	**72.39**	**0.640**

Table 4.2: CDE results comparison with LOSO on SMIC-HS, CASME II, SAMM and composite database (3 classes). Best performances are shown in bold, second best by square brackets enclosure. (*Reported by See et al. [152], **Reported by Xia et al. [201])

	Composite		SMIC-HS		CASME II		SAMM	
	UF1	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**
Handcrafted								
LBP-TOP*	0.588	0.579	0.200	0.528	0.703	0.743	0.395	0.410
Bi-WOOF*	0.630	0.623	0.573	0.583	0.781	0.803	0.521	0.514
Deep learning								
ResNet18**	0.589	0.563	0.461	0.433	0.625	0.614	0.476	0.436
DenseNet121**	0.425	0.341	0.460	0.333	0.291	0.352	0.565	0.337
Inception V3**	0.516	0.504	0.411	0.401	0.589	0.562	0.414	0.404
WideResNet28-2**	0.505	0.513	0.410	0.401	0.559	0.569	0.410	0.404
OFF-ApexNet* [46] (2019)	0.720	0.710	0.682	0.670	0.876	0.868	0.541	0.539
CapsuleNet [172] (2019)	0.652	0.651	0.582	0.588	0.707	0.701	0.621	0.599
Dual-Inception [233] (2019)	0.732	0.728	0.665	0.673	0.862	0.856	0.587	0.566
STSTNet [111] (2019)	0.735	0.761	0.680	0.701	0.838	0.869	0.659	0.681
EMR [114] (2019)	0.789	0.782	**0.746**	**0.753**	0.829	0.821	**0.775**	[0.715]
ATNet [138] (2019)	0.631	0.613	0.553	0.543	0.798	0.775	0.496	0.482
RCN [201] (2020)	0.705	0.716	0.598	0.599	0.809	0.856	0.677	0.698
AUGCN+AUFsuion [97] (2021)	[0.791]	**0.793**	0.719	[0.722]	[0.880]	[0.871]	[0.775]	**0.789**
SLSTT-Mean (Ours)	0.788	0.767	0.719	0.699	0.844	0.830	0.625	0.566
SLSTT-LSTM (Ours)	**0.816**	[0.790]	[0.740]	0.720	**0.901**	**0.885**	0.715	0.643

4.4.3 Results and discussion

I compare the performance of the proposed approach with baseline handcrafted feature extraction methods and the most prominent recent deep learning based methods on the widely used micro-expression databases, SMIC-HS, CASME II, and SAMM, described in the previous section, both in the SDE and the CDE settings. To ensure uniformity and fairness of the comparison, the SDE results for all methods were obtained in identical conditions, i.e. for the identical number of samples, the number of labels (classes), and using the same cross-validation approach. The details of the performance of our *SLSTT* on different emotion categories are shown in Figure 4.9.

As can be readily seen in Table 4.1 which presents a comprehensive overview of my experimental results in the SDE setting, the method proposed in the present chapter performs best (n.b. shown in bold) in all but one testing scenario, in which it is second best (n.b. second best performance is denoted by square brackets), trailing marginally behind the method introduced by Sun et al. [159]. What is more, in most cases my method outperforms rivals by a significant margin.

Moving next to the results of my experiments in the CDE setting, these are summarised in Table 4.2. It can be readily seen that my method's performance is again shown to be excellent. In particular, in most cases, my method again comes out either at the top or second best (as before the former being shown in bold and the latter denoted by square brackets enclosure). The only existing method in the literature which remains competitive against ours is that of Lei et al.'s [97]. To elaborate in further detail, my approach achieved the best results both in terms of UF1 and UAR on CASME II, and on UF1 on the full composite database, and second best on UAR on the composite database and on UF1 on SMIC-HS. The performance of all methods on CASME II is consistently higher than when applied to other data sets, which suggests that the challenge of MER is increased with the ethnic diversity of participants – this should be born in mind in future research and any comparative analysis. It is insightful to observe that in contrast with the results in the SDE setting already discussed (see Table 4.1), my method does not come out as dominant in the context of CDE. This suggests an important conclusion, namely that my method is particularly capable of nuanced learning over finer-grained classes and that its superiority is less able to come through in a simpler setting when only 3 emotional classes are used.

Taking into account the results from both the sole and the composite database experiments, it is useful to observe that when only short-range patterns are utilised, convolutional neural network approaches do not outperform methods based on handcrafted features. It is the inclusion of long-range spatial learning that is key, as shown by the marked improvement in the performance of the corresponding methods. Yet, the proposed method exceeds even their performance, owing to its use of MSM, thus demonstrating its importance in MER. The superiority of our short-

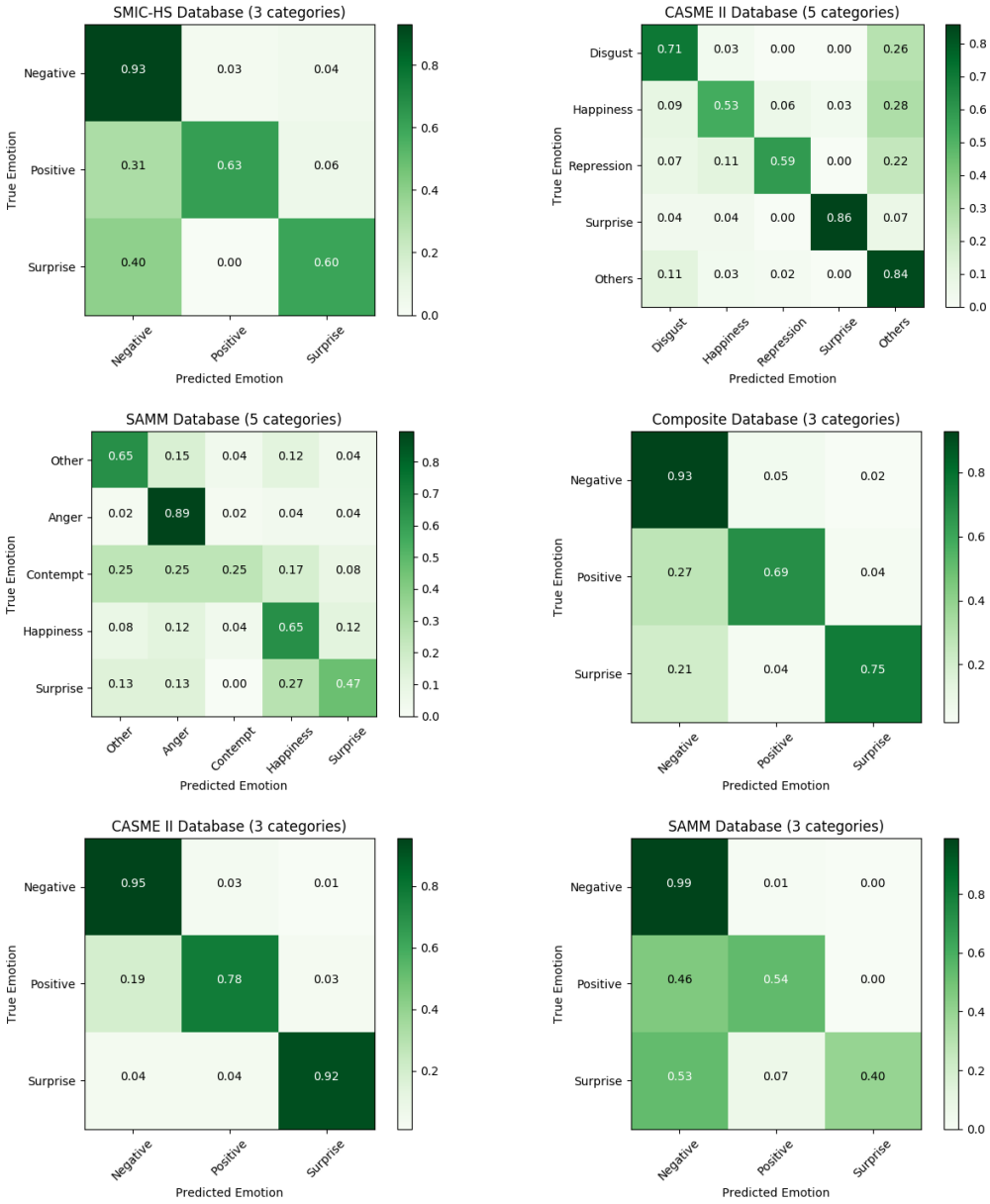


Figure 4.9: Confusion matrices corresponding to each of our experiments. Only one is shown for SMIC-HS because the SDE and the CDE are identical when this database is used alone.

and long-range relation based spatio-temporal transformer is further corroborated by the results shown in the latest two rows in both Table 4.1 and Table 4.2 which summarise our comparison of the proposed LSTM aggregator with the simpler mean operator aggregator.

From Figure 4.9, we could see in CASME II, that distinguishing whether a micro-expression is Disgust or Others is inherently difficult because the database contains multiple inconsistently labelled samples with only AU4 activated – some of them are labelled as Others, some as Disgust. It is also worth noting that in SAMM, some AU labels ('AU12 or 14') for the Contempt class were not manually verified, which also causes confusion with the Happiness class (mostly with AU12 labelled). In part, these labelling issues emerge from the fact that the mapping between facial AU activation and emotions (as understood by psychologists) is not a bijection. It is also the case that imperfect information is made use of because only visual data is used. Hence, it should be understood that the theoretical highest accuracy of automated micro-expression recognition on the MER corpora currently used for research purposes is not 100%. The micro-expression databases containing multi-modal signals [105, 101], which have begun emerging recently, seem promising in overcoming some of the limitations of the existing corpora.

In addition, I recently discovered that another study [61] proposed using transformer backbones for MER at around the same time as chapter 4. While my focus was on using the backbone for spatial feature extraction and designing a separate temporal aggregation block, they used video transformer backbones to process entire video clips. They also incorporated optical flow in the structure, but as an additional modality to learn micro-expression motion. In contrast, I implemented long-term optical flow and used it as the input for training. Despite the differences in our architectures, both studies demonstrate the potential of transformer backbones in this field.

4.5 Summary and conclusions

In this chapter, I proposed a novel transformer based spatio-temporal deep learning framework for micro-expression recognition, which is the first deep learning work in the field entirely void of convolutional neural network use. In my framework, both short- and long-term relations between pixels in spatial and temporal directions of the sample videos can be learned. I use transformer encoder layers with multi-head self-attention mechanism to learn spatial relations from visualised long-term optical flow frames and design a temporal aggregation block for temporal relations. Extensive experimental results using three large MER databases, both in the context of sole database evaluation and composite database evaluation settings and LOSO cross-validation protocol, consistently demonstrate that my approach is effective and outperforms the current state of the art. These findings strongly motivate further research on the use of transformer based

architectures rather than convolutional neural networks in micro-expression analysis, and I hope that my theoretical contributions will help direct such future efforts.

CHAPTER FIVE

MULTIMODAL LATENT EMOTION RECOGNITION FROM MICRO-EXPRESSION AND PHYSIOLOGICAL SIGNALS

In this chapter, in-depth exploration is undertaken to highlight the myriad advantages associated with leveraging multimodal data to enhance the accuracy of latent emotion recognition, specifically focusing on the integration of micro-expressions and physiological signals. To achieve this, a novel multimodal learning framework is proposed, which combines both input sources. This innovative approach encompasses a 1D separable and mixable depthwise inception network, a standardised normal distribution weighted feature fusion method, and depth/physiology guided attention modules for multimodal learning. Through rigorous experimentation, the results demonstrate the superior performance of the proposed approach compared to the benchmark method. Notably, the weighted fusion method and guided attention modules play indispensable roles in significantly elevating the overall performance of the system. By showcasing the effectiveness of this multimodal learning framework, this chapter contributes to the advancement of latent emotion recognition, paving the way for future research and applications in this domain.

5.1 Background

Emotional states have a significant impact on physical and psychological well-being, with recognition of emotions being crucial for effective communication and understanding of individuals' emotional states and mental well-being. The complex interplay between physiology and psychology in emotional responses has led to interdisciplinary research into accurate and rapid emotion recognition, which is increasingly important in multimedia and human-computer interaction. Real-time emotion recognition has potential applications in virtual and augmented reality, healthcare, education, and marketing. In interpersonal communication, facial expressions are a critical means of conveying emotions, with micro-expressions offering valuable insights into an individual's emotional state, including potential deception. Recognising micro-expressions enables experts to identify even the most subtle changes in an individual's facial expressions, potentially indicating their latent emotion. While facial expressions are the most reliable and universally accepted way of recognising emotions, vocal cues, body language, and physiological responses can also provide valuable information about a person's emotional state.

Enhancing emotion recognition accuracy entails exploring avenues beyond just improving the machine learning model, considering richer data types can also help achieve better performance. Human experience of the world is often multimodal, referring to how something happens or is experienced through multiple modalities. Incorporating multimodal signals can enable artificial intelligence to learn about the real world better. Relying solely on human physical signals, such as facial expression, speech, gesture, or posture, is not guaranteed as people can control these signals to hide their real emotions, especially during social communication. In contrast, *physiological signals*, which are in response to the central nervous system (CNS) and peripheral nervous system (PNS) of the human body, can provide reliable information about emotions. One significant advantage of using physiological signals is that they are largely involuntarily activated and, therefore, difficult to control, which is a similar characteristic to micro-expression. Researchers have attempted to establish standard relationships between emotional changes and various types of physiological signals.

Several studies have attempted to combine facial expressions with physiological data to create accurate emotion recognition systems [25, 67]. Both studies used late fusion techniques, such as voting and decision tree, to combine the decisions made from facial expressions and physiological signals, though Cimtay et al. use early fusion for different signals. However, people may conceal their true emotions behind fake facial expressions, whereas micro-expressions can reveal their genuine emotions, as can physiological signals. Therefore, combining micro-expressions and physiological signals through multimodal learning could be a better strategy for authentic emotion recognition. In previous research, colour images or videos have been the primary data

samples used due to the availability of micro-expression databases. However, the recent release of 4DME [105] and CAS(ME)^3 [101] has expanded the range of data available for emotion recognition related to micro-expressions. 4DME primarily focuses on 3D micro-expression data, while CAS(ME)^3 not only includes RGB-D micro-expression video clips but also includes physiological signals in one part of the database. This enables multimodal learning with micro-expressions for emotion recognition. Li et al. attempted to use voice, electrodermal activity (EDA) and depth information to assist MER. They converted the voice and EDA signals into 2D greyscale input channels and trained them with colour and depth information from the apex frame [101]. However, their results of combining EDA or speech signals were not satisfactory due to not addressing the noise in the signals or designing a specific network for physiological signals. Despite this, the database they provided is a valuable resource for researchers to optimise multimodal emotion recognition processing with micro-expressions. This chapter explores the benefits of incorporating multimodal data for improving latent emotion recognition accuracy, specifically with micro-expression and physiological signals. The main contributions of the present work are as follows:

1. I introduce a novel multimodal learning framework that combines micro-expression and physiological signals to enhance latent emotion recognition performance.

2. I design a 1D separable and mixable depthwise inception network that effectively extracts features from various physiological signals.

3. I propose a standardised normal distribution weighted feature fusion method that reconstructs informative maps from different frames of micro-expression video.

4. I develop a guided attention module that achieves multimodal learning for both micro-expression (colour and depth information) and latent emotion recognition (micro-expression and physiological signals).

5.2 Proposed method

Colour images are a crucial and widely used source for computer vision tasks [145, 155, 26], providing valuable information for deep learning models to analyse and interpret visual data. In addition, micro-expression has been extensively studied and recognised as a valuable source for authentic emotion recognition, as discussed in Section 5.1. Therefore, in my proposed framework, I consider colour images from micro-expression video clips as the primary source, with depth information used to guide the spatial features from each frame. Features are extracted from colour images and depth maps separately using backbone networks. To fuse the spatial features

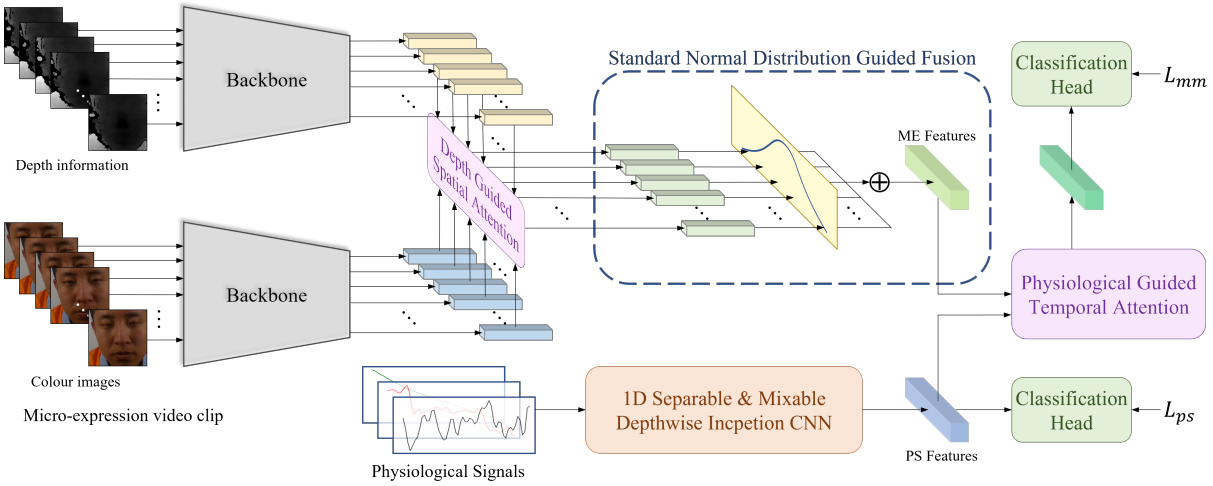


Figure 5.1: Architecture of the proposed framework comprising three main components: *micro-expression feature extraction branch*, *physiological signals feature extraction branch*, and *guided attention fusion module*. The final loss is $(L_{PS} + L_{mm})/2$, where L_{PS} and L_{mm} are both cross-entropy losses calculated from the physiological signals branch and whole multimodal learning, respectively.

extracted from each frame, I designed a standard normal distribution guided fusion method that pays more attention to the middle, where facial movements usually reach their apex, and less attention to the ends. Apart from micro-expression, physiological signals are used in another branch of the proposed framework to recognise latent emotions by my designed 1D separable and mixable depthwise inception network and enhance spatio-temporal features extracted from the micro-expression sample. By analysing both micro-expression and physiological signals, the network can gain a deeper understanding of the subject's emotional state and achieve more accurate latent emotion recognition results. Figure 5.1 shows the proposed framework.

5.2.1 1D separable & mixable depthwise inception CNN

In this framework, I designed a separable & mixable depthwise inception network that can effectively extract features from physiological signals. The network's structure is illustrated in Figure 5.2. The depthwise structure comprises separate convolutions for each group of channels, allowing for more precise feature extraction and capturing of spatial correlations. This enables the network to learn more complex and diverse representations of the input data. The inception block uses convolutional filters of varying sizes within a single layer to capture features at multiple scales without the need for multiple layers, which can be computationally expensive. The inception block can be thought of as an ensemble of smaller networks with different filter sizes, providing a form of regularisation that helps prevent overfitting and improves generalisation performance.

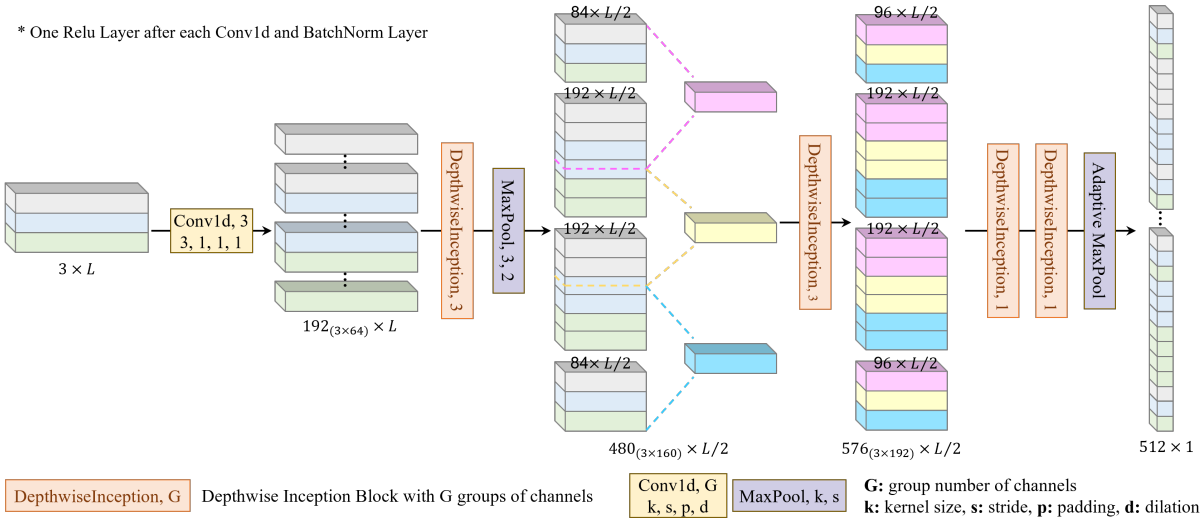


Figure 5.2: The separable and mixable network proposed for physiological signals, see Figure 5.3 for details of depthwise inception block, where L is the 1D input length of signals.

To effectively extract features from multiple input channels, I propose a method that involves the extraction of features from each channel individually, followed by their combination for further learning. To achieve this, I first use a depthwise convolutional layer to enhance the features from each input channel. The depthwise convolutional layer applies a single convolutional filter to each input channel separately, enabling it to capture more specific and precise features.

I then utilise a depthwise inception block with three groups to extract features from each physiological source, as illustrated in Figure 5.3. The depthwise inception block includes four branches of depthwise convolutional layers with varying kernel sizes to extract features from different scales. This design allows the network to capture features at different levels of abstraction and complexity, enabling it to learn more robust and generalisable representations of the input data. The resulting features from each branch are then concatenated together and fed into the next depthwise inception block.

As features from each branch in the previous block are concatenated together and the number of output channels in each branch of the inception block is different, the features from different sources are mixed together to form a new group of channels for further learning. This mixing process enables the network to capture a wide range of features from different sources, which can be combined and refined for improved performance. Finally, the mixed features are fed into the last two blocks as a whole group to extract the final features from the physiological sources. These blocks enable the network to effectively extract and combine features from different input sources.

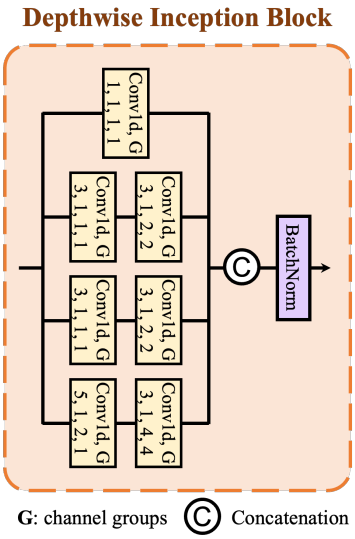


Figure 5.3: Illustration of the depthwise inception block's network design and layer hierarchy.

5.2.2 Standardised normal distribution weighted feature fusion

The sequence of extracted features is mapped to the micro-expression feature by utilising a standard normal distribution function. It is important to highlight that the standard normal distribution serves as a special case of the normal distribution, with mean $\mu = 0$ and variance $\sigma^2 = 1$. However, given the discrete nature of the frames, a slight adjustment is made to the function to effectively map a set of extracted features from all frames to a set of values within the range of $(0, 1)$. These resulting values signify the weight of the features within the set.

A widely accepted belief within the field is that during a micro-expression instance, the most pronounced facial movement typically occurs approximately in the middle of the timeframe. In other words, the apex frame of a micro-expression sample tends to fall roughly in the middle of the clip. Additionally, it is commonly observed that frames in close proximity to the apex frame tend to encompass more valuable features compared to those further away. With this in mind, my approach deviates from extracting spatial features solely from the apex frame and instead incorporates all features across several adjacent frames, spanning the entire duration. This approach aims to capture a more comprehensive representation of the micro-expression.

To assign weights to the features, I have moved away from employing a uniformly weighted function, where each feature carries the same weight. Instead, I adopt a weighting scheme that places greater emphasis on features extracted from frames closer to the middle of the clip, as these frames are considered more representative and valuable. This weighting scheme aligns with the aforementioned proposition. The function of weight for each frame f can be expressed

as follows:

$$W_f = \frac{\exp\left(-\frac{i^2}{2}\right)}{\sum_{f=0}^{F-1}\exp\left(-\frac{i^2}{2}\right)}, \quad \text{where} \quad i = -3\sigma + f \cdot \frac{6\sigma}{F-1}, \tag{5.1}$$

where F denotes the total number of frames within a single micro-expression sample. It is noteworthy that approximately 99.7% of the probability density for a standard normal distribution lies within 3 standard deviations of the mean. Hence, for the purpose of assigning weights, only the features within 3 standard deviations are considered, while the remaining 0.3% is deemed negligible.

5.2.3 The depth/physiology guided attention module

I opted for the guided attention model as my chosen multi-modal learning feature fusion strategy due to its inherent ability to intelligently weigh and integrate information from different modalities. Unlike simplistic concatenation methods, guided attention fusion enables the model to selectively focus on salient aspects of the input data, optimising the learning process.

The module is designed for feature fusion of both colour and depth information for each frame of micro-expression, as well as the final fusion of micro-expression and physiological signals features. For attention modules, formally I have a query Q, a key K, and a value V to calculate attention. The depth and physiological signals features could be considered as the input Q to guide attention learning. At the beginning of the module, the main features are copied as sources for both inputs K and V. After fully-connected layers, the scaled dot-product attention mechanism, denoted as SDP below, is run through several times in parallel. The scaled dot-product attention is an attention mechanism where the dot products are scaled down by $\sqrt{d}$:

$$\text{SDP}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{5.2}$$

d is the dimension of the queries and keys, and softmax denotes the softmax function. The dot product results of the attention mechanism are divided by $\sqrt{d}$ to maintain a variance of 1. The independent attention outputs are then concatenated and linearly transformed to the expected dimension. The multi-head attention mechanism is defined as follows:

$$\text{MultiHead}\,(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\text{head}_1,\ldots,\text{head}_h]\,\mathbf{W}_0, \tag{5.3}$$

$$\text{head}_i = \text{SDP}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right), \tag{5.4}$$

where $\mathbf{W}$ represents the learnable parameter matrices. The multi-head attention mechanism allows for different parts of the sequence to be attended to differently, such as longer-term

dependencies versus shorter-term dependencies.

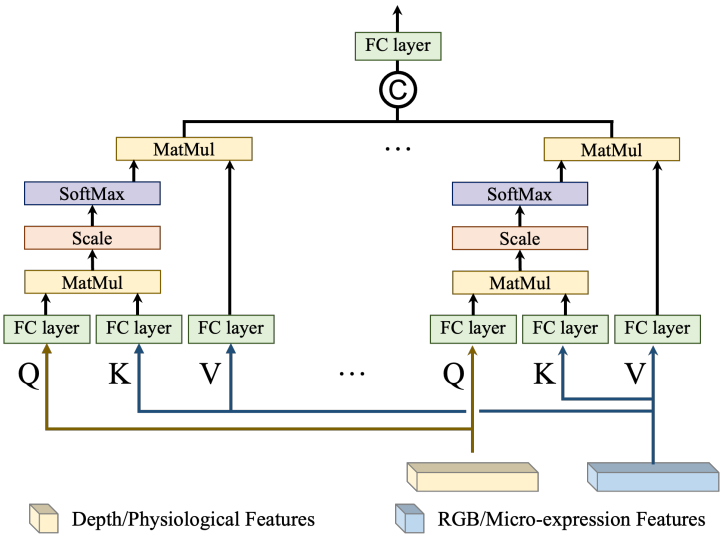


Figure 5.4: Structure of the guided attention module.

5.3 Empirical investigation

5.3.1 Data preparation and experiment setting

In these experiments, I use Part C of CAS(ME)^3 corpus [101] developed to address the challenges of micro-expression elicitation, collection, and annotation. CAS(ME)^3 is composed of three parts. Part A and B contain labelled and unlabelled long videos recorded in the same environment and labelled by the same labellers. CAS(ME)^3 uniquely introduces multimodality to micro-expression analysis in Part C, which is a third-generation multimodal spontaneous micro-expression database that goes beyond just RGB images and includes depth information, voice, and physiological signals. Part C used a third-generation of micro-expression eliciting paradigm, mock crime, with higher ecological validity. Participants were asked to steal a small amount of money from an envelope and were subsequently questioned about the theft. The scenario was designed to create a stressful situation that would elicit spontaneous micro-expressions associated with guilt or deception. Part C contains 166 micro-expressions from 31 subjects and makes it possible to enrich multimodal micro-expression analysis with physiological signals, including EDA, ECG, RSP, and pulse photoplethysmography (PPG).

The colour and depth frames in Part C are captured at a frame rate of 30 fps. Due to the definition of the happening time of a micro-expression, I selected only the samples with less than 15 frames (500ms) from the database. I cropped the facial region based on the landmarks detected from the

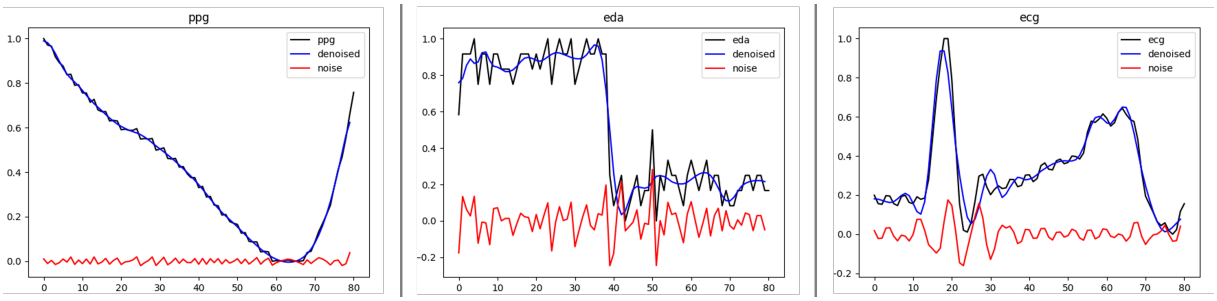


Figure 5.5: Examples of Daubechies wavelet denoising results for physiological signals.

onset colour image, and this cropping was applied to all subsequent colour and depth images. To process the data, I utilised pretrained VGG-Face [133] and VGG-16 [156] networks as backbone networks for colour and depth, respectively.

As for the physiological signals branch, I utilised EDA, ECG, and PPG signals as the three-channel input. To process the source signal data, I employed wavelet denoising on the segmented signal clips. This method is highly effective in denoising 1D signals due to its ability to capture both local and global features of the signal accurately, whilst maintaining a good balance between time and frequency localisation. Daubechies wavelets are orthogonal and form a complete basis set, allowing the signal to be decomposed into its wavelet coefficients, which can then be thresholded to remove noise, as shown in Figure 5.5.

5.3.2 Experimental results

The traditional evaluation approach for MER involves LOSO cross-validation, where a single subject's data is withheld and used as a validation data set, while all remaining subjects' data is used for training. The overall performance of a method is then assessed by aggregating the results of all different possible iterations of the process, i.e. of all subjects being withheld in turn. Li et al. also utilised LOSO in their experiments [101]. To ensure fair comparison and more accurate evaluation, I applied the LOSO approach in my experiments as well. Accuracy, UF1 and UAR, averaging the per-class recall and F1-score respectively, are used as metrics during evaluation.

My study aimed to investigate multimodal latent emotion recognition, and the main results can be seen in Table 5.1. To confirm the effectiveness of my proposed network structure and each designed module, I conducted ablation experiments. Table 5.2 presents the results related to the standardised normal distribution guided spatial feature fusion, while Table 5.3 displays the results of the depth/physiology guided attention module.

Table 5.1: Comparison of multimodal analysis for latent emotion recognition. "Colour" and "Depth" are from micro-expression samples, and "physiological signals" indicates the combination of EDA, PPG, and ECG in my results, while representing the use of only EDA in Li et al.'s results.

	Colour			Colour + Depth			Colour + Depth + physiological signals			Colour + physiological signals		
	Acc	UF1	UAR	Acc	UF1	UAR	Acc	UF1	UAR	Acc	UF1	UAR
Li et al. [101]	-	0.248	0.263	-	0.296	0.296	-	0.230	0.244	-	0.230	0.260
Mine	**0.640**	**0.353**	**0.345**	**0.640**	**0.315**	**0.318**	**0.738**	**0.586**	**0.563**	**0.750**	**0.642**	**0.578**

5.3.3 Analysis and discussion

All results of my experiments will be discussed in this section. I start by investigating the impact of multimodal learning on latent emotion recognition performance. Then, I discuss the effectiveness of each design inside the proposed framework.

The results presented in Table 5.1 demonstrate a significant improvement in performance compared to the benchmark results of Li et al. who used RGB and depth information from the apex frame only as 4-channel input for AlexNet. While producing worse performance than the proposed approach their work demonstrated the value of depth information. In contrast to Li et al., herein I used all frames from a video clip and a standard normal distribution feature fusion module to merge the features extracted from all frames. I note that although the use of depth information facilitates the learning of more expressive features, it may introduce noise, which is particularly problematic in micro-expression analysis wherein the signal corrupted by noise is weak; therefore, one of the potential avenues for further research in colour and depth MER could be finding a better approach to denoise the depth information. Regarding physiological signals, Li et al. used them as greyscale 2D input channels to the same backbone, without addressing their noise content or designing a specialised network to process them. In contrast, the proposed approach employed the Daubechies wavelet for denoising and introduced a 1D separable and mixable depthwise inception CNN for feature extraction. my results suggest that this network structure contributes to improved performance in recognising latent emotions.

5.3.4 Ablation study

The performance of the proposed standardised normal distribution weighted fusion method is compared with that of the uniform distributed fusion method in Table 5.2. The results demonstrate that the proposed method can fuse the features extracted from each frame of micro-expressions more effectively than simply adding them together. The weighted fusion method assigns different weights to different features based on their importance, allowing more important features to have a greater influence on the overall learning process. This emphasises the significance of each feature's contribution to the final result and enhances the performance of the model in

recognising micro-expressions.

Table 5.2: Results of the comparison study on standard normal distribution fusion for MER.

	Colour		
	Acc	UF1	UAR
Uniform Distribution	0.610	0.258	0.283
Standard Normal Distribution	**0.640**	**0.353**	**0.345**

Furthermore, a comparison study was conducted to evaluate the impact of depth/physiology guided attention modules on the performance of the proposed model. Incorporating attention mechanisms allows the model to assign varying levels of importance to features, facilitating a more subtle and effective fusion of diverse information. By dynamically adjusting attention weights, the model can emphasise relevant cues while suppressing noise or less informative signals from the input modalities. This adaptability is crucial for enhancing the model's overall performance in complex tasks like latent emotion recognition.

Table 5.3: Results of the comparison study on the impact of depth and spatial guided attention modules for multimodal latent emotion learning.

	Colour + Depth			micro-expression + physiological signals		
	Acc	UF1	UAR	Acc	UF1	UAR
Concatenation	0.604	0.262	0.288	0.701	0.492	0.468
Guided Attention module	**0.640**	**0.315**	**0.318**	**0.738**	**0.586**	**0.563**

I used concatenation as the baseline fusion method for multimodal learning and trained and tested four different configurations of the model. The results of the study, presented in Table 5.3, revealed that the depth-guided attention module outperformed concatenation in incorporating colour and depth information. Additionally, the physiology-guided attention module used for emotion recognition demonstrated significantly better results, indicating that these guided attention modules are capable of effectively fusing extracted features from multiple modalities to learn more beneficial mixed features and contribute to the improved performance of the proposed model in latent emotion recognition.

5.4 Conclusive remarks and reflections

Emotional states have a significant impact on physical and psychological well-being, and the recognition of emotions is essential for effective communication and understanding of an individual's emotional and mental state. However, relying solely on facial expressions is not sufficient as people can control these signals to hide their real emotions, especially during social

communication. Therefore, in this chapter, I propose a multimodal learning framework that combines micro-expressions and physiological signals to enhance latent emotion recognition performance. The proposed approach denoises the signals and uses a 1D separable and mixable depthwise inception CNN for physiological feature extraction. Furthermore, in this work, I propose a standardised normal distribution weighted feature fusion method and a guided attention module that achieves multimodal learning for both micro-expression and latent emotion recognition. The results show a significant improvement in performance compared to the benchmark results, demonstrating the potential benefits of incorporating multimodal data for improving latent emotion recognition.

CHAPTER SIX

CONCLUSIONS AND FUTURE WORK

In this chapter, an extensive and comprehensive overview of the contributions made in this thesis concerning latent emotion recognition is presented, accompanied by a meticulous and in-depth discussion. The study of micro-expressions and latent emotions is a burgeoning research area that is still in its early stages of development, rather than a fully matured and established field of study. Therefore, it is unsurprising that numerous challenges persist within this domain, underscoring the importance of continued investigation and exploration. To provide clarity on these existing hurdles, this chapter concisely outlines the current challenges while also offering valuable insights into potential directions for future research. These research avenues hold promise in overcoming the obstacles faced in the realm of latent emotion recognition. By addressing these challenges head-on and proposing possible solutions, this thesis contributes to the broader understanding and advancement of latent emotion recognition. Moreover, this work lays a solid foundation for further exploration and innovation in this captivating field. It serves as a springboard for future researchers to delve deeper into the intricacies of latent emotion recognition, building upon the insights and findings presented in this thesis. Through its comprehensive analysis and forward-looking perspective, this chapter consolidates the significance of the research conducted, highlighting its potential impact on the development and evolution of latent emotion recognition as a vital area of study.

6.1 Summary of the contributions

This section presents a comprehensive summary of the contributions made in this thesis, which aims to address three main hypotheses through the proposal of novel approaches. In the pre-

ceding chapters, each hypothesis is explored in detail, and the implementation results obtained provide substantial support for their validity. By delving into these hypotheses and employing innovative methodologies, this research sheds light on significant aspects of the subject matter and contributes to the existing body of knowledge in the field. The following subsections provide an overview of the key contributions made in each hypothesis, emphasising the novel approaches introduced and the corresponding experimental results that bolster their credibility.

6.1.1 Facial action unit detection in micro-expression

Chapter 3 illustrates a method for the detection of activated AU during micro-expression. The results support the **Hypothesis 1:** *computer vision methods can effectively detect and analyse facial muscle movements, such as AU, facilitating a better understanding of the relationship between micro-expressions and emotional states* and its sub-hypotheses. Let us scrutinise the contribution of each methodological aspect in substantiating these hypotheses:

1. **Segmentation of facial key subregions:** The proposed method incorporates a meticulous segmentation approach based on the precise localisation of activated facial muscles (AUs) and facial landmarks. By partitioning the face into distinct subregions, the method aims to enhance the veracity and efficacy of detecting micro AUs. This segmentation methodology aligns with Hypothesis 1.1, as it acknowledges the significance of concentrating on specific facial regions associated with AU activation. By isolating these regions, the method provides a targeted analysis of the pertinent facial muscle movements, thereby corroborating the hypothesis that the segmentation of key subregions augments AU detection and micro-expression analysis.

2. **Multi-label classification and region assignment:** Following the segmentation process, the proposed method executes a multi-label classification paradigm by allocating AUs to different localised areas within the segmented regions. This approach simplifies the task of micro-expression AU detection by decomposing the intricate multi-label classification predicament into more manageable constituents. By assigning AUs to distinct local areas, the method mitigates the complexity of the classification task, thus affirming Hypothesis 1.2. This hypothesis posits that the division of the multi-label classification problem into smaller segments based on segmented regions facilitates AU detection. The successful implementation of this classification and assignment scheme provides empirical support for the hypothesis.

3. **AU-independent cross-validation:** The evaluation framework of the proposed method employs an AU-independent cross-validation methodology. This approach ensures a robust

and reliable evaluation metric for assessing the performance of AU detection approaches in micro-expression analysis. By designing an evaluation framework that is not reliant on specific AUs, the method provides a comprehensive assessment of its performance across various micro-expression scenarios. This corroborates Hypothesis 1.3, which contends that the utilisation of an AU-independent cross-validation method furnishes a dependable metric for evaluating AU detection. The successful incorporation of this evaluation approach bolsters the main hypothesis, substantiating the claim that computer vision methods can effectively detect and analyse facial muscle movements in micro-expressions.

6.1.2 A novel advanced architecture for micro-expression recognition

In Chapter 4, I propose spatio-temporal transformer architecture substantiates the main **Hypothesis 2:** *advanced architectures that can learn both short and long-range spatio-temporal relationships from micro-expression data, can significantly improve the accuracy of automatic latent emotion recognition* by improving the accuracy of automatic latent emotion recognition from micro-expressions. I will now investigate how each methodological aspect lends support to its sub-hypotheses:

1. **Novel deep learning architecture:** The proposed spatio-temporal deep learning architecture represents the first purely transformer-based approach for MER, devoid of any convolutional network usage. This architectural innovation directly addresses Hypothesis 2.1, which suggests that novel approaches utilising transformer architectures for video-based MER achieve comparable or superior performance compared to existing deep learning methods with other architectures. Through comprehensive evaluations on widely-used micro-expression datasets, the proposed approach consistently outperforms the state-of-the-art, providing empirical evidence to support the hypothesis.

2. **Alternative input representations:** The proposed approach incorporates a modification called "long-term optical flow" to overcome the limitations of traditional optical flow in micro-expression videos. By calculating the optical flow between each sample frame and the onset frame of the micro-expression, the method captures the unique dynamics of micro-expressions more effectively. This supports Hypothesis 2.2, demonstrating that alternative input representations, such as modified optical flow, enhance the accuracy and robustness of MER systems. The findings validate the effectiveness of long-term optical flow in capturing the subtle spatio-temporal patterns of micro-expressions.

3. **Integration of temporal information and spatial relations:** The proposed method aligns with Hypothesis 2.3 by employing a spatial feature extraction technique based on transformer

encoders to extract long-range spatial relations from each frame of micro-expression video clips. The approach represents input frames as sequences of constituent patches and converts them into vector sequences, preserving the local spatial features of each image patch. These sequences are then processed by the transformer encoder to extract long-term spatial features. By incorporating the transformer encoder, the proposed method captures the intricate spatial relationships within micro-expressions, enhancing the discriminative power of the extracted features. This provides empirical evidence in support of Hypothesis 2.3, demonstrating that integrating temporal information from optical flow with spatial feature extraction using transformer encoders improves the effectiveness of micro-expression analysis for latent emotion recognition.

4. **Temporal aggregation mechanisms:** The proposed method substantiates Hypothesis 2.4 by incorporating temporal aggregation mechanisms that connect spatio-temporal features extracted from multiple frames in micro-expression videos. By leveraging these mechanisms, the approach addresses the challenge of effectively capturing both local and global spatio-temporal patterns. The temporal aggregation facilitates the integration of information across frames, allowing for a more comprehensive analysis of the temporal dynamics within micro-expressions. This integration enhances the overall effectiveness of the video-based MER system. The inclusion of temporal aggregation mechanisms contributes to the improved performance and effectiveness of the system in recognising latent emotions from micro-expressions, providing empirical evidence to validate the hypothesis.

6.1.3 Multi-modal latent emotion recognition

Chapter 5 presents a multimodal learning framework that supports the main idea stated in **Hypothesis 3:** *incorporating multimodal data, such as physiological signals and depth information, alongside spontaneous micro-expressions can enhance the accuracy and robustness of emotion recognition systems, providing a more comprehensive understanding of emotions.* This chapter also highlights the potential for further research and applications in this field. In the subsequent analysis, I will examine how each aspect of the methodology contributes to validating the related sub-hypotheses:

1. **Enhancing emotion recognition through multimodal integration:** The innovative multimodal learning architecture utilised in the proposed framework directly tackles Hypothesis 3.1. By integrating micro-expressions and physiological signals within this architecture, the framework enhances the accuracy of recognising latent emotions. Through extensive evaluations, it is consistently demonstrated that the proposed approach outperforms existing benchmark methods, offering empirical validation for the hypothesis.

2. **Robust feature extraction from physiological signals:** The integration of a separable and mixable network structure into the multi-modal learning framework further strengthens the support for Hypothesis 3.2. This advanced network architecture adeptly extracts a broad spectrum of features from various physiological signals, capitalising on their unique and interconnected characteristics to improve emotion recognition. The results obtained from rigorous evaluations provide compelling evidence that validates the effectiveness of the separable and mixable network flow in capturing discriminative patterns within physiological data.

3. **Enhancing map reconstruction in micro-expression video clips:** The standardised normal distribution weighted feature fusion method aligns with Hypothesis 3.3. This method specifically enhances the reconstruction of informative maps from distinct frames of micro-expression videos by leveraging their temporal dynamics. By integrating this fusion method into the framework, the representation of emotional cues is improved, resulting in a more accurate and robust recognition of emotions.

4. **Guided attention for both multimodal micro-expression and latent emotion recognition:** The external feature guided attention modules, integrated into the multimodal learning framework, substantiate Hypothesis 3.4. These attention modules facilitate multimodal learning for both micro-expressions (colour and depth information) and latent emotion recognition (micro-expressions and physiological signals). By guiding attention to relevant features and their relationships across modalities, the framework enhances the overall performance of the system in recognising and understanding emotions.

6.1.4 Further remarks

The contributions made in this thesis are not only valuable for academic research but also hold great promise for practical applications, such as in human-computer interaction, mental health assessment, and security. The incorporation of the micro-AU detection task could advance our understanding of the intricate relationships between action units (AU) and micro-expressions by providing a richer pool of AU evidence. Additionally, it has the potential to facilitate the development of training tools essential for individuals in professions that necessitate acute emotional perception, such as law enforcement and healthcare. The introduction of this novel architecture, boasting improved accuracy in discerning latent emotions from micro-expressions, holds promise for enhancing emotion recognition systems in practical, real-world applications. Moreover, integrating physiological signals into emotion recognition may pave the way for advancements in mental health assessment tools. The proposed methodologies and findings can guide future research, encouraging the exploration of combining novel approaches and the

integration of multimodal data for a more comprehensive classification of human emotions. This can lead to the development of empathetic interfaces and applications.

6.2 Outstanding challenges and future work

The field of micro-expression study is currently considered to be in its early stages, with ongoing research and development. Consequently, it is not yet considered a fully mature research field. As a result, numerous challenges persist within this domain. This section aims to provide a concise overview of the current challenges and potential avenues for future research that hold promise. Although significant progress has been made in understanding micro-expressions, several obstacles still impede further advancements.

6.2.1 Data and its limitations

As I discussed in Chapter 2, a major practical obstacle limiting research on micro-expressions concerns the availability, quality, and standardisation of data used by researchers. One of the fundamental issues stems from the fact that repeatable and uniform stimulation of spontaneous micro-expressions is challenging. In research to date, participants are usually exposed to emotional videos which are then expected to rouse participants' emotions, but which the participants are asked to attempt to conceal. Since in some instances, emotional arousal fails, many recordings end up being useless as they contain no micro-expression exhibition – this is one of the reasons why both the number of micro-expression corpora is small and why each of the data sets contains relatively few class examples.

Another practical difficulty, pervasive in data intensive applications, concerns the encoding or labelling of data, which is very time-consuming and laborious. The process requires a trained and skilled labeller, repeated examination of participants' recordings (often in slow motion), and the marking of the micro-expression onset, peak, and termination. Thus, in addition to the process being laborious and slow, it is also inexact, with inter-labelled variability being an issue. Closely connected to this problem is the fact that there is no uniform and widely accepted standard for the classification of micro-expressions. Therefore, the labelling approaches adopted for different databases are different (with similar micro-expressions treated as different depending on the data set used), posing challenges to understanding the performance of the state-of-the-art, relative performances of different methods. There is no doubt that further work in this area is badly needed and that contributions to standardisation would benefit the field enormously.

6.2.2 Real-time micro-expression recognition

In the realm of micro-expression analysis, the tasks of micro-expression classification and the mapping of the corresponding class clusters onto the space of emotions are certainly the most widely addressed ones in the literature, and arguably the most important ones. In some practical applications, it is desirable to be able to do this in real-time. Considering that the duration of a micro-expression is very short, from 1/25 to 1/5 of a second, it is clear that this is a major computational challenge, especially when the application is required to run on embedded or mobile devices. Although workarounds are possible in principle, e.g. by offloading computation to more powerful servers, this may not always be possible and new potential bottlenecks emerge due to the need to transmit large amounts of data. Given the lack of attention to the problems associated with computational efficiency in the existing literature and the aforementioned need for micro-expression analysis in real-time, this direction of research also offers a range of opportunities for valuable future contributions.

6.2.3 Standardisation of performance metrics

The standardisation of performance metrics in MER is a critical challenge that requires further improvement in the evaluation process. Future work should focus on addressing the time-consuming nature of cross-database evaluations and developing new evaluation metrics that encompass both single-dataset and cross-database assessments. Through collaborative efforts and the adoption of standardised evaluation protocols, researchers can enhance the reliability and comprehensiveness of MER methods, leading to significant advancements in the field. The commonly employed method, LOSO cross-validation, while effective, can become time-consuming, particularly when dealing with a large number of subjects. The computational burden of withholding data from each subject individually in multiple iterations hinders the evaluation process. Therefore, it is necessary to explore alternative strategies or optimisations to expedite evaluations without compromising reliability and comprehensiveness.

To address this challenge, future work in the field of MER should focus on developing new standard evaluation metrics. These metrics should not only account for the accuracy of single-dataset evaluations but also encompass cross-database evaluations. By establishing common evaluation protocols and benchmarks that can be consistently applied across datasets, researchers and practitioners can ensure comparability and standardisation of results. Collaborative efforts among the MER community are crucial in defining these frameworks, promoting transparency and the adoption of best practices. Cross-database evaluations offer several advantages over single-dataset evaluations, such as increasing sample diversity and size, and providing insights into method performance across different populations and contexts. By overcoming the limita-

tions of small dataset sizes, these evaluations help uncover potential biases and limitations in MER methods. With standardised performance metrics and evaluation protocols, researchers can make meaningful comparisons, identify areas for improvement, and develop more robust and reliable MER techniques.

6.2.4 Multi-modal latent emotion recognition with contactless bio-signal measurement

The integration of micro-expression analysis and contactless bio-signals in multi-modal latent emotion recognition holds immense potential for advancing our understanding of human emotions. By combining the subtle facial expressions captured through micro-expression analysis with non-intrusive bio-signals, such as heart rate variability or electrodermal activity, a more comprehensive and accurate assessment of emotional states can be achieved. This novel approach not only allows for a more refined recognition of latent emotions but also opens doors to various applications in fields such as psychology, human-computer interaction, and affective computing. Although there currently exists a micro-expression database with physiological signals, its limitations lie in the restricted sample size and the lack of contactless data collection. Thus, the development of a comprehensive database that encompasses both micro-expressions and contactless bio-signals would serve as a crucial foundation for future research in this domain. Such a database would enable researchers to explore new avenues in emotion recognition, devise more robust algorithms, and ultimately contribute to the advancement of our understanding of human emotions.

APPENDIX A

PARTIAL SUMMARY OF MICRO-EXPRESSION RECOGNITION WORK ON SPONTANEOUS DATABASES: A COMPREHENSIVE TABLE

Table A.1: Partial Summary of Micro-Expression Recognition Work on Spontaneous Databases.

Paper	Feature	Method	Database	Best Result
2011 Pfister et al. [140]	Hand-crafted	LBP-TOP	Early SMIC	Acc: 71.4%
2013 Li et al. [103]	Hand-crafted	LBP-TOP	SMIC-VIS	Acc: 52.11%
2014 Guo et al. [55]	Hand-crafted	LBP-TOP	SMIC	Acc: 65.83%
2014 Wang et al. [184]	Hand-crafted	TICS	CASME CASME II	Acc: 61.85% Acc: 58.53%
2014 Wang et al. [183]	Hand-crafted	DTSA	CASME	Acc: 46.90%
2014 Yan et al. [206]	Hand-crafted	LBP-TOP	CASME II	Acc: 63.41%
2015 Huang et al. [64]	Hand-crafted	STLBP-IP	SMIC CASME II	Acc: 57.93% Acc: 59.51%
2015 Huang et al. [65]	Hand-crafted	STCLQP	SMIC CASME	Acc: 64.02% Acc: 57.31%

Paper	Feature	Method	Database	Best Result
			CASME II	Acc: 58.39%
2015 Le et al. [95]	Hand-crafted	DMDSP+LBP-TOP	CASME II	F1-score: 0.52
2015 Le et al. [96]	Hand-crafted	LBP-TOP+STM	SMIC CASME II	Acc: 44.34% Acc: 43.78%
2015 Liong et al. [107]	Hand-crafted	OSW-LBP-TOP	SMIC CASME II	Acc: 57.54% Acc: 66.40%
2015 Lu et al. [117]	Hand-crafted	DTCM	SMIC CASME CASME II	Acc: 82.86% Acc: 64.95% Acc: 64.19%
2015 Wang et al. [186]	Hand-crafted	TICS, CIELuv and CIELab	CASME CASME II	Acc: 61.86% Acc: 62.30%
2015 Wang et al. [191]	Hand-crafted	LBP-SIP and LBP-MOP	CASME	Acc: 66.8%
2016 Ben et al. [8]	Hand-crafted	MMPTR	CASME	Acc: 80.2%
2016 Chen et al. [22]	Hand-crafted	3DHOG	CASME II	Acc: 86.67%
2016 Kim et al. [87]	Deep Learning	CNN+LSTM	CASME II	Acc: 60.98%
2016 Liong et al. [108]	Hand-crafted	Optical Strain	SMIC CASME II	Acc: 52.44% Acc: 63.41%
2016 Liu et al. [113]	Hand-crafted	MDMO	SMIC CASME CASME II	Acc: 80% Acc: 68.86% Acc: 67.37%
2016 Oh et al. [126]	Hand-crafted	I2D	SMIC CASME II	F1-score: 0.44 F1-score: 0.41
2016 Talukder et al. [165]	Hand-crafted	LBP-TOP	SMIC-NIR	Acc: 62%
2016 Wang et al. [187]	Hand-crafted	STCCA	CASME CASME II	Acc: 41.20% Acc: 38.39%
2016 Zheng et al. [229]	Hand-crafted	LBP-TOP, HOOF	CASME CASME II	Acc: 69.04% Acc: 63.25%
2017 Happy and Routray [59]	Hand-crafted	FHOFO	SMIC CASME CASME II	F1-score: 0.524 F1-score: 0.549 F1-score: 0.525
2017 Liong et al. [109]	Hand-crafted	Bi-WOOF	SMIC-VIS CASME II	Acc: 53.52% F1-score: 0.59
2017 Peng et al. [137]	Deep Learning	DTSCNN	CASMEI/II	Acc: 66.67%
2017 Wang et al. [193]	Hand-crafted	LBP-TOP	CASME II	Acc: 75.30%
2017 Zhang et al. [216]	Hand-crafted	LBP-TOP	CASME II	Acc: 62.50%
2018 Ben et al. [9]	Hand-crafted	HWP-TOP	CASME II	Acc: 86.8%
2018 Hu et al. [63]	Hand-crafted	LGBP-TOP and CNN	SMIC CASME II	Acc: 65.1% Acc: 66.2%
2018 Khor et al. [85]	Deep Learning	ELRCN	CASME II SAMM	F1-score: 0.5 F1-score: 0.409
2018 Li et al. [104]	Hand-crafted	HIGO	SMIC-HS	Acc: 68.29%

Paper	Feature	Method	Database	Best Result
			CASME II	Acc: 67.21
2018 Liong et al. [110]	Hand-crafted	Bi-WOOF	SMIC-HS	F1-score: 0.62%
			CASME II	F1-score: 0.61
2018 Su et al. [158]	Hand-crafted	DS-OMMA	CASME II	F1-score: 0.724
			CAS(ME)^2	F1-score: 0.737
2018 Zhu et al. [237]	Hand-crafted	LBP-TOP and OF	CASME II	Acc: 53.3%
2018 Zong et al. [238]	Hand-crafted	STLBP-IP	CASME II	Acc: 63.97%
2019 Gan et al. [46]	Deep Learning	OFF-ApexNet	SMIC	Acc: 67.6%
			CASME II	Acc: 88.28%
			SAMM	Acc: 69.18%
2019 Huang et al. [66]	Hand-crafted	DiSTLBP-RIP	SMIC	Acc: 63.41%
			CASME	Acc: 64.33%
			CASME II	Acc: 64.78%
2019 Li et al. [100]	Deep Learning	3D-FCNN	SMIC	Acc: 55.49%
			CASME	Acc: 54.44%
			CASME II	Acc: 59.11%
2019 Liong et al. [111]	Deep Learning	STSTNet	Composite	UF1: 0.735 and UAR: 0.760
			SMIC	UF1: 0.680 and UAR: 0.701
			CASME II	UF1: 0.838 and UAR: 0.869
			SAMM	UF1: 0.659 and UAR: 0.681
2019 Liu et al. [114]	Deep Learning	EMR	Composite	UF1: 0.789 and UAR: 0.782
			SMIC	UF1: 0.746 and UAR: 0.753
			CASME II	UF1: 0.829 and UAR: 0.821
			SAMM	UF1: 0.775 and UAR: 0.715
2019 Peng et al. [139]	Hand-crafted	HIGO-TOP, ME-Booster	SMIC-HS	Acc: 68.90%
			CASME II	Acc: 70.85%
2019 Peng et al. [138]	Deep Learning	Apex-Time Network	SMIC	UF1: 0.497 and UAR: 0.489
			CASME II	UF1: 0.523 and UAR: 0.501
			SAMM	UF1: 0.429 and UAR: 0.427
2019 Van Quang et al. [172]	Deep Learning	CapsuleNet	Composite	UF1: 0.652 and UAR: 0.651
			SMIC	UF1: 0.582 and UAR: 0.588
			CASME II	UF1: 0.707 and UAR: 0.701
			SAMM	UF1: 0.621 and UAR: 0.599
2019 Xia et al. [199]	Deep Learning	MER-RCNN	SMIC	Acc: 57.1%
			CASME	Acc: 63.2%
			CASME II	Acc: 65.8%
2019 Zhao and Xu [227]	Hand-crafted	NMPs	SMIC	Acc: 69.37%
			CASME II	Acc: 72.08%
2019 Zhou et al. [233]	Deep Learning	Dual-Inception	Composite	UF1: 0.732 and UAR: 0.728
			SMIC	UF1: 0.665 and UAR: 0.673
			CASME II	UF1: 0.862 and UAR: 0.856
			SAMM	UF1: 0.587 and UAR: 0.566
2020 Wang et al. [176]	Deep Learning	ResNet, Micro-Attention	SMIC	Acc:49.4%
			CASME II	Acc:65.9%
			SAMM	Acc: 48.5%

Paper	Feature	Method	Database	Best Result
2020 Xie et al. [203]	Deep Learning	AU-GACN	CASME II SAMM	Acc:49.2% Acc: 48.9%
2020 Buhari et al. [13]	Deep Learning	FACS-based graph	CASME II SMIC SAMM	Acc: 75.05% Acc: 70.25% Acc: 87.33%
2020 Cen et al. [16]	Hand-crafted	Enhanced LCBP	CASME II SMIC SAMM	Acc: 78.45% Acc: 79.26% Acc: 79.41%
2020 Chen et al. [20]	Deep Learning	CBAMNet	CASME II SMIC	Acc: 69.92% Acc: 54.84%
2020 Choi and Song [24]	Deep Learning	LFM	CASME II SMIC	Acc: 73.98% Acc: 71.34%
2020 Gao et al. [47]	Hand-crafted	LNMF	CASME II SAMM	Acc: 72.60% Acc: 73.30%
2020 Lai et al. [91]	Deep Learning	MACNN	CASME II	Acc: 72.26%
2020 Le et al. [93]	Deep Learning	VGG19 + ROI	CASME II	Acc: 78.50%
2020 Liu et al. [112]	Deep Learning	Optical flow + CNN	CASME II	Acc: 64.63%
2020 Lo et al. [115]	Deep Learning	MER-GCN	CASME II	Acc: 58.82%
2020 Pan et al. [129]	Hand-crafted	H-SVM	CASME II SMIC SAMM	Acc: 73.10% Acc: 73.10% Acc: 58.46%
2020 Sun et al. [159]	Hand-crafted	MAP-LBP-TOP	CASME II SAMM	Acc: 77.30% Acc: 58.82%
2020 Takalkar et al. [163]	Deep Learning	LBP-TOP + CNN	CASME II SMIC SAMM	Acc: 86.20% Acc: 93.30% Acc: 91.70%
2020 Verma et al. [175]	Deep Learning	Affective Net	CASME II SAMM	Acc: 68.74% Acc: 58.12%
2020 Wang et al. [179]	Deep Learning	2D-3D CNN	SAMM	Acc: 85.19%
2020 Xia et al. [200]	Deep Learning	STRCN	CASME II SMIC SAMM	Acc: 80.30% Acc: 72.30% Acc: 78.60%
2020 Zhu et al. [236]	Deep Learning	DSTICNN	CASME II SMIC	Acc: 82.21% Acc: 78.78%
2021 Gajjala et al. [45]	Deep Learning	MERANet	CASME II	Acc: 91.70%
2021 Guermazi et al. [53]	Hand-crafted	LBPAccPu2	CASME II SMIC	Acc: 80.81% Acc: 76.59%
2021 Gupta et al. [57]	Deep Learning	MERASTC	CASME II SMIC SAMM	Acc: 85.40% Acc: 79.30% Acc: 83.80%
2021 Lei et al. [97]	Deep Learning	AUGCN + AUFsuion	CASME II SAMM	Acc: 80.80% Acc: 74.26%

Paper	Feature	Method	Database	Best Result
2021 Li et al. [106]	Deep Learning	LGCcon	CASME II	Acc: 65.02%
			SAMM	Acc: 40.90%
2021 Nie et al. [125]	Deep Learning	GEME	CASME II	Acc: 75.20%
			SMIC	Acc: 64.63%
			SAMM	Acc: 55.88%
2021 Pan et al. [130]	Deep Learning	TSDN	CASME II	Acc: 71.49%
			SMIC	Acc: 68.90%
2021 Takalkar et al. [164]	Deep Learning	LGAttNet	CASME II	Acc: 94.20%
			SAMM	Acc: 86.70%
2021 Wang et al. [190]	Deep Learning	DSTAN	CASME II	Acc: 75.20%
			SMIC	Acc: 77.40%
2021 Wang et al. [178]	Deep Learning	STM-Net	CASME II	Acc: 84.84%
			SMIC	Acc: 78.66%
			SAMM	Acc: 81.13%
2021 Wang et al. [189]	Deep Learning	MESNet	SAMM	Acc: 97.73%
2021 Yang et al. [207]	Deep Learning	MERTA	CASME II	Acc: 60.54%
2021 Zhao et al. [222]	Deep Learning	MERSiamC3D	CASME II	Acc: 80.05%
			SAMM	Acc: 64.03%
2021 Zhao et al. [226]	Deep Learning	ARS + CropNet	CASME II	Acc: 86.20%
2022 Chen et al. [19]	Deep Learning	BDCNN	CASME II	UF1: 0.950 and UAR: 0.952
			SMIC	UF1: 0.786 and UAR: 0.787
			SAMM	UF1: 0.819 and UAR: 0.799
2022 Fan et al. [44]	Deep Learning	ViT	MMEW	Acc: 87.00%
2022 Indolia et al. [71]	Deep Learning	DSResNetAtt	CASME	Acc: 90.34%
			CASME II	Acc: 80.08%
			SAMM	Acc: 95.50%
2022 Wang et al. [180]	Deep Learning	FPR	CASME II	Acc: 70.00% and F1-score: 0.71
			SMIC	Acc: 69.00% and F1-score: 0.69
			CASME	Acc: 74.00% and F1-score: 0.74
2022 Wei et al. [196]	Hand-crafted	LBP-FIP	CASME II	Acc: 79.00%
			SMIC	Acc: 67.86%
2022 Wei et al. [197]	Deep Learning	AMAN	CASME II	Acc: 75.40% and F1-score: 0.713
			SMIC	Acc: 79.87% and F1-score: 0.771
			SAMM	Acc: 68.85% and F1-score: 0.668
2022 Zhao et al. [224]	Deep Learning	Cascade-MEMN + SCL	CASME II	UF1: 0.915 and UAR: 0.915
			SMIC	UF1: 0.717 and UAR: 0.733
			SAMM	UF1: 0.759 and UAR: 0.724
2022 Zhao et al. [223]	Deep Learning	ME-PLAN	CASME II	UF1: 0.894 and UAR: 0.896
			SMIC	UF1: 0.713 and UAR: 0.726
			SAMM	UF1: 0.736 and UAR: 0.769
2022 Zhou et al. [234]	Deep Learning	FeatRef	CASME II	Acc: 68.38%
			SMIC	Acc: 57.90%
			SAMM	Acc: 60.13%

APPENDIX B

LIST OF PUBLICATIONS

1. Liangfei Zhang and Ognjen Arandjelović. Review of automatic micro-expression recognition in the past decade. *Machine Learning and Knowledge Extraction*, 3(2):414–434, 2021

2. Liangfei Zhang, Ognjen Arandjelović, and Xiaopeng Hong. Facial action unit detection with local key facial sub-region based multi-label classification for micro-expression analysis. In *ACM international conference on Multimedia Workshops*, 2021

3. Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelović, and Guoying Zhao. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1973–1985, 2022

4. Liangfei Zhang, Ognjen Arandjelović, Sonia Dewar, Arlene Astell, Gayle Doherty, and Maggie Ellis. Quantification of advanced dementia patients' engagement in therapeutic sessions: An automatic video based approach using computer vision and machine learning. In *Proceedings International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 5785–5788. IEEE, 2020

5. Yifei Qian, Liangfei Zhang, Xiaopeng Hong, Carl Donovan, and Ognjen Arandjelovic. Segmentation assisted u-shaped multi-scale transformer for crowd counting. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022

6. Qingshu Guan, Xiaopeng Hong, Wei Ke, Liangfei Zhang, Guanghui Sun, and Yihong Gong. Kohonen self-organizing map based route planning: A revisit. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7969–7976. IEEE, 2021

APPENDIX C

THE ETHICS APPROVAL

University of St Andrews | FOUNDED 1413

School of Computer Science Ethics Committee

30 October 2023

Dear Liangfei

Thank you for submitting your ethical application which was considered by the School Ethics Committee.

The School of Computer Science Ethics Committee, acting on behalf of the University Teaching and Research Ethics Committee (UTREC), has approved this application:

Approval Code:	CS17299	**Approved on:**	30.10.23	**Approval Expiry:**	30.10.28
Project Title:	Automatic Inference of Latent Emotion from Spontaneous Facial Micro-Expressions				
Researcher(s):	Liangfei Zhang				
Supervisor(s):	Ognjen Arandelovic				

The following supporting documents are also acknowledged and approved:

1. Application Form
2. License Agreement – CASME 1
3. License Agreement – CASME 2
4. License Agreement - SMIC
5. Release Agreement

Approval is awarded for 5 years, see the approval expiry data above.

If your project has not commenced within 2 years of approval, you must submit a new and updated ethical application to your School Ethics Committee.

If you are unable to complete your research by the approval expiry date you must request an extension to the approval period. You can write to your School Ethics Committee who may grant a discretionary extension of up to 6 months. For longer extensions, or for any other changes, you must submit an ethical amendment application.

You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
  - the details provided in your ethical application
  - the University's Principles of Good Research Conduct
  - the conditions of any funding associated with your work
- that you obtain all applicable additional documents (see the 'additional documents' webpage for guidance) before research commences.

You should retain this approval letter with your study paperwork.

REFERENCES

[1] Ognjen Arandjelović. Colour invariants under a non-linear photometric camera model and their application to face recognition from video. *Pattern Recognition*, 45(7):2499–2509, 2012.

[2] Ognjen Arandjelović, Duc-Son Pham, and Svetha Venkatesh. Cctv scene perspective distortion estimation from low-level motion features. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(5):939–949, 2015.

[3] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[4] Amal Azazi, Syaheerah Lebai Lutfi, Ibrahim Venkat, and Fernando Fernández-Martínez. Towards a robust affect recognition: Automatic facial expression recognition in 3d faces. *Expert Systems with Applications*, 42(6):3056–3066, 2015.

[5] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *Proceedings of 7th International Conference on Learning Representations, ICLR*, 2019.

[6] Muzammil Behzad, Nhat Vo, Xiaobai Li, and Guoying Zhao. Automatic 4d facial expression recognition via collaborative cross-domain dynamic image network. *arXiv preprint arXiv:1905.02319*, 2019.

[7] Muzammil Behzad, Nhat Vo, Xiaobai Li, and Guoying Zhao. Towards reading beyond faces for sparsity-aware 3d/4d affect recognition. *Neurocomputing*, 458:297–307, 2021.

[8] Xianye Ben, Peng Zhang, Rui Yan, Mingqiang Yang, and Guodong Ge. Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation. *Neural Computing and Applications*, 2016.

[9] Xianye Ben, Xitong Jia, Rui Yan, Xin Zhang, and Weixiao Meng. Learning effective binary descriptors for micro-expression recognition transferred by macro-information. *Pattern Recognition Letters*, 2018.

[10] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (9):5826–5846, 2022.

[11] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint*, page 1703.01210, 2017.

[12] Susana Brás, Jacqueline HT Ferreira, Sandra C Soares, and Armando J Pinho. Biometric and emotion identification: An ecg compression based method. *Frontiers in psychology*, 9:467, 2018.

[13] Adamu Muhammad Buhari, Chee-Pun Ooi, Vishnu Monn Baskaran, Raphaël C. W. Phan, KokSheik Wong, and Wooi-Haw Tan. FACS-based graph features for real-time micro-expression recognition. *Journal of Imaging*, 6(12), 2020.

[14] Walter B Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4):106–124, 1927.

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[16] Shixin Cen, Yang Yu, Gang Yan, Ming Yu, and Qing Yang. Sparse spatiotemporal descriptor for micro-expression recognition using enhanced local cube binary pattern. *Sensors*, 20(16):4437, 2020.

[17] Ya Chang, Marcelo Vieira, Matthew Turk, and Luiz Velho. Automatic 3d facial expression analysis in videos. In *Analysis and Modelling of Faces and Gestures: Second International Workshop, AMFG 2005, Beijing, China, October 16, 2005. Proceedings 2*, pages 293–307. Springer, 2005.

[18] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, 2009.

[19] Bin Chen, Kun-Hong Liu, Yong Xu, Qing-Qiang Wu, and Jun-Feng Yao. Block division convolutional network with implicit deep features augmentation for micro-expression recognition. *IEEE Transactions on Multimedia*, 25:1345–1358, 2022.

[20] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020.

[21] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[22] Mengting Chen, Heather T. Ma, Jie Li, and Huanhuan Wang. Emotion recognition using fixed length micro-expressions sequence and weighting method. In *2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016*, 2016.

[23] Zhixing Chen, Di Huang, Yunhong Wang, and Liming Chen. Fast and light manifold cnn based 3d facial expression recognition across pose variations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 229–238, 2018.

[24] Dong Yoon Choi and Byung Cheol Song. Facial micro-expression recognition using two-dimensional landmark feature maps. *IEEE Access*, 8:121549–121563, 2020.

[25] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, 8:168865–168878, 2020.

[26] Alessandro Conti, Paolo Rota, Yiming Wang, and Elisa Ricci. Cluster-level pseudo-labelling for source-free cross-domain facial expression recognition. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 1995.

[28] Charles Darwin. *The expressions of the emotions in man and animals*. J. murray, 1872.

[29] Adrian K. Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. SAMM: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, 2018.

[30] Silvia De Nadai, Massimo D'Incà, Francesco Parodi, Mauro Benza, Anita Trotta, Enrico Zero, Luca Zero, and Roberto Sacile. Enhancing safety of transport by road by on-line monitoring of driver emotions. In *2016 11th system of systems engineering conference (SoSE)*, pages 1–4. Ieee, 2016.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova Google. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[33] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge (emotiw) challenge and workshop summary. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 371–372, 2013.

[34] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and Others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[36] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, and Stefano Berretti. 3d dynamic expression recognition based on a novel deformation vector field and random forest. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1104–1107. IEEE, 2012.

[37] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[38] Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion review*, 3(4):364–370, 2011.

[39] Paul Ekman and Wallace V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[40] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 1971.

[41] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[42] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. *Facial Action Coding System: Manual and Investigator's Guide*, 2002.

[43] Junjie Fan and Ognjen Arandjelović. Employing domain specific discriminative information to address inherent limitations of the lbp descriptor in face recognition. In *International Joint Conference on Neural Networks*, pages 1–7. IEEE, 2018.

[44] Yali Fan, Minghan Jia, Yifan Zhang, and Yang Yu. Micro-expression recognition using pre-trained model and transformer. In *2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pages 1404–1408, 2022.

[45] Viswanatha Reddy Gajjala, Sai Prasanna Teja Reddy, Snehasis Mukherjee, and Shiv Ram Dubey. Meranet: Facial micro-expression recognition using 3d residual attention network. In *Proceedings of the twelfth Indian conference on computer vision, graphics and image processing*, pages 1–10, 2021.

[46] Y. S. Gan, Sze Teng Liong, Wei Chuen Yau, Yen Chang Huang, and Lit Ken Tan. OFF-ApexNet on micro-expression recognition system. *Signal Processing: Image Communication*, 2019.

[47] Junli Gao, Huajun Chen, Xiaohua Zhang, Jing Guo, and Wenyu Liang. A new feature extraction and recognition method for microexpression based on local non-negative matrix factorization. *Frontiers in Neurorobotics*, 14:579338, 2020.

[48] Rohit Girdhar, Joao Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[49] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[50] Ardeshir Goshtasby. Image registration by local approximation methods. *Image and Vision Computing*, 1988.

[51] Simeng Gu, Wei Wang, Fushun Wang, and Jason H Huang. Neuromodulator and emotion biomarker for stress induced mental disorders. *Neural plasticity*, 2016, 2016.

[52] Qingshu Guan, Xiaopeng Hong, Wei Ke, Liangfei Zhang, Guanghui Sun, and Yihong Gong. Kohonen self-organizing map based route planning: A revisit. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7969–7976. IEEE, 2021.

[53] Radhouane Guermazi, Taoufik Ben Abdallah, and Mohamed Hammami. Facial micro-expression recognition based on accordion spatio-temporal representation and random forests. *Journal of Visual Communication and Image Representation*, 79:103183, 2021.

[54] Rui Guo, Shuangjiang Li, Li He, Wei Gao, Hairong Qi, and Gina Owens. Pervasive and unobtrusive emotion sensing for human mental health. In *2013 7th international conference on pervasive computing Technologies for Healthcare and Workshops*, pages 436–439. IEEE, 2013.

[55] Yanjun Guo, Yantao Tian, Xu Gao, and Xuange Zhang. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *Proceedings of the International Joint Conference on Neural Networks*, 2014.

[56] Yingchun Guo, Cuihong Xue, Yingzi Wang, and Ming Yu. Micro-expression recognition based on CBP-TOP feature with ELM. *Optik*, 2015.

[57] Puneet Gupta. Merastc: Micro-expression recognition using effective feature encodings and 2d convolutional neural network. *IEEE Transactions on Affective Computing*, 2021.

[58] Ernest A Haggard and Kenneth S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*. Springer, 1966.

[59] S. L. Happy and Aurobinda Routray. Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing*, 2017.

[60] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[61] Jiuk Hong, Chaehyeon Lee, and Heechul Jung. Late fusion-based video transformer for facial micro-expression recognition. *Applied Sciences*, 12(3):1169, 2022.

[62] Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. Lbp-top: a tensor unfolding revisit. *ACCV Workshop on Spontaneous Facial Behavior Analysis*, 2016.

[63] Chunlong Hu, Dengbiao Jiang, Haitao Zou, Xin Zuo, and Yucheng Shu. Multi-task micro-expression recognition combining deep and handcrafted features. In *Proceedings - International Conference on Pattern Recognition*, pages 946–951. Institute of Electrical and Electronics Engineers Inc., 2018.

[64] Xiaohua Huang, Su Jing Wang, Guoying Zhao, and Matti Piteikainen. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[65] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 2015.

[66] Xiaohua Huang, Su Jing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng, and Matti Pietikainen. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. *IEEE Transactions on Affective Computing*, 2019.

[67] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience*, 2017, 2017.

[68] Kelly Hubble, Katharine L Bowen, Simon C Moore, and Stephanie HM Van Goozen. Improving negative emotion recognition in young offenders reduces subsequent crime. *PLoS one*, 10(6), 2015.

[69] Petr Husak, Jan Cech, and Jiri Matas. Spotting facial micro-expressions “in the wild”. In *Proc. Computer Vision Winter Workshop*, 2017.

[70] Xuan-Phung Huynh, Tien-Duc Tran, and Yong-Guk Kim. Convolutional neural network models for facial expression recognition using bu-3dfe database. In *Information Science and Applications (ICISA) 2016*, pages 441–450. Springer, 2016.

[71] Sakshi Indolia, Swati Nigam, and Rajiv Singh. Integration of transfer learning and self-attention for spontaneous micro-expression recognition. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 325–330, 2022.

[72] Carroll E Izard. *Human emotions*. Springer Science & Business Media, 1977.

[73] Carroll E Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3): 260–280, 2007.

[74] Carroll E Izard. Forms and functions of emotions: Matters of emotion–cognition interactions. *Emotion review*, 3(4):371–378, 2011.

[75] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[76] Rachael E Jack, Oliver GB Garrod, and Philippe G Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192, 2014.

[77] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7680–7689, June 2021.

[78] William James. *The Principles of Psychology*, volume 1. Henry Holt, New York, 1890.

[79] Asim Jan and Hongying Meng. Automatic 3d facial expression recognition using geometric and textured feature fusion. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 5, pages 1–6. IEEE, 2015.

[80] Asim Jan, Huaxiong Ding, Hongying Meng, Liming Chen, and Huibin Li. Accurate facial parts localization and deep learning for 3d facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 466–472. IEEE, 2018.

[81] Hirotaka Kaji, Hisashi Iizuka, and Masashi Sugiyama. Ecg-based concentration recognition with multi-task regression. *IEEE Transactions on Biomedical Engineering*, 66(1):101–110, 2018.

[82] Juan Karsten and Ognjen Arandjelović. Automatic vertebrae localization from ct scans using volumetric descriptors. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 576–579. IEEE, 2017.

[83] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[84] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[85] Huai Qian Khor, John See, Raphael Chung Wei Phan, and Weiyao Lin. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In *Proceedings of 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2018.

[86] Huai Qian Khor, John See, Sze Teng Liong, Raphael C.W. Phan, and Weiyao Lin. Dual-stream shallow networks for facial micro-expression recognition. In *Proceedings of International Conference on Image Processing, ICIP*, 2019.

[87] Dae Hoe Kim, Wissam J. Baddar, and Yong Man Ro. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, 2016.

[88] Dae Hoe Kim, Wissam J. Baddar, and Yong Man Ro. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. *Proceedings of the 2016 ACM Multimedia Conference*, pages 382–386, 2016.

[89] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009.

[90] Ankith Jain Rakesh Kumar and Bir Bhanu. Micro-expression classification based on landmark relations with graph attention convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1511–1520, 2021.

[91] Zhenyi Lai, Renhe Chen, Jinlu Jia, and Yurong Qian. Real-time micro-expression recognition based on resnet and atrous convolutions. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2020.

[92] Peter J Lang. The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

[93] Trang Thanh Quynh Le, Thuong-Khanh Tran, and Manjeet Rege. Rank-pooling-based features on localized regions for automatic micro-expression recognition. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 11(4):25–37, 2020.

[94] Vuong Le, Hao Tang, and Thomas S Huang. Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 414–421. IEEE, 2011.

[95] Anh Cat Le Ngo, Sze Teng Liong, John See, and Raphael Chung Wei Phan. Are subtle expressions too sparse to recognize? In *International Conference on Digital Signal Processing, DSP*, 2015.

[96] Anh Cat Le Ngo, Raphael Chung Wei Phan, and John See. Spontaneous subtle expression recognition: Imbalanced databases and solutions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[97] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1571–1580, 2021.

[98] Robert W Levenson. Basic emotion questions. *Emotion review*, 3(4):379–386, 2011.

[99] Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831, 2017.

[100] Jing Li, Yandan Wang, John See, and Wenbin Liu. Micro-expression recognition based on 3d flow convolutional neural network. *Pattern Analysis and Applications*, 22(4):1331–1339, 2019.

[101] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2782–2800, 2023.

[102] Weijian Li, Di Huang, Huibin Li, and Yunhong Wang. Automatic 4d facial expression recognition using dynamic geometrical image network. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 24–30. IEEE, 2018.

[103] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Proceedings of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG*, 2013.

[104] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 2018.

[105] Xiaobai Li, Shiyang Cheng, Yante Li, Muzammil Behzad, Jie Shen, Stefanos Zafeiriou, Maja Pantic, and Guoying Zhao. 4DME: A spontaneous 4D micro-expression dataset with multimodalities. *IEEE Transactions on Affective Computing*, pages 1–18, 2022.

[106] Yante Li, Xiaohua Huang, and Guoying Zhao. Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 30, 2021.

[107] Sze Teng Liong, John See, Raphael C.W. Phan, Anh Cat Le Ngo, Yee Hui Oh, and Kok Sheik Wong. Subtle expression recognition using optical strain weighted features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[108] Sze Teng Liong, John See, Raphael C.W. Phan, Yee Hui Oh, Anh Cat Le Ngo, Kok Sheik Wong, and Su Wei Tan. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Processing: Image Communication*, 2016.

[109] Sze Teng Liong, John See, Koksheik Wong, and Raphael Chung Wei Phan. Automatic micro-expression recognition from long video using a single spotted apex. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[110] Sze Teng Liong, John See, Kok Sheik Wong, and Raphael C.W. Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 2018.

[111] Sze Teng Liong, Y. S. Gan, John See, Huai Qian Khor, and Yen Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[112] Nian Liu, Xinyu Liu, Zhihao Zhang, Xueming Xu, and Tong Chen. Offset or onset frame: A multi-stream convolutional neural network with capsulenet module for micro-expression recognition. In *2020 5th international conference on intelligent informatics and biomedical sciences (ICIIBMS)*, pages 236–240. IEEE, 2020.

[113] Yong Jin Liu, Jin Kai Zhang, Wen Jing Yan, Su Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 2016.

[114] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *Proceedings of 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2019.

[115] Ling Lo, Hong Xia Xie, Hong Han Shuai, and Wen Huang Cheng. MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks. In *Proceedings of 3rd International Conference on Multimedia Information Processing and Retrieval, MIPR*, 2020.

[116] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of 5th International Conference on Learning Representations, ICLR*, 2017.

[117] Zhaoyu Lu, Ziqi Luo, Huicheng Zheng, Jikai Chen, and Weihong Li. A delaunay-based temporal coding model for micro-expression recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[118] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

[119] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021.

[120] David Matsumoto, Jeff LeRoux, Carinda Wilson-Cohn, Jake Raroque, Kristie Kooken, Paul Ekman, Nathan Yrizarry, Sherry Loewinger, Hideko Uchida, Albert Yee, et al. A new test to measure emotion recognition ability: Matsumoto and ekman's japanese and caucasian brief affect recognition test (jacbart). *Journal of Nonverbal behavior*, 24:179–209, 2000.

[121] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14:261–292, 1996.

[122] Michael W. Morris and Dacher Keltner. How emotions work: The social functions of emotional expression in negotiations. *Research in Organizational Behavior*, 2000.

[123] Roberto Munoz, Rodrigo Olivares, Carla Taramasco, Rodolfo Villarroel, Ricardo Soto, Thiago S Barcelos, Erick Merino, and María Francisca Alonso-Sánchez. Using black hole algorithm to improve eeg-based emotion recognition. *Computational intelligence and neuroscience*, 2018, 2018.

[124] Bahareh Nakisa, Mohammad Naim Rastgoo, Dian Tjondronegoro, and Vinod Chandran. Evolutionary computation algorithms for feature selection of eeg-based emotion recognition using mobile sensors. *Expert Systems with Applications*, 93:143–155, 2018.

[125] Xuan Nie, Madhumita A. Takalkar, Mengyang Duan, Haimin Zhang, and Min Xu. Geme: Dual-stream multi-task gender-based micro-expression recognition. *Neurocomputing*, 427, 2021.

[126] Yee Hui Oh, Anh Cat Le Ngo, Raphael C.W. Phari, John See, and Huo Chong Ling. Intrinsic two-dimensional local structures for micro-expression recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016.

[127] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[128] Oyebade K Oyedotun, Girum Demisse, Abd El Rahman Shabayek, Djamila Aouada, and Bjorn Ottersten. Facial expression recognition via joint deep learning of rgb-depth map latent representations. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3161–3168, 2017.

[129] Hang Pan, Lun Xie, Zeping Lv, Juan Li, and Zhiliang Wang. Hierarchical support vector machine for facial micro-expression recognition. *Multimedia Tools and Applications*, 79:31451–31465, 2020.

[130] Hang Pan, Lun Xie, Juan Li, Zeping Lv, and Zhiliang Wang. Micro-expression recognition by two-stream difference network. *IET Computer Vision*, 15(6):440–448, 2021.

[131] Pallavi Pandey and KR Seeja. Emotional state recognition with eeg signals using subject independent approach. In *Data Science and Big Data Analytics: ACM-WIR 2018*, pages 117–124. Springer, 2019.

[132] Jaak Panksepp and Douglas Watt. What is basic about basic emotions? lasting lessons from affective neuroscience. *Emotion review*, 3(4): 387–396, 2011.

[133] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[134] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. psychology press, 2001.

[135] Devangini Patel, Xiaopeng Hong, and Guoying Zhao. Selective deep features for micro-expression recognition. In *Proceedings of 23rd international conference on pattern recognition (ICPR)*, pages 2258–2263. IEEE, 2016.

[136] Christopher J Patrick. Emotion and psychopathy: Startling new insights. *Psychophysiology*, 31(4):319–330, 1994.

[137] Min Peng, Chongyang Wang, Tong Chen, Guangyuan Liu, and Xiaolan Fu. Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in Psychology*, 2017.

[138] Min Peng, Chongyang Wang, Tao Bi, Yu Shi, Xiangdong Zhou, and Tong Chen. A novel apex-time network for cross-dataset micro-expression recognition. *Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction, ACII*, 2019.

[139] Wei Peng, Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. A boost in revealing subtle facial expressions: A consolidated eulerian framework. In *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, 2019.

[140] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[141] Duc-Son Pham, Ognjen Arandjelović, and Svetha Venkatesh. Detection of dynamic background due to swaying movements from motion features. *IEEE Transactions on Image Processing*, 24(1):332–344, 2014.

[142] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[143] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *IET Seminar Digest*, 2009.

[144] Stephen Porter and Leanne Ten Brinke. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological science*, 19(5):508–514, 2008.

[145] Yifei Qian, Liangfei Zhang, Xiaopeng Hong, Carl Donovan, and Ognjen Arandjelovic. Segmentation assisted u-shaped multi-scale transformer for crowd counting. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[146] Fangbing Qu, Su Jing Wang, Wen Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 2018.

[147] Sai Prasanna Teja Reddy, Surya Teja Karri, Shiv Ram Dubey, and Snehasis Mukherjee. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, 2019.

[148] James A Russell. Culture and the categorization of emotions. *Psychological bulletin*, 110(3):426, 1991.

[149] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 406–413. IEEE, 2011.

[150] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 30(10):762–773, 2012.

[151] Stanley Schachter. The interaction of cognitive and physiological determinants of emotional state. In *Advances in experimental social psychology*, volume 1, pages 49–80. Elsevier, 1964.

[152] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. Megc 2019–the second facial micro-expressions grand challenge. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[153] Dongmin Shin, Dongil Shin, and Dongkyoo Shin. Development of emotion recognition interface using complex eeg/ecg bio-signal for interactive contents. *Multimedia Tools and Applications*, 76:11449–11470, 2017.

[154] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 2011.

[155] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny P L Lo. Revisiting self-supervised contrastive learning for facial expression recognition. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[156] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[157] Ruchir Srivastava and Sujoy Roy. Utilizing 3d flow of points for facial expression recognition. *Multimedia tools and applications*, 71: 1953–1974, 2014.

[158] Wenchao Su, Yanyan Wang, Fei Su, and Zhicheng Zhao. Micro-expression recognition based on the spatio-temporal feature. In *2018 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2018*. Institute of Electrical and Electronics Engineers Inc., 2018.

[159] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. Dynamic micro-expression recognition using knowledge distillation. *IEEE Transactions on Affective Computing*, 2020.

[160] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[161] Yi Sun and Lijun Yin. Facial expression recognition based on 3d dynamic range model sequences. In *ECCV (2)*, pages 58–71, 2008.

[162] Sima Taheri, Qiang Qiu, and Rama Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Transactions on Image Processing*, 23(8):3590–3603, 2014.

[163] Madhumita A Takalkar, Min Xu, and Zenon Chaczko. Manifold feature integration for micro-expression recognition. *Multimedia Systems*, 26:535–551, 2020.

[164] Madhumita A Takalkar, Selvarajah Thuseethan, Sutharshan Rajasegarar, Zenon Chaczko, Min Xu, and John Yearwood. Lgattnet: Automatic micro-expression detection using dual-stream local and global attentions. *Knowledge-Based Systems*, 212:106566, 2021.

[165] B. M.S.Bahar Talukder, Brinta Chowdhury, Tamanna Howlader, and S. M.Mahbubur Rahman. Intelligent recognition of spontaneous expression using motion magnification of spatio-temporal data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[166] Silvan Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962.

[167] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[168] Jessica L Tracy and Daniel Randles. Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt. *Emotion review*, 3(4):397–405, 2011.

[169] Filareti Tsalakanidou and Sotiris Malassiotis. Robust facial action recognition from real-time 3d streams. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 4–11. IEEE, 2009.

[170] Filareti Tsalakanidou and Sotiris Malassiotis. Real-time 2d+ 3d facial action and expression recognition. *Pattern Recognition*, 43(5): 1763–1775, 2010.

[171] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE transactions on knowledge and data engineering*, 23(7):1079–1089, 2010.

[172] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *Proceedings of 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2019.

[173] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

[174] Gyanendra K Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102:162–172, 2014.

[175] Monu Verma, Santosh Kumar Vipparthi, and Girdhari Singh. Affectivenet: Affective-motion feature learning for microexpression recognition. *IEEE MultiMedia*, 28(1):17–27, 2020.

[176] Chongyang Wang, Min Peng, Tao Bi, and Tong Chen. Micro-attention for micro-expression recognition. *Neurocomputing*, 2020.

[177] Fushun Wang and Alfredo Pereira. Neuromodulation, emotional feelings and affective disorders. *Mens sana monographs*, 14(1):5, 2016.

[178] Jie Wang, Xiao Pan, Xinyu Li, Guangshun Wei, and Yuanfeng Zhou. Single trunk multi-scale network for micro-expression recognition. *Graphics and Visual Computing*, 4:200026, 2021.

[179] Lin Wang, Jingqian Jia, and Nannan Mao. Micro-expression recognition based on 2d-3d cnn. In *2020 39th Chinese control conference (CCC)*, pages 3152–3157. IEEE, 2020.

[180] Mingzhong Wang, Qi Wang, Qingshan Wang, and Zhiwen Zheng. A fixed-point rotation-based feature selection method for micro-expression recognition. *Pattern Recognition Letters*, 164:261–267, 2022.

[181] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Proceedings of 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822. Association for Computational Linguistics, 2019.

[182] Shuai Wang, Jiachen Du, and Ruifeng Xu. Decision fusion for eeg-based emotion recognition. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 883–889. IEEE, 2015.

[183] Su Jing Wang, Hui Ling Chen, Wen Jing Yan, Yu Hsin Chen, and Xiaolan Fu. Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural Processing Letters*, 2014.

[184] Su Jing Wang, Wen Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In *Proceedings - International Conference on Pattern Recognition*, 2014.

[185] Su-Jing Wang, Wen-Jing Yan, Guoying Zhao, Xiaolan Fu, and Chun-Guang Zhou. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In *European Conference on Computer Vision*, volume 8925, pages 325–338, 2014.

[186] Su Jing Wang, Wen Jing Yan, Xiaobai Li, Guoying Zhao, Chun Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. *IEEE Transactions on Image Processing*, 2015.

[187] Su Jing Wang, Wen Jing Yan, Tingkai Sun, Guoying Zhao, and Xiaolan Fu. Sparse tensor canonical correlation analysis for micro-expression recognition. *Neurocomputing*, 2016.

[188] Su Jing Wang, Bing Jun Li, Yong Jin Liu, Wen Jing Yan, Xinyu Ou, Xiaohua Huang, Feng Xu, and Xiaolan Fu. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing*, 312, 2018.

[189] Su-Jing Wang, Ying He, Jingting Li, and Xiaolan Fu. Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos. *IEEE Transactions on Image Processing*, 30:3956–3969, 2021.

[190] Yan Wang, Yikun Huang, Can Liu, Xiaoying Gu, Dandan Yang, Shuopeng Wang, and Bo Zhang. Micro expression recognition via dual-stream spatiotemporal attention network. *Journal of Healthcare Engineering*, 2021, 2021.

[191] Yandan Wang, John See, Raphael C.W. Phan, and Yee Hui Oh. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PLoS ONE*, 2015.

[192] Yandan Wang, John See, R. Raphael, and Yee Hui Oh. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[193] Yandan Wang, John See, Yee Hui Oh, Raphael C.W. Phan, Yogachandran Rahulamathavan, Huo Chong Ling, Su Wei Tan, and Xujie Li. Effective recognition of facial micro-expressions with video motion magnification. *Multimedia Tools and Applications*, 2017.

[194] Yiting Wang, Wei-Bang Jiang, Rui Li, and Bao-Liang Lu. Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1575–1578. IEEE, 2021.

[195] Gemma Warren, Elizabeth Schertler, and Peter Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 2009.

[196] Jinsheng Wei, Guanming Lu, Jingjie Yan, and Huaming Liu. Micro-expression recognition using local binary pattern from five intersecting planes. *Multimedia Tools and Applications*, 81(15):20643–20668, 2022.

[197] Mengting Wei, Wenming Zheng, Yuan Zong, Xingxun Jiang, Cheng Lu, and Jiateng Liu. A novel micro-expression recognition approach using attention-based magnification-adaptive networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2420–2424. IEEE, 2022.

[198] Jacob Whitehill, Zewelanji Serpell, Yi Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 2014.

[199] Zhaoqiang Xia, Xiaoyi Feng, Xiaopeng Hong, and Guoying Zhao. Spontaneous facial micro-expression recognition via deep convolutional network. In *Proceedings of 8th International Conference on Image Processing Theory, Tools and Applications, IPTA*. Institute of Electrical and Electronics Engineers Inc., 2019.

[200] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia*, 22(3), 2020.

[201] Zhaoqiang Xia, Wei Peng, Huai Qian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29, 2020.

[202] Fu Xiaofeng and Wei Wei. Centralized binary patterns embedded with image euclidean distance for facial expression recognition. In *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, 2008.

[203] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[204] Feng Xu, Junping Zhang, and James Z. Wang. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing*, 2017.

[205] Wen Jing Yan, Qi Wu, Yong Jin Liu, Su Jing Wang, and Xiaolan Fu. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

[206] Wen Jing Yan, Xiaobai Li, Su Jing Wang, Guoying Zhao, Yong Jin Liu, Yu Hsin Chen, and Xiaolan Fu. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE*, 2014.

[207] Bing Yang, Jing Cheng, Yunxiang Yang, Bo Zhang, and Jianxin Li. Merta: micro-expression recognition with ternary attentions. *Multimedia Tools and Applications*, 80:1–16, 2021.

[208] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[209] Huiyuan Yang and Lijun Yin. Cnn based 3d facial expression recognition using masking and landmark features. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, pages 556–560. IEEE, 2017.

[210] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10494–10503, 2019.

[211] Lijun Yin, Xiaozhou Wei, Peter Longo, and Abhinesh Bhuvanesh. Analyzing facial expressions using intensity-variant 3d data for human computer interaction. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1248–1251. IEEE, 2006.

[212] Liangfei Zhang and Ognjen Arandjelović. Review of automatic micro-expression recognition in the past decade. *Machine Learning and Knowledge Extraction*, 3(2):414–434, 2021.

[213] Liangfei Zhang, Ognjen Arandjelović, Sonia Dewar, Arlene Astell, Gayle Doherty, and Maggie Ellis. Quantification of advanced dementia patients' engagement in therapeutic sessions: An automatic video based approach using computer vision and machine learning. In *Proceedings International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 5785–5788. IEEE, 2020.

[214] Liangfei Zhang, Ognjen Arandjelović, and Xiaopeng Hong. Facial action unit detection with local key facial sub-region based multi-label classification for micro-expression analysis. In *ACM international conference on Multimedia Workshops*, 2021.

[215] Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelović, and Guoying Zhao. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1973–1985, 2022.

[216] Shiyu Zhang, Bailan Feng, Zhineng Chen, and Xiangsheng Huang. Micro-expression recognition by aggregating local spatio-temporal patterns. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[217] Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[218] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[219] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946, 2016.

[220] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[221] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 2090–2099, 2018.

[222] Sirui Zhao, Hanqing Tao, Yangsong Zhang, Tong Xu, Kun Zhang, Zhongkai Hao, and Enhong Chen. A two-stage 3d cnn based learning method for spontaneous micro-expression recognition. *Neurocomputing*, 448:276–289, 2021.

[223] Sirui Zhao, Huaying Tang, Shifeng Liu, Yangsong Zhang, Hao Wang, Tong Xu, Enhong Chen, and Cuntai Guan. Me-plan: A deep prototypical learning with local attention network for dynamic micro-expression recognition. *Neural Networks*, 153:427–443, 2022.

[224] Wei Zhao, Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition by combining progressive-learning intensity magnification with self-attention-convolution classification. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022.

[225] Xi Zhao, Di Huang, Emmanuel Dellandréa, and Liming Chen. Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model. In *2010 20th International Conference on Pattern Recognition*, pages 3724–3727. IEEE, 2010.

[226] Yuan Zhao, Zhuang Chen, and Song Luo. Micro-expression recognition based on pixel residual sum and cropped gaussian pyramid. *Frontiers in Neurorobotics*, 15:746985, 2021.

[227] Yue Zhao and Jiancheng Xu. An improved micro-expression recognition method based on necessary morphological patches. *Symmetry*, 2019.

[228] Qingkai Zhen, Di Huang, Yunhong Wang, and Liming Chen. Muscular movement model-based automatic 3d/4d facial expression recognition. *IEEE Transactions on Multimedia*, 18(7):1438–1450, 2016.

[229] Hao Zheng, Xin Geng, and Zhongxue Yang. A relaxed K-SVD algorithm for spontaneous micro-expression recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[230] Zheng Zheng, Simeng Gu, Yu Lei, Shanshan Lu, Wei Wang, Yang Li, and Fushun Wang. Safety needs mediate stressful events induced mental disorders. *Neural plasticity*, 2016, 2016.

[231] Huang Zhewei, Zhang Tianyuan, Heng Wen, Shi Boxin, and Zhou Shuchang. RIFE: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294*, 2020.

[232] Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N Metaxas. Learning multiscale active facial patches for expression analysis. *IEEE transactions on cybernetics*, 45(8):1499–1510, 2014.

[233] Ling Zhou, Qirong Mao, and Luoyang Xue. Dual-inception network for cross-database micro-expression recognition. In *Proceedings of 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG*. Institute of Electrical and Electronics Engineers Inc., 2019.

[234] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022.

[235] Kangkang Zhu, Zhengyin Du, Weixin Li, Di Huang, Yunhong Wang, and Liming Chen. Discriminative attention-based convolutional neural network for 3d facial expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[236] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[237] Xuena Zhu, Xianye Ben, Shigang Liu, Rui Yan, and Weixiao Meng. Coupled source domain targetized with updating tag vectors for micro-expression recognition. *Multimedia Tools and Applications*, 2018.

[238] Yuan Zong, Xiaohua Huang, Wenming Zheng, Zhen Cui, and Guoying Zhao. Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. *IEEE Transactions on Multimedia*, 2018.