C02047: Doctor of Philosophy
CRICOS Code: 0000
Subject Code: 33875
June 2020

Deciphering True Emotions:
Micro-Expression Detection and Recognition using Deep Nets

Madhumita Abhijeet Takalkar

School of Electrical and Data Engineering
Faculty of Engg. & IT
University of Technology Sydney
NSW - 2007, Australia

Deciphering True Emotions:
Micro-Expression Detection and Recognition using Deep Nets

A dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Computer Systems

by

Madhumita Abhijeet Takalkar

under the supervision of

A/Prof. Dr. Min Xu and **Dr. Zenon Chaczko**

to

School of Electrical and Data Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

June 2020

Certificate of Original Authorship

I, *Madhumita Abhijeet Takalkar* declare that this thesis is submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Electrical and Data Engineering*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: _____

[Madhumita Abhijeet Takalkar]

DATE: 14 th September, 2020

PLACE: Sydney, Australia

Abstract

Micro-expressions are anticipated as the outcome of deliberate manipulation or involuntary repression of emotions when an individual feels emotion but tries to conceal the facial movements. The micro-expression interpretation tends to recognise a person's deceit and actual mental state. Therefore, micro-expression detection and recognition has significant opportunities for emotion analysis in psychotherapy, forensics, border protection, and negotiations, among others. Since such gestures are quick and hard to spot with the naked eyes, the inclination towards automated micro-expression recognition is an obvious step forward in the domain. Micro-expression research has drawn various interests within the computer vision field notable in localisation, magnification and recognition. Earlier studies primarily implemented single handcraft descriptors and classifiers for recognising micro-expressions. Modern techniques emphasise on deploying Convolutional Neural Networks (CNNs) or hybrid strategies that integrate handcraft descriptors and CNNs. Owing to the existence of a few datasets, the recognition of micro-expressions is still a concern. Nevertheless, efficiency is often influenced by the feature selection and training approach.

Our work, presented in this thesis, introduces various approaches that we have developed to detect and recognise facial micro-expressions using deep networks. In the initial stages of this work we design a dual-stream model with attention networks for the task of micro-expression detection from images. We implement Local- and Global-level Attention Networks (LGAttNet) to concentrate on local facial regions as well as full face to boost the chances of extracting relevant micro-expression features. Unlike previous detection methods where frame difference is calculated to detect micro-expressions, our framework uses attention network to focus on various parts of a face to identify the presence of the micro-expression. We developed LGAttNet to be a supervised detection framework where a traditional Artificial Neural Network (ANN) is trained as a binary classifier. LGAttNet is a novel documented approach that utilises attention network for micro-expression detection from image and video frame sequence.

The next stage of this thesis focuses on recognising micro-expression from an image using CNN. We propose to implement a CNN network by performing fine-tuning on a pre-trained CNN network. Fine-tuning is carried out to retrain the last convolutional layer of the CNN network to be able to learn appropriate micro-expression features and predict the micro-expression classes accurately. This fine-tuned CNN network gained acceptable accuracy for recognising micro-expressions from image frames. Thirdly, we extend the outcome of this stage to be implemented on video data; hence we explore the

approach of combining handcrafted descriptors with the CNN derived features. Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) and VGGFace CNN network are combined in late fusion technique to extract a comprehensive feature representation of the video. Softmax and SVM are trained for classification. The employed hybrid approach is one of the first attempts to implement handcrafted descriptors and deep features for micro-expression recognition.

Finally, we consider the factor of gender affecting the tendency to express micro-expressions. We have built a multi-task learning architecture with two streams extracting different features to achieve the same task of micro-expression recognition based on gender, GEME. We incorporated a dynamic image concept to convert a video into a single frame, and gender features and micro-expression features are added at each level and given to the micro-expression stream. Inclusion of the gender features with the micro-expression features elevates the feature details respective to the individual participant, and the network learns unique gender features while extracting micro-expression features.

Concisely, we have introduced four novel concepts for micro-expression detection and recognition. The work described in this thesis establishes a connection between computer vision and psychotherapy, and aids to expedite the micro-expression analysis process for quick assessment wherever necessary.

Dedication

In memory of my father, to whom I promised to dedicate this thesis before he left this world...

Acknowledgments

Pursuing this PhD has been a truly life-changing journey for me, and I am blessed with all the encouragement and help from so many people without whom I could not have achieved it.

At this moment of accomplishment, I would like to express my most profound appreciation to my two supervisors, principal supervisor **Dr. Min Xu** and co-supervisor **Dr. Zenon Chaczko**, for their constant guidance and motivation they provided me during my PhD study. This PhD would not have been attainable without their continuous input and support. Dr. Xu has, knowingly and implicitly, taught me how to evaluate my work in order to enhance it and develop it further. Dr. Chaczko was able to direct me in making my work a more realistic solution. Their experience and advice rendered me capable of working in both computational and application-oriented areas. I am thankful for their persistence, time, suggestions, inspiration and tremendous expertise that they shared with me. I am very grateful to them for allowing me the opportunity to determine my area of interest and to encourage me to focus on it.

Besides, I would like to acknowledge **Mengyang Duan** and **Selvarajah Thuseethan** for their helpful advice and collaborations. Collaborating alongside them for two projects has been an outstanding experience. It was a privilege to discuss the research with them, and I was pleased to understand various facets of critical methodology and reflective analysis from them. My sincere thanks to **Dr. Wenjing Jia, Dr. Qiang Wu** and **Dr. Guoqiang Zhang**, who acted as panel members for my candidature assessments.

It's my fortune to gratefully acknowledge by bachelor's supervisor, **Asst. Prof. Yogita Pagar**, who introduced me into the research community and encouraged and mentored me in my first research project. Her charismatics and determination to pursue goals have always inspired me.

There is no pleasure to have a ride without friends. My bumpy research ride at UTS was made easy by my fellow labmates, **Haimin Zhang, Zhiyuan Shi (Zhi), Lingxiang Wu (Lynn), Ruiheng Zhang, Zhongqin Wang, Tianrong Rao (Ron), Xiaoxu Li (Sam), and everyone in the team**. I will always recall the fun we had together and the kindness and support everyone showed in tough times. Thanks in particular to **Haimin Zhang** for sharing his profound expertise and technical cooperation. I cannot forget my crazy friends from UTS **Dr. Ashish Nanda, Alina Rakhi, Manisha Pratihast, Sara Farahmandian** with whom I had been enjoying my research tenure and who became a part of my life. And a huge thanks to all my friends back home who in the good and rough moments were beside me to drive me and inspire me.

I owe a great deal to the most beautiful couple I have ever met, **Mummy** and **Baba**, **Smt. Veena** and **Late Shri. Ramesh**, for all the selfless love, care, pain and sacrifice they have done to make me who I am today. They have taught me, *'he that ventures not fails not'* and made me worthy of facing the world. It was my father's dream to see me with a doctorate degree before he was separated from us last year owing to a severe cardiac arrest. This thesis is a tribute to my late father under whose diligent protection I have enjoyed my life. Special gratitude to my Mummy and little brother **Neeraj** for having always trusted me and inspired me to pursue my dreams. My deepest reverence goes to my **father-in-law** for his moral encouragement and to all my in-laws for supporting in any way they can through this challenging time.

Last but definitely not the least, I owe it to an exceptional person, my better half, **Abhijeet**, who lived by my side every moment of my doctorate, and without whom I would not have had the confidence to embark on this journey in the first place. His ongoing and unfailing love, encouragement and appreciation during my pursuit of PhD degree has made it possible to complete the thesis. At times, when I felt it was hard to keep moving, you have always helped me to keep things in perspective. I immensely respect his effort and truly admire his belief in me.

This thesis is dedicated to my parents and husband for their affection and trust in me. Without them, I would be nothing.

Thank you.

List of Publications

Related to the Thesis :

1. **Madhumita A. Takalkar**, Min Xu, Qiang Wu, Zenon Chaczko, *A survey: facial micro-expression recognition [J]*, Multimedia Tools and Applications (MTA) 2017. (***Published***)

2. **Madhumita A. Takalkar**, Min Xu, *Image based facial micro-expression recognition using deep learning on small datasets [C]*, In 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE. (***Published***)

3. **Madhumita A. Takalkar**, Haimin Zhang, Min Xu, *Improving Micro-expression Recognition Accuracy Using Twofold Feature Extraction [C]*, In 2019 International Conference on Multimedia Modeling, Springer, Cham. (***Published***)

4. **Madhumita A. Takalkar**, Min Xu, Zenon Chaczko, *Manifold Feature Integration for Micro-Expression Recognition [J]*, Multimedia Systems (MS) 2020. (***Published***)

5. **Madhumita A. Takalkar**, Selvarajah Thuseethan, Sutharshan Rajasegarar, Zenon Chaczko, Min Xu, John Yearwood, *LGAttNet: Automatic Micro-expression Detection using Dual-Stream Local and Global Attentions [J]*, Knowledge-Based Systems (KBS). (***Under Review***)

6. Xuan Nie, **Madhumita A. Takalkar**, Mengyang Duan, Haimin Zhang, Min Xu, *GEME: dual-stream multi-task GEnder-based Micro-Expression recognition*, Neurocomputing. (***Under Review***)

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Background

Facial gestures perform a crucial part in interacting with others throughout our everyday lives. By means of facial expressions, people communicate their emotions as well as interpret the sentiments of others.

Typically, facial expressions are loosely classified as macro- expressions and micro-expressions. Unlike the long-lasting and highly perceptible macro-expressions, micro-expressions may be characterised by quick transitions of less than half a second with low intensity on the facial regions.

Ekman and Frisen [43] introduced the theory of "micro-expressions", and researchers universally endorsed it. Figure 1.1 illustrates seven universal micro-expressions and action units (AU) corresponding to each face muscle movement. Micro-expressions (MEs) can be, commonly, seen in cases when individuals try to regulate or suppress their feelings. There are three fundamental approaches to deceive facial expressions [45]:(1) **Stimulated expressions:** an expression is induced to convey a feeling with expression while experiencing nothing; (2) **Neutralised expressions:** showing to be a neutral face, though having a particular feeling; (3) **Masked expressions:** masked in such a way that the perceived emotion is hidden with another expression. Micro-expression is the outcome of the disturbance of the facial muscles that appears when a perceived facial emotion understating reactions, neutralises or hides [45].

For certain people, the primary aim of lying to improve social relations and obtain

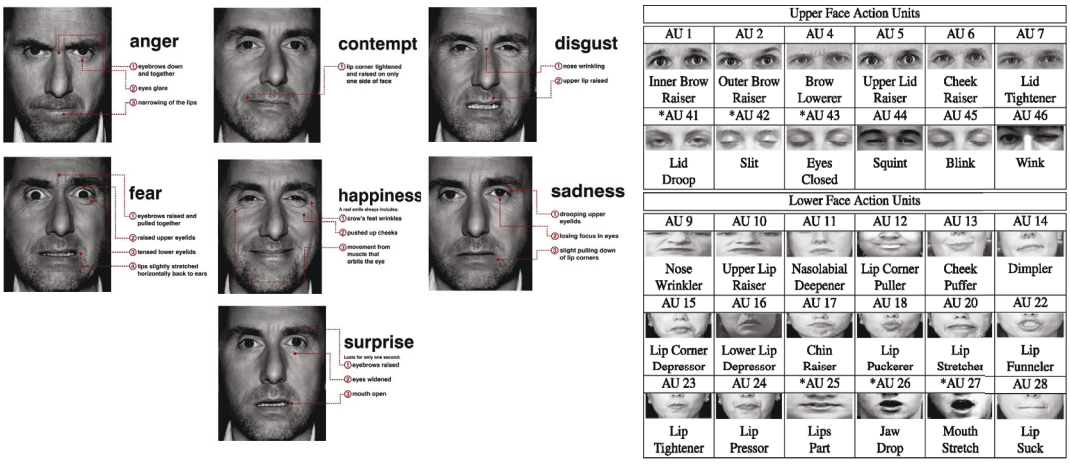


Figure 1.1: Seven universal facial micro-expressions

other people's; respect and affection. Nevertheless, detecting lies with aggressive intent can be critical. Liars cannot conceal facial movements entirely showing a disclosure or deceit as depicted by the presence of micro-expressions. Such characteristics encourage the prospective usage of micro-expressions in (1) high-risk scenarios such as crime investigations, airport and group transit control points, counter-terrorist activities; (2) industries like marketing, teaching, consulting, administration, leadership, trade negotiations; (3) medical services, such as physician-patient consultation. A few professionals have developed the technique to catch micro-expressions with bare eyes given the attempts made by Ekman [41] educating people using the micro-expression training tool. In recent years, the necessity for automated micro-expression recognition has risen immensely, owing to the challenges faced due to manual micro-expression detection. While computer vision technology has evolved extensively, the study of micro-expressions remains a challenge due to its brief and low-intensity properties.

Micro-expression research aims to identify and recognise micro-expressions. Over the past decade, a broad spectrum of methods to identify and recognise micro-expressions have been used.

1.2 Factors Influencing Recognition of Micro-Expressions

Micro-expression is contained in the flow of expressions when individuals are trying to repress their emotions. According to studies in [180, 239], it is observed that certain factors affect recognition of micro-expressions.

1.2.1 Emotional Context

Existing studies have employed neutral expressions before and after the emotion. The study suggests that micro-expressions may be encapsulated either in neutral expressions or in certain facial expressions accompanying emotions such as sadness and happiness. According to the theory of emotional regulation, in the priming task, primes presented for longer duration may lead to greater priming effect. Moreover, it is observed that emotional information influences attention [239]. The aims of this research are:

* to examine the effect of emotional context on micro-expressions;
* to explore if the effect of the context was limited to a particular material, and
* to investigate the reason of the effect.

The findings will lead researchers to predict that the emotional context would indeed influence micro-expression recognition.

1.2.2 Duration of Expression

The significant difference between a micro- and a macro- expression is the length for which the expression lasts. Many specific assessments of a micro-expression's length have been developed. Therefore, there is also a difference of agreement as to the time span of the micro-expression duration. Although the difference in duration might not be significantly noticeable, for micro-expressions it needs to be taken into account.

To verify the effect of duration on micro-expressions recognition, researchers conducted two experiments [180] asking the participants to recognise the micro-expressions in the images shown to them. In Experiment 1 expression images were shown to participants for 40,120,200 or 300 ms. The researchers employed Brief Affect Recognition Test (BART) for Experiment 1. In Experiment 2 the participants were given the micro-expression recognition training using the Micro Expression Training Tool (METT) [41] paradigm which played a significant role in recognition of the micro-expressions. The outcome of the experiments indicated that the participants could recognise the micro-expression in the images in 200 ms without training and 160 ms after training. The results suggest that the critical time point that differentiates micro-expressions was about 200 ms or less. Thus, in conclusion, the accuracy of the micro-expression recognition is a function of the duration of the expressions.

1.3 Challenges

1.3.1 Environmental factor

Environmental variability in the study of micro-expressions is the most intractable problem which entails variability in the light as well as head position. The majority of the features of illumination variability highly rely on the pixel intensity shifts. Dynamic lighting conditions could be triggering inaccurate feature assessment, head motion or adjustments in the head location, which may be mistaken as micro-expression. A slight head motion greatly influences the shifts of perceived expressions, thereby reducing detection precision.

1.3.2 Spontaneous and subtle motion of the facial movement

A significant obstacle for micro-expression recognition is a low intensity subtle and involuntary facial movements that makes detection of emotions not identifiable via the naked eyes. For certain instances, the classifier might even confuse motions wrongly as a neutral face. Techniques for magnifying and intensifying subtle emotions are thus essential at the pre-processing level.

1.3.3 Imbalanced dataset

SMIC, CASME, CASME II, and CAS(ME)^2 are some of the limited publicly accessible datasets concentrating on micro-expression recognition. Despite being recommended for assessing micro-expression recognition framework, the uneven distribution of samples through expressions can contribute to the biases on performance. Besides, the captured samples for the available datasets are typically obtained in regulated settings with even lighting and fixed locations. Therefore, current state-of-the-art algorithms validated on these datasets might not be ideal in real-world and raises a requirement for more real-world based datasets.

1.4 Research Questions and Hypothesis

We aim to propose the best possible solutions through our research to the problems we identified as necessary to be addressed. Our research questions addressed in this study are as follows:

$Q1$: What information about micro-expressions does an image provides?

H_1: *An image can provide the information related to the presence of micro-expression in the image.*

$Q2$: As in facial macro-expressions, can we recognise facial micro-expressions from an image?

H_1: *If we can extract accurate facial movement details, then we will be able to use an image for recognising micro-expressions.*

$Q3$: How can we improve micro-expression recognition accuracy in videos?

H_1: *If we can identify different feature descriptors and combine them, then it can lead to better recognition accuracy.*

$Q4$: Which human characteristic affects the way micro-expressions are expressed, contributing to improving micro-expression recognition accuracy?

H_1: *If we can extract the additional facial properties unique to each individual, then we will be able to improvise micro-expression recognition accuracy.*

1.5 Influence of Micro-Expressions on different domains

We frequently encounter a query as to how the people interpret “micro-expressions” and their importance in our general perception of body language, and most specifically, their applicability in deceit detection. Decades of studies indicate that training in emotion recognition will boost and reinforce the capacity to discern whether someone is dishonest, understand how other people feel and enable you to consider the effect of your actions on others. Nowadays, micro-expressions are adopted as subjects and are granted separate courses in various institutes and schools. Numerous professions where human intercommunication is routinely required, the student must study micro-expressions in order to train themselves for varied career directions. Micro-expressions being ubiquitous and hard to deceive, they play a vital position in the analysis of deception and criminal investigations [219]. Recognition of micro-expressions may be used to identify harmful atrocities [219]. Micro-expressions can also aid in interacting and recognising the purpose of others in several areas, such as business, medicine, law and national security

[78]. Identification of micro-expressions provides a wide variety of possible uses spanning from understanding consumer responses to diverse circumstances and advertisement strategies [220] to the intercession of terrorist attacks [219]. It can also be adopted to determine the probability of an individual harming their children [6].

1.5.1 Social Impact

For those engaged in law enforcement, sales and marketing, recruitment, business and negotiations, leadership, coaching and teaching, micro-expressions provide a crucial edge.

The FBI and CIA in the U.S.A. and other ***law enforcement agencies*** around the globe have adopted methods to train and enable law enforcement officers to detect micro-expressions. The identification of terrorists in line at the airport is only one serious application of the potential to detect micro-expressions [51, 219]. Individuals who are skilled ***liars*** seek to intimidate us with false news but relinquishing their dishonest act with their micro-expressions. Professional psychologists and prosecutors globally adopt the technique of rapid-fire questions to apprehend liars [42, 52].

Business executives and delegates will foster mutually valuable alliances if they are capable of interpreting others thoughts and feelings [220]. Learning what a micro-expressions is, as a ***salesperson***, acts as the first move towards a more nuanced approach to better "interpret" your consumers and prospective clients. In order to be efficient, the vendor must follow a personalised approach and be vigilant and attentive to the consumer by observing their facial expressions to realise the way a consumer perceives the conversation. ***Product analysts*** may enhance the qualitative feedback collected from the customers by interpreting customer feelings while testing products, providing clues of what they believe about what they claim.

Employers usually discover soon after ***recruiting*** those with outstanding credentials that they are toxic to the workplace and that they are annoying, egotistic, seeking personal prestige over team goals. It would be immensely beneficial when recruiting individuals, not just for work but elected positions, jury duties, or several other situations where it is required to dig deep into someone being assessed [138].

Teachers should interpret their students' thoughts to gain insight into how their lesson plans are progressing and that they can adapt better and execute them efficiently. Researchers also attempted to analyse the ***pre-schoolers'*** micro-expression recognition ability that could enable them to grasp adult interactions and properly respond to emerging situations and experiences [168]. ***School managers*** who could understand

their teachers' feelings will be able to scale down the burden consequently sustaining and enhancing teacher performance.

Parents, partners, friends and anyone willing to maintain positive and constructive relationships will profit from enhancing their capacity to read emotions [6, 42].

1.5.2 Scientific Impact

The identification of micro-expressions facilitates an exposure into a very intimate domain of life: the emotions that they do not want others to realise they are having. Despite being an intrusion in privacy, it may serve the interest of the public. This allows the doctors, nurse or health care professionals to adapt for improved treatment.

In a recorded interview in 1969, Ekman and Friesen [43] discovered that a psychologically unstable patient was attempting to conceal a deep depressive sentiment from her therapist to reassure that she was not suicidal anymore. As the interview footage was viewed in slow motion, only two frames were seen revealing the patient's depressed face accompanied by a fake grin for a longer time. Such facial expressions are termed as micro-expressions and Haggard and Isaacs [62] first identified them three years before the incident. Haggard and Isaacs clarified how, throughout their research, such "micromomentary" gestures were witnessed when analysing hours of videos of psychotherapy sessions, aiming for signs of non-verbal conversation between the patient and psychiatrist.

Medical practitioners can strengthen interactions with patients, compassionately communicate with empathy and concern, and gathering undeniable evidence to guarantee the right diagnosis. The studies involving patients with ***Schizophrenia*** [175, 241] illustrates that those qualified to interpret micro-expressions are more likely to identify others' feelings. A pilot study [47], emphasising on facial micro-expressions as non-verbal indications, was carried out. This research analysed first-year medical students' ability to perceive emotional micro-expressions as strong or weak communicators before and after the METT [39] training. The research demonstrated that the learning component of METT supported certain students who exhibited reliable professional contact while under-performing students did not benefit.

Ekman initiated one of the very first attempts to enhance the human capacity to micro-expressions, where he established the METT to teach people to identify seven micro-expression classes [41]. Frank et al. [51], however, noticed that the efficiency of micro-expression identification by undergraduate students only exceeded a high of 40% with the aid of METT, whereas unassisted U.S. coastguards achieved no more than

50%. There is, therefore, a strong demand for an automated ME recognition method, particularly with recent developments in computing capacity and multi-core parallel functionalities, to assist in catching micro-expressions involved in lies and alarming behaviours. Researchers, today, have extended past psychology to use of computer vision and video processing to simplify the identification of micro-expressions.

1.6 Contributions and Outline

This thesis introduces four approaches for micro-expression detection and recognition.

The traditional pipeline of processes followed for micro-expression analysis comprises of micro-expression detection, feature extraction and recognition. A similar sequence of operations using proposed techniques are depicted in the following chapters of this thesis. Any video to be annotated as one of the emotion categories, it is essential to detect that the video contains some facial movements that resemble the micro-expressions. Chapter 3 introduces a technique to detect such facial signs from video frames. Upon detection, the relevant facial features are extracted to classify the micro-expression into its correct emotion category. In Chapter 4, micro-expression related spatial and spatio-temporal facial features are extracted from images and sequence of images, respectively to recognise the emotion class. Apart from the micro-expression specific features, and the additional human-specific element is extracted that is likely to impact the way micro-expressions are expressed. Chapter 5 demonstrates the impact of gender on micro-expression recognition. Chapters 4 and 5 illustrate different techniques of feature extraction and introduce a new feature for micro-expression recognition. Each of the following chapters is a part of the micro-expression analysis pipeline, and each block in this analysis pipeline plays a vital role. The workflow depicted in Figure 1.2 explains relation between the proposed methods and that without any one of these blocks, the micro-expression analysis is incomplete.

The main contributions of the thesis are outlined in chapters as follows:

* **Chapter 2** introduces the background: an extensive study of the databases, the state-of-the-art features for micro-expression detection and recognition. This chapter introduces several fundamental feature descriptors both handcrafted and deep learning as well as hybrid approaches and classifiers used in the micro-expression analysis. We have also listed and explained widely used benchmark micro-expression databases. Some content of this chapter is published in MTA

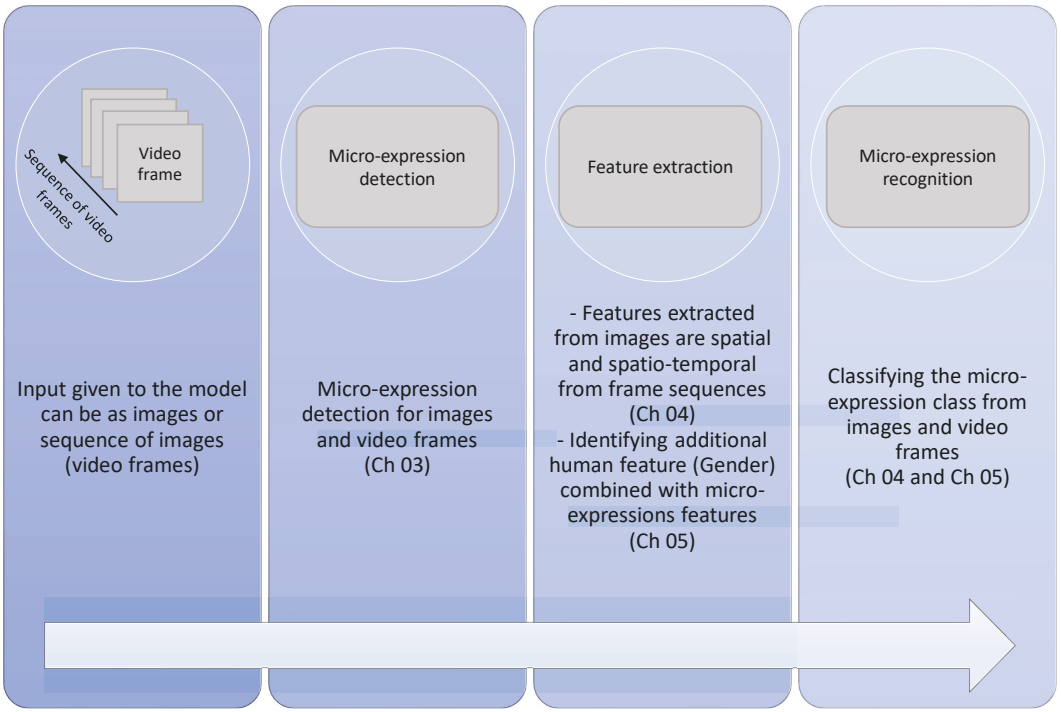


Figure 1.2: Relation between proposed facial micro-expression analysis methods

2017[1].

* **Chapter 3** explores a novel idea of detecting micro-expressions from images. The existing methods used feature differencing or frame differencing to detect the presence of micro-expressions in the videos. In this chapter, we propose a supervised framework of deep network using attention network, called LGAttNet, to detect if the frame contains micro-expression or not. Three attention networks are fed with the feature maps from different parts of the face as well as the whole facial region to extract local and global attention feature maps, respectively. The output from these attention networks is added and passed to a standard ANN. This ANN is used to perform binary classification to output if the class of the frame is a ME class or non-ME. The content presented in this chapter is based on the work submitted to Knowledge-Based Systems[2].

* **Chapter 4** presents our two new approaches for micro-expression recognition

[1]Madhumita A. Takalkar, Min Xu, Qiang Wu, Zenon Chaczko, *A survey: facial micro-expression recognition*, Multimedia Tools and Applications (MTA) 2017.

[2]Madhumita A. Takalkar, Selvarajah Thuseethan, Sutharshan Rajasegarar, Zenon Chaczko, Min Xu, John Yearwood, *LGAtteNet: Automatic Micro-expression Detection using Dual-Stream Local and Global Attentions*, Knowledge-Based Systems (KBS).

based on deep learning. The first method deals with developing a system to work with static images for recognition of micro-expression using convolutional neural networks (CNN). An extension to this work is the second method that uses a video frame sequence for micro-expression recognition. The principal idea of this approach is to identify additional features contributing to the improvement of the accuracy results. Hence in this chapter, the second approach is the combination of handcrafted LBP-TOP features with deep CNN features. The classification is carried out using Softmax and SVM with different kernels to identify the best suitable classifier for the model. The content of this chapter is published in DICTA 2017[3], MMM 2019[4] and some sections are based on the work published in MS 2020[5].

* **Chapter 5** originates the adoption of a distinctly human characteristic: gender. We claim that gender influences the way every individual shows the micro-expressions. The experiments presented in this chapter confirm our hypothesis. We have built a dual-stream CNN framework called GEME with one stream classifying gender and the other classifying micro-expression. This framework uses a concept of a dynamic image which is a single image depicting a micro-expression video. The gender features extracted from one stream are added with the micro-expression features at each level, and the new feature map is given to the next block of the micro-expression stream. This way, we are utilising the gender characteristics to train the network for unique features distinct to each gender. Incorporating the gender features also aided to improve micro-expression recognition accuracy. The content of this chapter is based on the work submitted to Neurocomputing [6].

* **Chapter 6** summarises the key novel contributions of this thesis. The chapter also highlights some unsettled and open issues generally faced in micro-expression analysis and gives the future perspective work based on our thesis.

[3]Madhumita A. Takalkar, Min Xu, *Image based facial micro-expression recognition using deep learning on small datasets*, In 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE.

[4]Madhumita A. Takalkar, Haimin Zhang, Min Xu, *Improving Micro-expression Recognition Accuracy Using Twofold Feature Extraction*, In 2019 International Conference on Multimedia Modeling, Springer, Cham.

[5]Madhumita A. Takalkar, Min Xu, Zenon Chaczko, *Manifold Feature Integration for Micro-Expression Recognition*, Multimedia Systems (MS) 2020.

[6]Xuan Nie, Madhumita A. Takalkar, Mengyang Duan, Haimin Zhang, Min Xu, *GEME: dual-stream multi-task GEnder-based Micro-Expression recognition*, Neurocomputing.

Part I

Background

CHAPTER 2

Related Research Review

Facial expression recognition plays a pivotal role in a wide range of applications of psychotherapy, security systems, marketing, commerce and much more. Detecting a macro-expression, which is a direct representation of an 'emotion', is a relatively straightforward task. Playing a pivotal role as macro-expressions, micro-expressions are more accurate indicators of a train of thoughts or even subtle, passive or involuntary thoughts. Compared to macro-expressions, identifying micro-expressions is a much more challenging research question because their time spans are narrowed down to a fraction of a second, and can only be defined using a broader classification scale. This chapter is all-inclusive survey-cum-analysis of the various micro-expression processing techniques. We analyse the general framework for micro-expression analysis system by decomposing the pipeline into fundamental components, namely face detection, pre-processing, facial feature detection and extraction, datasets and classification. We discuss the role of these elements and highlight the models and new trends that are followed in their design. We also discuss the new deep learning networks applied for facial micro-expression analysis systems. This survey focuses on the methodologies applied, existing databases, performance and comparing these to distil the gaps in the efficiencies, and research potentials. Through this survey, we intend to look into micro-expression analysis problem and develop a comprehensive and efficient recognition scheme.

2.1 Introduction

Emotions play a very prominent and purposeful role in day-to-day life. There is a high possibility of ambiguity, in guessing the hidden emotion, within an expression that elicits during situations of low or normal stakes. High stake situations provide more probability in predicting the emotion correctly as compared to low and normal stakes. Occurring in high stake situations, micro-expressions are the basis for expressing involuntary feelings. Micro-expressions happen in a fraction of a second and are hard to be recognised in real time especially lacking related expertise. Macro-expressions are usually displayed for $3/4^{th}$ of a second to 2 seconds. Although there are many different categories of emotion, there can be six universal expressions: anger, disgust, fear, happiness, sadness and surprise [36]. Macro-expressions occur over a single or multiple regions of the face depending on the category of expression.

Micro-expressions are described as a habitual pattern of the human face that is observable but too brief to convey an emotion. Micro-expressions are extremely fast facial expressions that usually last for 1/25 second to 1/5 second [69, 239]. They can easily be neglected during the casual conversations.

Since micro-expressions are barely perceptible to humans, a Micro-Expression Training Tool (METT) has been developed by Ekman [39] to teach a human how to spot and respond to micro-expressions. However, micro-expressions can seldom be falsified, and the essential difference between macro- and micro-expressions is the duration instead of the intensity of the expressions [180]. Currently, although experts can identify the existence and recognise micro-expressions, the accuracy is only about 47% [69]. Thus, having a system to improve the micro-expression analyses and help identify and categorise a person's feelings automatically and correctly is desirable.

Automated macro facial expression recognition has an enormous amount of existing research. Researchers have developed many algorithms, which have achieved a recognition accuracy of over 90% [2, 36, 85], for above mentioned six standard posed macro facial expressions. A recent study in [85] proposes a novel technique which when compared with existing state-of-the-art technique indicates a better result. On the contrary, micro-expressions have not yet been explored extensively due to several challenges.

One of the challenges, most of the researchers face, is the lack of a standard micro-expression database, which makes it difficult to obtain dynamic facial features to train an accurate micro-expression recognition system. There is no significant research on the dynamics of the micro-expressions. Since the appearance of the micro-expression

completely resembles the six primary macro-expressions, it is possible for researchers to train a system based on the existing macro facial expression databases by utilising the appearance information and ignoring the dynamic information. Other challenges include the development of robust methods to tackle the short span and low-intensity of micro-expressions.

There is a vast range of applications that can benefit from the study of micro-expressions. A primary purpose for the active involvement in micro-expressions is that it proves to be an important sign for detecting lies. For example, in situations when the suspects are being questioned, a micro-expression fleeting across the face can tell the Police that the criminal is pretending to be innocent. It can also benefit the border security officers for identifying suspicious behaviour of the individuals during usual interviews of checking for potential dangers. In the study of psychotherapy, micro-expressions have been proved very helpful in understanding genuine emotions of the patients. Micro-expression recognition systems are sometimes also used as an additional module for user authentication [169]. In other fields, such as marketing, distance learning, and many more, micro-expressions can be used as recognition to reflect human reactions and feedback to advertisements, products, services and learning materials.

This chapter provides a comprehensive survey of the existing micro-expression recognition methods along with their outcomes, to offer a convenient introduction to the recent developments in this domain.

2.2 Micro-Expression Databases

The success in macro- facial expression recognition primarily relies on sufficient facial expression databases, such as the popularly used Extended Cohen-Kanade (CK+) [133], Multimedia Understanding Group (MUG) [4], MMI [199], JApanese Female Facial Expression (JAFFE) [135], Multi- Pose, Illumination and Expression (Multi-PIE) [57] and also several 3-D facial expression databases. In contrast, there are very few well-developed micro-expression databases, which have hindered the development of micro-expression recognition research.

The foundation of the micro-expression detection and recognition system is a well established database. It is a complex and difficult task to build a database that meets various criteria and will be used broadly to test new algorithms. For recognition of micro-expressions, the creation of normalised database presents various challenges, including how to evoke expression and to choose micro-expression from raw videos. First,

the method of evocation requires the correct emotional catalyst option that with high ecological legitimacy. After capturing, there is a need for psychologists or trained experts to verify the labelling of these micro-expression samples. The first non-spontaneous micro-expression recognition databases used in earlier research are Polikovsky dataset [166], USF-HD [181], and YorkDDT [218]. Because of their inadequacies such datasets are uncommon. A significant issue is the ability to present expressions spontaneously. Non-spontaneous or posed expressions are the formulated gestures given by a person when he or she is asked to do so which is normally in the laboratory environment. Spontaneous expressions, on the other hand, are those voluntarily displayed when people engage in natural talks, watch movies, and so on. Micro-expressions that are posed are quickly recognisable, whereas it is hard to create and locate spontaneous ones. In terms of nature and temporal dynamics, spontaneous micro-expressions are distinct. Implementation of an empirical micro-expression system ensures recognising spontaneous instead of posed micro-expressions. As a result, researchers today have begun to work towards the development of the databases with spontaneous micro-expressions and improving spontaneous analysis of micro-expression. According to [11] *"a standardised training and testing database contains images and video sequences (at different resolutions) of people displaying spontaneous expressions under different conditions (lighting conditions, occlusions, head rotations, etc)"*. In recent years, nine micro-expression repositories have been developed. Three of them are extensively used to evaluate micro-expressions: 1) the Chinese Academy of Sciences Micro-Expressions (CASME) [231]; 2) the improved CASME (CASME II) [229] and 3) the Spontaneous Micro-expression Corpus (SMIC) [112]. The other spontaneous micro-expression databases are the Database of Spontaneous Macro-expressions and Micro-expressions (CAS(ME)^2 [167]) and the Spontaneous Actions and Micro-Movements (SAMM) [29].

The CASME database includes 195 samples of 1500 facial expressions categorised in 8 groups emotion classes of 19 valid subjects at 60 frames per second. These micro-expressions are action unit (AU) labelled with a facial region of interest cropped to size of 150×190 pixels. The participants are paid to hide all their facial expressions, if failing to do so, the amount of token given will be subtracted. They are shown videos of around $1-4$ min duration to induce micro-expressions. The AU identification is analysed from collected video data.

CASME II is a revised version of CASME database in which the number of samples are increased to 247 with 26 substantial subjects. In order to select the best examples out of 2500 facial expressions, a thorough selection was made. The videos are recorded at

200 frames per second in order to capture facial expression under restricted environment. The emotions are classified into seven emotion classes with facial region cropped to 280×340 pixels.

The CAS(ME)^2 database contains lengthy video of both macro- and micro- expressions. The database is divided into two sections; section A consists of 87 of both expression kinds and section B is split into two parts, with 300 of cropped spontaneous macro-expression and 57 of spontaneous micro-expression provided by the 22 participants. Samples are recorded at 30 frames per second in a rather low frame rate in a comparatively few samples of prior databases.

To overcome the issue of lack of ethnic diversity, SAMM is introduced. In this database, subjects from 13 diverse ethnicities have participated. Facial motion is recorded in a composed lab setup at 200 frames per second. Unlike earlier database collection, participants are initially requested to complete a set of questions before they advance to experiments. Depending on the participants' response, the conductor of the research showed videos that are consistent to the answers. FACS is coded with less attention to affect labelling in all recorded videos.

The SMIC (HS), the expanded version of the initial 77 samples SMIC dataset, includes 16 substantial individuals with 164 instances recorded in a confined setup at 100 frames per second . This dataset is split into three main negative, positive and surprise categories. The positive class contains happy emotion whereas negative class combines four feelings, i.e. sadness, anger, fear, and disgust. The remaining category is a surprise with surprise emotions. The micro-emotions in this dataset are captured in the same way as CASME dataset. The dataset is, however, not annotated with action unit, and the apex frame index remains unknown.

Alternatively, SMIC contains samples collected with 25 frames per second using standard speed camera (VIS) and infrared cameras (NIR) from 8 participants. Total of 71 samples consisting of micro-expressions are collected for both types of cameras, respectively.

2.2.1 Database Comparison

A comparison summary of non-spontaneous and spontaneous micro-expression databases is shown in Table 2.1. Considering the unavailability of non-spontaneous databases, a comprehensive analysis of those databases could not be done. CASME II collects a high number of CASME-like micro-expression samples that are 195 samples of 35 participants. CASME and CASME II, where all subjects are Chinese, have no variation throughout

Table 2.1: Summary of micro-expression databases

Dataset		Participants	FPS	Resolution	Samples	Emotion Class	FACS Coded	Ethnicities
Polikovsky		10	200	640\times480	13	7	Yes	3
USF-HD		\	29.7	720\times1280	100	6	No	\
YorkDDT		9	25	320\times240	18	\	No	\
CASME		35	60	640\times480, 1280\times720	195	7	Yes	1
CASME II		35	200	640\times480	247	7	Yes	1
CAS(ME)^2		22	30	640\times480	250 macro, 53 micro	8	Yes	1
SAMM		32	200	2040\times1088	159	8	Yes	13
SMIC	HS	16	100	640\times480	164	3	No	3
	NIR	8	25		71			
	VIS	8	25		71			

ethnic groups. The disadvantage of lack of diversity in ethnic backgrounds observed in CASME, CASME II and SMIC is mitigated by the introduction of SAMM database with participants from 13 different nationalities. In terms of age distribution with a median age of 33.24 years (SD: $\pm$11.32), SAMM also has the edge over the other. Samples for CASME II and SAMM are captured with a high frame rate of 200 fps. SAMM is currently the first and only database with high-resolution of 2040×1088 pixels with a facial region measuring 400×400. The CAS(ME)^2 has only 53 reported micro-expression samples. FACS coding is used for labelling CASME, CASME II and SAMM databases.

The emphasis of the researchers is on CASME II and SAMM, which have all necessary criteria to recognise micro-expressions: emotion groups, large frame rates, a rich spectrum of micro-expressions and the severity for facial movements are special.

2.3 Approaches for Facial Micro-Expressions Analysis

Micro-expression recognition systems are developed by considering many factors and parameters. Many studies have been undertaken and still undergoing in delivering better detection, spotting and recognition accuracy. Apparently, the identification process entails the extraction and categorisation of features. Nonetheless, before the actual feature extraction, the quality of descriptive data to be extracted by descriptors could be enhanced in the pre-processing stage. All the above phases, as shown in Figure 2.1, are addressed in this segment of chapter.

Figure 2.1: A general framework for micro-expression recognition analysis

2.3.1 Face detection

Face detection is the primary stage of the recognition process. Human face(s) is located in the digital images or image sequences. This step is useful for selecting the region(s) of interest (ROI) in the images or selects ROI in the first frame and track the face in the remaining frames in case of image sequences.

There are several face detection methods enforced till date [89, 97, 104, 170, 189, 203, 207]. Some of the latest face detection techniques are summarised here. Viola et al. [203] introduced the first framework to provide competitive detection rates in real-time since 2001. This framework is capable of processing images rapidly while achieving high detection rates. Wang et al. [207] suggested a coupled network of encoder-decoder to identify faces and locate facial landmarks together. The encoder and decoder produce response maps to locate the facial key points. They designed a coherent architecture for cascading multi-scale face detection through the combination of feature maps. With the help of deep learning, Sun et al. [189] introduced a new facial detection scheme and achieved the cutting-edge detection performance on the renowned FDDB face detection evaluation. A variety of techniques incorporating the feature concatenation, hard negative mining, multi-scale training, model pre-training and careful tuning of crucial parameters reinforced the faster RCNN framework.

2.3.2 Pre-processing

Pre-processing is the common name for operations performed on images at the lowest level. The aim is to achieve improvement of the image data that suppresses unwanted distortions or enhances some features for further processing. The sequences for micro-expressions are of very short duration wherein the intensity of the facial movements is low. There are several methods implemented to normalise the input data so that sufficient details about the micro-expression are extracted for further processing. Some of the commonly used pre-processing methods are discussed below.

2.3.2.1 Facial Action Coding System (FACS)

The Facial Action Coding System (FACS) is described as an anatomy based system for comprehensively describing all facial movements. FACS devised by Ekman and Friesen [53], provides an objective means for measuring the facial muscle contractions involved in a facial expression. FACS was developed to allow researchers to measure the activity of facial muscles from video images of faces. Each noticeable component of facial movement is called an Action Unit (AU). Ekman and Friesen [53] defined 46 distinct action units, each of which corresponds to displacement in a specific muscle or muscle group, and produces facial feature deformations which can be identified in the images. Recently, considering the significance of action units in the detection and recognition of micro-expressions, Li et al. [114] introduced a new Spatio-Temporal Adaptive Pooling (STAP) network for localising action units in micro-expressions.

2.3.2.2 Temporal normalisation (TIM)

Micro-expression video clips are unevenly small (or long). This can lead to two conflicting situations: (a) for short-duration videos that restrict the use of feature extraction strategies requiring variable time window duration (e.g. LBP dependent modes which can render binary patterns from different radius), (b) for long-duration videos which could degrade the recognition efficiency through recurrent or duplicate frames (captured with high speed camera). For answering this issue, the TIM (Temporal Interpolation method) is used to render clips of identical frame lengths either by up-sampling (too short clips) or down-sampling (too long clips).

Pfister et al. [164] suggested to standardise the video frames in a certain time period, using a TIM. TIM uses graph embedding to interpolate images at arbitrary positions within micro-expressions. This interpolation allows inputting a sufficient number of frames to the feature descriptor. TIM is a manifold-based interpolation method that inserts a curve in a low-dimensional space after embedding of an image sequence. In Figure 2.2, a micro-expression video is represented as a set of images sampled along the curve creating a low-dimensional manifold by delineating the micro-expression video as a path of the graph with vertices. The interpolated frames are mapped back to a high dimensional space to form the temporally normalised image sequence [111].

Figure 2.2: (a) An example of micro-expression being interpolated through graph embedding; (b) Temporal interpolation method. The video is represented onto a curve along which a new video is sampled [164]

2.3.2.3 Integral projection

Huang et al. [75] proposed a new framework to obtain the horizontal and vertical projections using the integral projection method based on calculating the difference of images, which helps to retain the shape aspect of facial images. The integral projection generates a one-dimensional pattern by summing the given set of pixels along a given direction. The integral projections can extract common structure for the same person. In a micro-expression video clip, supposing that a frame is neutral, the difference between neutral face image and the expression image derive new images. The new derived facial images help reduce the influence of face identity on recognition methods. The integral projection itself does not define the presence and movement of facial images. It is, therefore combined with feature extraction method, e.g. LBP-TOP (as discussed in Section 2.4.3), to get the appearance and motion features. To preserve sufficient information in the process of projection, a new spatiotemporal method based on integral projection is introduced in [75]. Hence, the method is called as Spatiotemporal Local Binary Pattern with Integral Projection (STLBP-IP). Figure 2.3 shows the procedure to encode integral projection by using LBP. STLBP-IP achieves state-of-the- art performance compared to TIM.

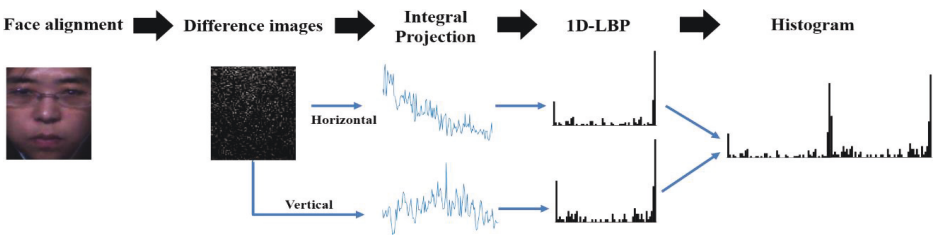


Figure 2.3: The procedure of encoding difference-image based integral projection on the spatial domain [75]

2.3.2.4 Color space model

Color is a fundamental aspect of human perception, and its effects on cognition and behaviour have attracted interests of many generations of researchers. Recent research revealed that colour might supply useful data for face recognition.

Wang et al. [213] demonstrated a Tensor Discriminant Color Space (TDCS) model that uses a 3rd-order tensor to represent a color facial image. To make the model robust to noise, they [212] also used an elastic net to propose a Sparse Tensor Discriminant Color Space (STDCS). Lajevardi and Wu [99] also addressed a color facial expression image as a 3rd-order tensor and presented that the perceptual color spaces (CIELab and CIELuv) are better overall for recognition of facial expression than most color spaces .

A new color space model called tensor independent color space model (TICS) [210, 211] reveals that a micro-expression color video sequence is conceived as a fourth-order tensor, i.e., a four-dimension array. The initial two dimensions cater the spatial details; the third delivers the temporal data, and the fourth gives the color specifications. Wang et al. [211] transformed the fourth dimension from RGB into TICS, wherein the color elements become as separate as possible. In a color micro-expression video clip, the correlated R, G and B components in RGB space are transformed into a series of uncorrelated components T1, T2 and T3, and extract the dynamic texture features from each uncorrelated component to obtain better results. These research measured the impact of various colour spaces on detection based on facial motion rather than skin colour change.

Recently, Shahar et al. [178] attempted to explore a feature of human face which is much more complicated to conceal, which is the facial colour alteration induced by the flow of blood while expressing emotions. They recommended a method that measures shift of colour during micro-emotion, which overlooks movement-related facets of expression and solely focusing on face colour and distinguish emotional types successfully.

2.3.2.5 Motion magnification

The complexity and subtlety of the micro-expressions is one of the contributing factors making it difficult to recognise them automatically. Given the very weak intensity of facial micro-expression movement, the distinction between micro-expression types is extremely difficult. One approach to this issue is the distortion or amplification of these facial micro-movements.

Recent works [111, 156, 215, 238] have used the Eulerian Motion Magnification

(EMM) [225] technique for magnifying the subtle movements of micro-expression videos. With the use of band-pass filters, the EMM technique derives the frequency bands of interests from the numerous spatial frequency bands generated from the disintegration of the input video, which are augmented with a magnification factor in order to intensify the motions at a different spatial level. Li et al. [111] proved that the EMM approach leads to increase in the disparity between the various micro-expression groups (i.e. inter-class variation) and thereby improve the recognition rate.

Nevertheless, greater amplification factors may trigger unwanted distorted noises (movements that are non-ME induced) may impede the recognition efficiency. Le et al. [101] technically calculated the upper limits of impact to avoid over-amplification of micro-expression samples. In addition, the authors also analysed the efficiency of the amplitude-based EMM (A-EMM) and phase-based EMM (P-EMM). Park et al. [156] suggested a magnification regime that modified the most selective frequency band required for EMM to enhance the subtle facial movements, in order to mask the unique temporal features of various micro-expression groups. A study by Le et al. [100] demonstrated the Global Lagrangian Motion Magnification (GLMM), in particular, with higher magnification, may lead to greater recognisability than local Eulerian approaches.

2.3.2.6 Other techniques

Active Appearance Models (AAM) is a statistically based template matching method, where a representative training set takes the variability of shape and texture. A group of images with landmark coordinates that appear in all of the images is given to the training supervisor. Edwards, Cootes, and Taylor [39] were the first to introduce the model in the context of face analysis. The method is widely used for matching and tracking faces and for medical image interpretation [23]. The algorithm applies the difference between the current estimate of appearance and the target image to derive an optimisation process. To match an image, the current residuals are measured and use the framework to anticipate changes to the present parameters, leading to a better match. A good overall match is obtained in a few iterations, even from poor starting estimates. AAMs, learn what are the valid shapes and intensity variations from their training set.

Active Shape Model (ASM) algorithm is a fast and robust method of matching a set of points controlled by a shape model to a new image. Cootes et al. [24] proposed the active shape model where shape variability is learned through observation. ASM is again a statistical model of the shape of objects which iteratively deform to fit an example of

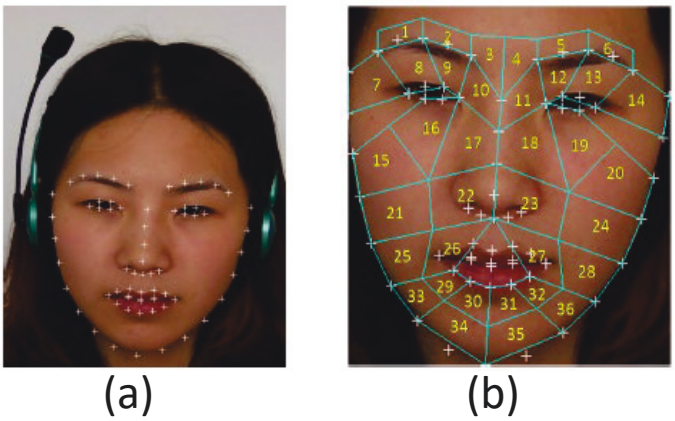


Figure 2.4: (a) 66 feature points using DRMF; (b) 36 regions-of-interest (ROIs) [127]

the object in a new image. The technique relies on each object or image structure being represented by a set of points. The points can represent a boundary, internal features, or even external ones, such as the centre of a concave section of the border. Points are placed in the same way on each of a training set of examples of the object. The sets of points are aligned automatically to minimise the variance in the distance between similar points. By analysing the statistics of the positions of the labelled points a "Point Distribution Model (PDM)" is derived. The model gives the average positions of the points and has some parameters which control the main modes of variation found in the training set [24].

Registering and tracking a non-rigid object has significant variations in shape and appearance. Discriminative Response Map Fitting (DRMF) [7] is one of holistic texture based methods, which relies on shape initialisation. Moreover, as a discriminative regression-based approach, DRMF performs impressively well in the generic face fitting scenario. DRMF is used to identify a series of facial landmarks in the facial area of the first frame in every micro- expression video sequence. DRMF located 68 feature points in a facial region. With the help of Facial Action Coding System (FACS), 36 Regions of Interest (ROI) are marked, and the face region is partitioned as shown in Figure 2.4 [127, 160].

2.3.3 Features

The extraction of features ensures that the volume of the data needed to reflect a broad range of dataset is minimised. Perhaps the most crucial phase in micro-expression recognition is the facial feature extraction. Various viable features are adopted by

several researchers to reveal facial characteristics. Micro-expression feature extraction mechanisms are categorised as geometric-based and appearance-based. Geometrical features represent face morphology, such as curves and facial landmarks, and they involve accurate landmark localisation and alignment techniques. Conversely, appearance-based features define colour and textural details namely wrinkles and hue shifts and are therefore more resilient for alterations in lighting and orientation. There was, therefore, a rise in prevalence of appearance-based micro-expression recognition approaches [106], including LBP-TOP [243], HOG 3D [166], HOOF [20] and deep learning.

2.3.3.1 Traditional Approaches

Local Binary Pattern- Three Orthogonal Planes (LBP-TOP) and its variants The basic idea behind Local Binary Pattern (LBP) is to compare the centre pixel value with the neighbourhood pixel values. A binary code is generated by assigning the value one to the greater neighbour pixel value and assigning zero to the rest. The obtained binary code is converted to decimal to get the local binary pattern value of the centre pixel. The calculation process of a basic LBP operator can be understood from Figure 2.5.

Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) combines the temporal features along with the spatial features from LBP of the image sequence. A video is a chronological succession of frames with three dimensions, i.e. two dimensions (X and Y) are the spatial information and the third dimension (T) is time. The three orthogonal plane corresponds to the combination of spatial and temporal planes: XY, YT, and XT. LBP is firstly computed on these three planes. After that, histograms corresponding to each plane are obtained, which are concatenated to describe the dynamic texture of the micro-expression video [217, 243].

A feature vector is generated by calculating a histogram of LBPs over a whole image. The LBP method is effective for describing 2D textures of static images, but to analyse time-dependent textures (i.e. changing expressions in the video), LBP method needs to be broadened. To extend the LBP method, computation of LBP histograms is done in

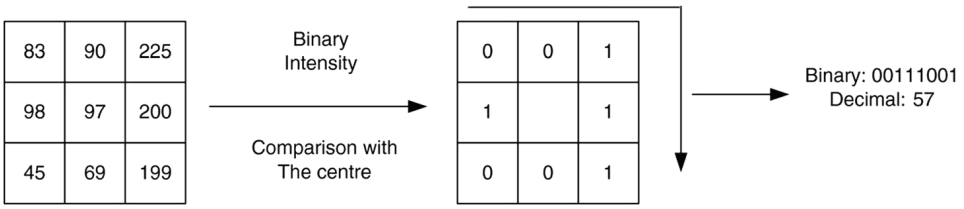


Figure 2.5: The calculation process of a basic LBP operator

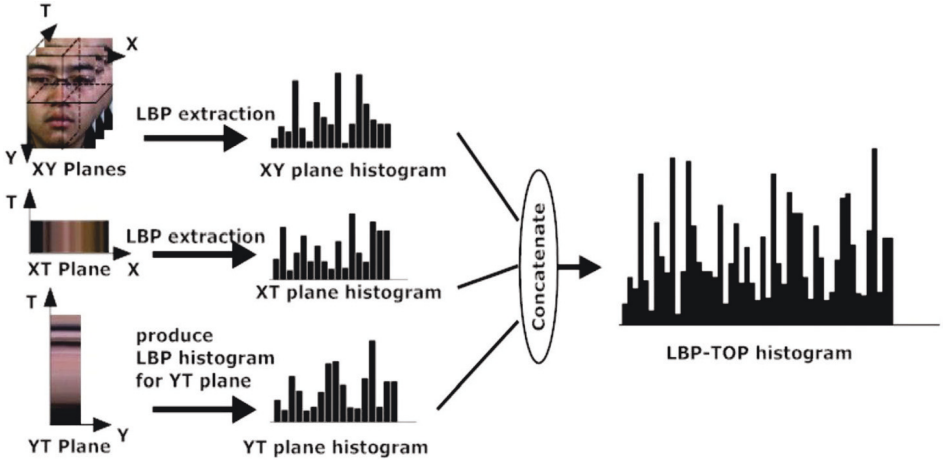


Figure 2.6: LBP-TOP example: scan all the pixels to calculate their LBP histograms on the XY, XT, and YT plane respectively. The data of the pattern frequency is counted in each corresponding histogram and then concatenated as one [217].

three orthogonal planes. For a video with time duration T, LBP is calculated for the XY-, XT-, and YT- planes. The XY- plane describes the spatial changes, while the XT- and YT- planes describe the spatial-temporal change in each respective dimension. The calculated histograms are then merged to generate the final LBP-TOP feature vector. The frequency of patterns at each of the three planes is counted to prevail the corresponding histograms, which are then integrated to describe the dynamic texture of the video (Figure 2.6).

LBP-TOP has been by far the most common method combined with various learning techniques for database evaluation and classification [112, 229]. LBP-TOP has evolved as a reliable way for spontaneous ME research following the Pfister et al. [164] groundbreaking contribution to the field and several variations were suggested. Also suggested by Pfister et al. [163] is a system for discriminating between spontaneous and posed facial expressions (SVP). The Complete Local Binary Patterns (CLBP), which Guo et al. [61] proposed, was enhanced and is referred to as CLBP from Three Orthogonal Planes (CLBP-TOP) to deal with complex texture descriptor. The three intersecting lines that pass over the centre point is the cornerstone of LBP Six Intercepting Points (LBPSIP) [216]. Another variant LBP Mean Orthogonal Planes (LBPMOP) [209] initially determines the mean plane for three orthogonal planes and further calculates the LBP on the three mean orthogonal planes. LBPSIP and LBPMOP obtained improved performance by reducing duplicate details. The additional parameters of magnitude and directions are also considered in Spatio-Temporal Completed Local Quantized Patterns (STCLQP) [77]. The Radon Transform is employed in Spatio-Temporal Local Radon Binary Pattern

(STRBP) [76] for collecting reliable structural features whereas Spatiotemporal Local Binary Pattern with Integral Projection (STLBPIP) [75] implements integrated projections to retain shape properties. Guo et al. [58], in 2019, suggested an easy, effective yet robust feature descriptor for micro-expression recognition named Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP). Along with LBPTOP, ELBPTOP comprises of two additional new binary descriptors referred to as Radial Difference LBPTOP (RDLBPTOP) and Angular Difference LBPTOP (ADLBPTOP) that analyses details of local secondary order in the radial and angular directions of the micro-expression video sequences. ELBPTOP effective in computing and raises LBPTOP computing costs just fractionally, while it still allows micro-expression recognition extremely successful.

Histogram of Oriented Gradients (HOG) HOG is another popular feature descriptor used in computer vision and image processing for the use of object detection. The procedure counts occurrences of gradient orientation in localised portions of an image. According to the study by Dalal et al. [26], the local object appearance and shape within am image can be depicted by the distribution of intensity gradients or edge directions. The image is divided into small spatial regions called cells, and for the pixels within each cell, a histogram of gradient direction is compiled. The descriptor is a concatenation of these histograms. For better precision, the local histograms can be contrast-normalised by measuring an intensity factor over a wider spatial region of the image, termed as block, and then normalise all the cells within the block using this magnitude. These normalised blocks are referred to as Histogram of Oriented Gradient (HOG) descriptors [26].

Li et al. [39] claim that on the XY image plane, a 2D HOG can be constructed. In the first step, the horizontal and vertical representations are derived by sampling image using kernels. The next step is to develop the histogram, determined by the gradient directions and its magnitudes. The pixel vote is divided equally into two bins when the gradient path is between the two bins of the histogram. In brief, a quantised orientation channel is built according to each pixel's weight within a block dependent on the corresponding gradient calculation. The Histogram of Image Gradient Orientation (HIGO) is also a variant of gradients implemented and introduced in [39]. HIGO is a less complex and a simpler variant of HOG. HIGO avoids the weighting of magnitude and thus subdue the lighting effect. HIGO is, therefore, one of the most precise descriptors today. Nonetheless, it should be mentioned that the HOG gradient is an edge-based gradient descriptor. When not filtered, it is responsive to noise and application of low-

pass filtering in micro-expression recognition may result in the loss of details about the subtle motion shifts. Moreover, the estimation process is long and tedious, contributing to a slow speed.

The related HOG and HIGO were advanced to 3D with three orthogonal planes instead of the XY planes used in 2D solutions in the case of spatio-temporal feature extraction. Initially, HOG 3D [166] is utilised for identification of posed micro-expressions, followed by a baseline approach for spontaneous micro-expressions. Polikovsky et al. [166] segmented the facial area into 12 regions, utilising manually annotated points centring a rectangle on these points. Implementation of 3D HOG was done to detect the movements in every region. The fact that different facial segments contribute differently to micro-expressions, as shown by Chen et al. [21] have been commonly disregarded in previous studies. They recommended implementing 3D HOG weighted approach along with fuzzy classifier for micro-expression recognition.

Optical Flow-based methods Optical flow infers target movement by evaluating the shift in pixel amplitude between two image frames over time. Lucas-Kanade [132] assumes the displacement of the pixels between two nearby frames is small and nearly constant. Horn-Schunck introduces a global constraint of smoothness to solve the aperture problem. The method assumes precision in the movement over the whole image, trying to minimise distortions in the flow [68]. Usually, optical flow method is extracted and analysed for cropped and pre-processed images to identify pose and face variations [181].

Unlike the LBP variants, optical flow aims at tracking and capturing non-rigid facial part movement [64]. The Optical Strain Map (OSM), proposed by Liong et al. [120], is determined from the severity of Optical Strain Feature (OSF), Optical Strain Weight (OSW) and the blended variant of the two features. For every pixel of a video, OSM often includes graphical representations of motion intensity. OSM locates the highest or lowest motion projected area of the image frame when the spatial displacement is considered. Liong et al. [120] coupled optical strain weight and optical strain features to provide high efficiency relative to LBP-TOP and its versions. In order to achieve improved performance in recognition, histogram bins which lead to noise are disregarded.

Amongst all optical flow variants, Histogram of Oriented Optical Flow (HOOF) [20] is one of the foundation approaches used in micro-expression recognition. Facial Dynamics Map (FDM) [228] outlined the spatial facial dynamics with the extraction of each cuboids' primary optical flow vector. Likewise, [127] has developed the Main

Directional Mean Optical flow (MDMO) characteristics where the action unit details are used through the facial region segmented into 36 regions of interest. Contrary to these approaches, Consistent Optical Flow Maps [5] calculates the optical flow to describe facial motions from 25 ROIs and that the optical flow of each segment can be measured in several directions. Lately, only the peak frame and the onset frame are being used by Bi-Weighted Oriented Optical Flow (Bi-WOOF) [122]. Most of the optical flow-dependent approaches have to segment the facial region accurately in order to use AU details. It boosts efficiency, but raises pre-processing complexities. In [244], the LBP-TOP and HOOF hybrid features are determined for automated Necessary Morphological Patches (NMPs) extraction, which incorporates the AU-based and feature selection approaches.

2.3.3.2 Deep Approach

Deep learning success has sparked the community to seek unique and innovative ways of improving feature extraction. Patel et al. [159] initiated to use the deep features, in their approach, transferred from pre-trained ImageNet models. They realised it is not feasible to optimise the network with an inadequate amount of data available from micro-expression datasets instead preferred a feature selection strategy. Kim et al. [88] applied CNN and LSTM for interpreting spatial and temporal characteristics, respectively. This approach has taken advantage of the derived features to distinguish micro-expressions while transferring the model to long-short term memory (LSTM) recurrent neural network for evaluating the temporal features of the data.

Reddy et al. [173] suggested MicroExpSTCNN approach focused on 3D-CNN design implemented over the entire face. Wang et al. [205] suggested the addition of a remaining block-based attention unit to the proposed CNN approach to aid network to focus on vital regions. Motivated by visual-attention and commonly used CNN systems, Yang et al. [232] built an attention-based CNN network called MERTA, a deep learning model for reliable extraction of particular features for precise classification of micro-expressions. The design consists of three VGGNets and one Long Short-Term Memory (LSTM). Three VGGNets aim to collect static and dynamic details where three kinds of attention processes are combined to render visual representations more careful distinctions. The spatial characteristics of the micro-expression sequence are provided in a sequential order to LSTM in order to derive spatio-temporal properties and estimate micro-expression category. Verma et al. [202] lately introduced a workaround for CNN named as Lateral Accretive Hybrid Network (LEARNet). The input can be resumed by introducing an accretion layer for optimising the salient expression features. Quang et al. [200] modified

the CapsuleNet framework for micro-expression recognition by choosing the apex frame, the most significant frame from the micro-expression sequence. Transfer learning from ImageNet and data augmentation is implemented due to the lack of data. Wang et al. [209] addressed the issue of small sample size by implementing transfer learning in order to pre-train a deep neural network and suggested a micro-expression recognition system referred to as Transferring Long-Term Convolutional Neural Network (TLCNN). TLCNN involves two transfer learning phases: (1) transfers from expression data and (2) transfer from a single frame from micro-expression video, which can be considered as "big data". TLCNN also incorporates LSTM to capture temporal features in micro-expression clips from mid-level range image representation for every frame.

Learning micro-expressions' distinctive features from three vital video frames for recognising micro-expressions was achievable using a recently introduced Three-Stream Convolutional Neural Network (TSCNN) by Song et al. [185]. TSCNN is constructed using a dynamic-temporal stream, static-spatial stream, and local-spatial stream module in order to learn and incorporate, in micro-expression videos, respectively, temporal, whole facial and local facial area references for recognising micro-expressions. TSCNN also developed a robust apex frame detection method for micro-expression recognition instead of using the index value of the apex frames.

2.3.3.3 Hybrid Approach

The fundamental principle behind hybrid frameworks is about using handcrafted techniques together with deep learning methodologies. A novel mechanism for recognising micro-expressions was introduced by Hu et al. [71] that integrates handcrafted features with deep features. Local Gabor Binary Pattern from Three Orthogonal Planes (LGBP-TOP) is the handcrafted feature descriptor used in the implemented hybrid system. In order to encrypt local facial motions, LGBP-TOP incorporates spatial and temporal processing. Convolutional Neural Network (CNN), a class of deep neural network, model trained on micro-expression dataset is the other descriptor adopted. Then, with the adaptive repressive parameters, the sparse multi-task learning system is used to eliminate the less important details from the integrated LGBP-TOP and CNN features. Hybrid solutions like such, nevertheless, render it as a computationally challenging approach. Thanks to the prominent hardware and software advances, these approaches can be accomplished.

Khor et al. [87] suggested an Enriched Long-term Recurrent Convolutional Network (ELRCN). Initially, a variety of Optical flow forms (Horizontal, Longitudinal, Magnitude

and Strain) are computed. They then provided two separate CNN structures, one for obtaining spatial characteristics (the input was the image combined with the results from different optical flow) while the other extracts temporal features where every optical flow outcome was given to a specific convolutional block of a 3 block CNN module. The CNN structures conclude classification task through a fully connected layer. A new technique referred to as Dual Temporal Scale Convolutional Neural Network (DTSCNN) was introduced by Peng et al. [161] in 2017. The insufficiency of the data in existing datasets resulted in the design of a shallow neural network comprising of only four layers for convolutional and pooling to recognise micro-expressions. DSTCNN is a two-stream network as the name suggests.

Gan et al. [55] suggested Off-ApexNet model that functions in three stages of offset and peak frames detection, horizontal and vertical flow computation and eventually providing all this to a CNN. A fully connected layer performs the classification process. STSTNet [118] is an enhanced alternative to Off-ApexNet where optical strain is incorporated along with the horizontal and vertical optical flow in order to achieve better performance. The Spatiotemporal Recurrent Convolutional Network (STRCN) has been suggested by Xia et al. [226]. In this study, two variants of the network are formulated: STRCN with Appearance-based Connectivity (STRCN-A) utilising a different image representation as a vector thus passing entire sequence as a matrix to STRCN which is essentially a recurrent CNN block. The second variant is Geometric-based Connectivity (STRCN-G) which involves the application of optical flow before feeding the STRCN block. Several other suggested studies [86, 148] summarises the methods in two phases: evaluating optical flow or LBP, which is supplied to a CNN or RNN architecture for extracting the corresponding spatio-temporal features.

2.3.4 Classification

Image classification analyses the statistical attributes of different image features and regulates data into categories. Classification is typically a two-step process: training and testing. Most of the research on micro-expression recognition applies existing classification methods as discussed below.

2.3.4.1 Support Vector Machine (SVM)

SVMs [69, 75, 81, 164, 229] are based on the concept of decision planes that define decision boundaries, which primarily perform classification tasks by constructing hy-

perplanes and a multidimensional space that separate samples of different class labels. SVMs correlate to the general category of kernel methods, which can operate in a high-dimensional, implicit feature space, through applying kernel functions. SVM has two advantages: Firstly, SVM can generate non-linear decision boundaries using methods intended for linear classifiers. Secondly, the use of kernel functions grants the user to implement a classifier to data that have no demonstrable fixed dimensional vector space representation [9]. Some kernels can be used in SVMs, e.g., linear, polynomial, Radial Basis Function (RBF) and sigmoid. The RBF is one of the most widely used kernel types in SVMs mainly because of their localized and limited responses across the entire range of the real X-axis.

2.3.4.2 Extreme Learning Machine (ELM)

ELM is a single hidden layer feed-forward neural network which has extremely fast learning speed [208]. ELM has better generalisation performance. ELM classifier provides a unified learning platform with popular features mappings and can be directly applied in the regression and multiclass classification [60]. According to the research, ELM is proved to have better generalisation performance, much faster training and learning speed than the traditional SVM. This characteristic of ELM proves vital in the micro-expression recognition.

2.3.4.3 Nearest Neighbor Algorithm (NNA)

NNA depends on limited adjacent samples, so as compared to other methods, NNA is more efficient for the sample set with class fields cross [134]. In this study [134], the system integrated the gradient magnitude weighted into Nearest Neighbour Algorithm for classification. The idea of the nearest neighbor method is to compare the distances between unknown samples with the entire known sample set and to judge the distances between samples. Euclidean distance is one of the measurements of similarity among samples [59]. If the distance of two samples in the feature space is close, then the samples may have the same label. The fine calculation ability of NNA can help the micro-expression recognition system to classify accurately.

2.3.4.4 Multiple Kernel Learning (MKL)

MKL is developed for supervised, semi-supervised and unsupervised learning. The fundamental idea behind MKL is to add an extra parameter to the minimisation problem

of the learning algorithm. MKL determines weights for linear/non-linear combinations of kernels through various domains by optimising a cost function [164]. Compared to SVM, MKL can provide better micro-expression recognition in some cases.

2.3.4.5 Random Forest (RF)

Random Forest, also known as Random Decision Forest, is an ensemble learning method that is used for classification and can be thought of as a form of the nearest neighbor predictor. Ensemble learning is a divide-and-conquer method used to improve performance. The basic principle of ensemble methods is that a set of "weak learners" can come together to form a "strong learner" [12]. Several researchers [30, 131, 163, 164] opted to use RF classifier for facial micro-expression recognition.

2.3.4.6 Other classifiers

Some researchers attempted to work with relaxed K-SVD, Sparse Representation Classifier (SRC) and Group Sparse Learning (GSL) strategies to tackle the sparseness for MEs. Nonetheless, each solution deals differently with micro-expression's sparseness. A sparse dictionary is used by relaxed K-SVD [245] to differentiate various micro-expressions by reducing the variance of sparse coefficients. The SRC [234] employed in [240] is a sparse linear combination of all the training samples representing a given test sample; the sparse non-zero coefficients will, therefore, probably focus on training and test samples of the same class. The aim of kernelised GSL [248] is to encourage the technique of learning a series of essential weights from hierarchical spatio-temporal descriptors, which can assist in selecting important block from multiple facial blocks.

2.4 Discussions

This chapter gives a far-reaching overview of cutting edge facial micro-expression analysis approaches including handcrafted, deep learning-based methods that structure the key components of the micro-expression recognition system. The handcrafted micro-expression recognition approaches have been there for a significantly long time and accomplished surprising outcomes on available benchmark datasets. However, most successful handcrafted recognition methods are based on the local densely-sampled descriptors. In these methodologies, the necessary features are extracted from a sequence of video frames to generate the feature vector using human engineered feature detectors

and descriptors. Later, the classification is performed by training a generic classifier. These approaches include space-time, appearance, geometry, and local binary patterns based methods.

On the other hand, learning-based micro-expression recognition method uses trainable feature extractors followed by the trainable classifier, which prompts the idea of end-to-end learning or learning from pixel level to micro-expression class identification. This rules out the need for handcrafted feature detectors and descriptors utilised for micro-expression recognition. These approaches include deep learning-based approaches. These approaches give high performance as compared to their handcrafted counterparts on micro-expression datasets. A few of the deep learning methods are still taking help from the handcrafted features. Such methods are the hybrid methods where either the feature detection or selection is done using handcrafted methods and given to the deep methods or the features extracted using handcrafted methods are combined with the features extracted by deep methods.

Part II

Proposed Methods for Micro-Expression Detection and Recognition

CHAPTER 3

LGAttNet: Automatic Micro-Expression Detection using Dual-Stream Local and Global Attentions

Research in the field of micro-expressions has gained significance in recent years. Many researchers have concentrated on classifying micro-expressions in different emotion classes, while detecting the presence of micro-expression in the video frames is considered as a pre-requisite step in the recognition process. Hence, there is a need to introduce more advanced detection models for micro-expressions. In order to address this, we propose a dual attention network based micro-expression detection architecture called LGAttNet. LGAttNet is one of the first to utilise a dual attention network grouped with 2D-CNN to perform frame-wise automatic micro-expression detection. This method divides the feature extraction and enhancement task into two different CNN network modules; Sparse Module (SM) and Feature Enhancement Module (FEM). One of the key modules in our approach is the attention network which extracts local and global facial features, namely Local Attention Module (LAM) and Global Attention Module (GAM). The attention mechanism adopts the human characteristic of focusing on the specific regions of micro-movements, which enables the LGAttNet to concentrate on particular facial regions along with the full facial features to identify the micro-expressions in the frames. Experiments performed on widely used publicly available databases demonstrate the robustness and superiority of our LGAttNet when compared to state-of-the-art approaches.

3.1 Introduction

One of the most natural ways for individuals to communicate their feelings and thoughts is through facial expressions. Perhaps the correct perception of feelings from facial expressions is the most significant social activity that people, as social beings, perform [151]. Not all feelings, though, will be reflected on the face. Given the attempts to conceal, studies have discovered that real feelings are always leaked. Such leaked feelings typically manifest as micro-expressions (MEs) [42]. Micro-expression is a brief facial expression, lasts for the overall duration of less than 500 milliseconds and the onset duration of less than 260 milliseconds [230].

Typically, it happens in circumstances of high stakes, particularly for people who win or risks something valuable [42]. The precise identification of such micro-expressions provides a tremendous ability for those with face-to-face communications expertise, including health care professionals, psychotherapists, educators and law enforcement officers because of their involuntary nature [47, 137]. Moreover, recognising micro-expressions is regarded as one of the most accurate tools for identifying deceit, owing to the near association between micro-expressions and deception [52].

Similar to macro-expressions, the micro-expression research has gained popularity in recent years. Automatic macro-expression detection and recognition can be accomplished with the advent of technology in real-time and is effectively implemented in industry since macro-expressions are clear to recognise and last for 500 milliseconds to 4 seconds [31]. In contrast to macro-expression, a micro-expression is harder to recognise due to its subtle presence in the facial regions, which makes the detection and recognition using naked eye challenging to accomplish. According to Ekman[1], sometimes the micro-expressions can be quicker than usual, and even occur for less than 40 milliseconds. Ekman further indicated that the detection and recognition of micro-expression is much more challenging in comparison to spotting. Detecting micro-expressions has to be performed using the images while neglecting the temporal information connected to micro-expressions. Moreover, much of the literature related to micro-expression focuses on spotting [205, 232], and a little research has been dedicated to the detection, which is the foundation of this study.

The continuing technological innovation in computer vision and machine learning helps in boosting the recognition efficiency, and alleviate the issues related to micro-expression detection. As spontaneous micro-expressions can often be seen in real life

[1]https://www.paulekman.com/resources/micro-expressions/

and reveal better affective knowledge about humans, this work concentrates on the issue of detecting spontaneous micro-expressions from video frames. Thus far, many micro-expression detection approaches analyse the disparity in the features between the first frame and the other frames in a time span [13, 15, 112]. In contrast, this work aims to detect the micro-expressions from the spatial features that can be extracted from a single video frame.

In the case of micro-expressions, it is interesting to note that most of the clues originate from a few facial regions such as the mouth and eyes. Ideally, this suggests that the machine learning models must concentrate only on the relevant facial areas and be less responsive to the other facial regions. The predictions made by CNNs are based on the posterior probability functions, whereas, the professionals, typically render judgements that can be clarified more clearly depending on the selective local facial regions of interests (RoIs). Similar to this human behaviour, the attention mechanism can also concentrate on specific regions of images. Much research lately aims to incorporate attention mechanisms with deep networks [136, 143, 188]. Through deep learning, the attention model directly simulates the human brain's attention mechanism. In earlier research, the importance of incorporating the attention system has been thoroughly discussed [83, 117, 201, 205, 206, 214, 224, 232].

In this chapter, a deep learning framework for micro-expression detection is proposed, which applies an attention mechanism to concentrate on the salient parts of the face. The following are the novel and key contributions of this chapter: [5]:

* We propose an attention driven detection mechanism, called LGAttNet, to identify the frame-wise micro-expression. According to the best of our knowledge, this is the first approach to use a dual attention network for building a micro-expression detection framework. LGAttNet is designed to be an automatic micro-expression detection model which focuses on facial regions with specific information related to micro-expressions on top of the full face information.

* The attention networks in LGAttNet are structured as dual-stream local and global attention blocks. The local attention stream of the architecture focuses on the Regions of Interest (RoIs) that exist only within local facial areas for associated micro-muscle movements, whereas the global attention stream considers the full face, establishing a relation between the local facial RoIs. Further, to extract the

[5] Madhumita A. Takalkar, Selvarajah Thuseethan, Sutharshan Rajasegarar, Zenon Chaczko, Min Xu, John Yearwood, *LGAttNet: Automatic Micro-expression Detection using Dual-Stream Local and Global Attentions*, Multimedia Tools and Applications. (*Submitted*)

global and local attention maps, a combination of deep and shallow networks is used instead of using a single very deep CNN, wherein the deep network provides a sparse representation of the features, that are subsequently passed on to a traditional shallow CNN with sigmoid function.

* We have evaluated the accuracy of our architecture, LGAttNet, on publicly available and widely used CASME, CASME II, CAS(ME)^2 and SAMM databases using a leave-one-subject-out (LOSO) cross-validation. An ablative study to manifest that the idea of building a dual-stream network using local and global attention networks achieves improvement in the micro-expression detection accuracy is conducted. A cross-database analysis is also performed to verify the efficiency of the proposed architecture. Furthermore, a comparison of the LGAttNet with state-of-the-art approaches is performed to demonstrate that the LGAttNet performs remarkably well in detecting the micro-expressions in video frames.

The remaining parts of this chapter are structured in the following order. Section 3.2 summarises the related studies performing binary classification between ME and non-ME video frames. Section 3.3 introduces our framework, called LGAttNet, and its supporting components. Experimentation performed on the model along with the outcomes are presented in Section 3.4. Finally, the chapter concludes in Section 3.5 with future research directions.

3.2 Related Research

In spite of the fact that automatic micro-expression detection and recognition is not broadly analysed in contrast to macro-expression study, a number of works have addressed this problem with the recent advances in computer vision. For micro-expression analysis, micro-expression detection is a crucial and essential pre-processing phase in defining a corresponding series of frames from a given long video containing micro-expressions. Given that micro-expression is an uncontrolled facial expression, the micro-expression detection study has been conducted on the publicly available spontaneous micro-expression repositories. It is difficult to distinguish concise facial gestures from neutral faces, especially in real life videos and to prevent false alarms triggered by global facial actions, speaking and occlusions.

Many approaches proposed for micro-expression detection primarily focus on assessing the discrepancy between their own features, which indicates the disparity in the

time window from the first and the other frames. Since the span of the spontaneous micro-expression is relatively short, only a few frames are available in a video that reveal the micro-expressions rendering the detection of spontaneous micro-expression extremely hard. The techniques used in the literature to detect these micro-expressions are broad, including optical flow [121], Local Binary Patterns (LBP) [105], Histogram of Oriented Gradients (HOG) [27] and integral projection [130].

Davison et al. [28, 29] implemented ‚Äúindividualised baselines‚Äù determined by taking the participant‚Äôs neutral video sequence and using the Chisquare distance to achieve the initial features for baseline sequence. Lu et al. [130] introduced a low-computing cost approach focusing on differences in the integral projection (IP) of sequential ME frames for detection. Li et al. [112] published a Spontaneous Micro-Expression Database (SMIC) offering the benchmark observations for micro-expression detection and recognition. The authors observed that not every subject exhibited micro-expressions while capturing samples for SMIC. Micro-expression detection was carried out in a two-class classification process by differentiating the micro-expression clip from a randomly selected non-micro-expression clips. Throughout this analysis, the researchers applied an Active Shape Model (ASM), which normalises and monitors all faces to focus on spatial feature variations and Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) for extraction of features. In order to facilitate feature extraction, Temporal Interpolation Model (TIM) was implemented to adjust frame numbers, and the classification was performed using Support Vector Machine (SVM).

Huang et al. [77] implemented Spatio-temporal completed local quantization patterns (STCLQP) to perform the detection by extracting sign, magnitude and orientation as features. The movement magnitude across frames has been used by Borza et al. [121] with simple absolute frame variations, along with an Adaboost algorithm to identify micro-expression frames. Deep learning based techniques have been used for micro-expression detection in the past. Li et al. [113] introduces a deep multi-task approach with HOOF analysis for ME detection, using CNN for preprocessing ME data to recognise the location of the facial landmarks and split the facial area into regions of interest. The sliding window based technique proposed by Borza et al. [15] preserves the present frame, past and future frame at equivalent intervals, with the discrepancy between these being given to the CNN that categorises the period as ME or non-ME. In addition, Zhang et al. [242] have implemented deep learning to identify MEs from longer videos for the first time. A novel convolutional neural network called SMEConvNet (Spotting Micro-Expression Convolutional Network) was developed for extraction of features from

video clips. For a long video apex frame spotting, the feature matrix processing method using sliding window was also proposed to consider micro-expression characteristics in order to search for the apex frame.

Besides the existing deep learning systems, the network can also focus on certain facial regions by incorporating the attention mechanism to the micro-expression recognition architecture. Attention enhances the representation of interests besides simply showing where to concentrate. Fernandez et al. [136] introduced a CNN-based end-to-end approach utilising attention methodology to address facial expression recognition problems for representation and classification jointly. Likewise, in 2D+3D FER, Jiao et al. [83] suggested enhanced facial attention-based convolutional neural network (FA-CNN). The facial attention mechanism allows the network to automatically identify the discriminative regions without dense landmark annotations from multi-modal expressions. Wang et al. [205] designed a novel attention model named micro-attention to help emphasise on the facial region of interest. For precise micro-expression recognition, Yang et al. [232] applied visual attention to developing an attention-based CNN network called MERTA. Although attention has been actively applied for face recognition [117], facial expression recognition [83, 136] and recently also for micro-expression recognition [205, 232], it has not yet been considered for facial micro-expression detection.

Nonetheless, the aforementioned studies are correlated with micro-expression recognition while the actual work for detecting micro-expression using attention network has never been studied. The occurrence of micro-expression in small sections of the face and the insufficient size of available repositories hinder the precision of recognition. In this work, we propose a mechanism to incorporate attention network for micro-expression detection with the available amount of data samples.

3.3 LGAttNet Detection Model Description

LGAttNet model is the first to utilise the sparse representation and attention mechanism for micro-expression detection. The first phase of our architecture is the pre-processing phase. Subsequently, the pre-processed image is utilised in three different ways. The image is initially divided into two parts, where the first part focuses on the eye regions and the second part focuses on the mouth region. Further, the full facial image is used for processing. The output of the final module represents the results in different metrics, predicting the existence of micro-expressions.

Our proposed model, dual-stream LGAttNet, comprises five modules: Sparse module

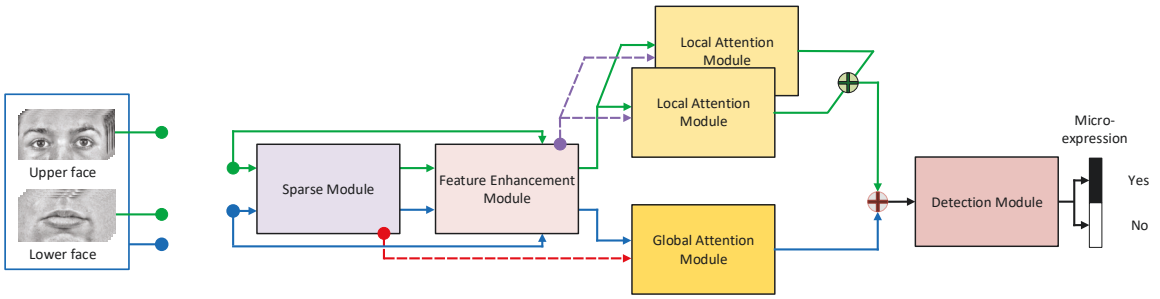


Figure 3.1: Local and Global Attention Network (LGAttNet). The flow of the upper and lower face images are indicated by the green color arrows, while the flow of entire face is indicated by blue arrows. Both dotted lines indicate the attention inputs to global and local attention modules.

(SM), Feature Enhancement module (FEM), Local Attention module (LAM), Global Attention module (GAM) and Detection module (DM) as presented in Figure 3.1.

The input to the LGAttNet is a pre-processed video frame. The input pre-processed face image is divided into two sub parts: upper face, focusing mainly on the eyes and eye-brows muscles of the face, and lower face, which concentrates on the mouth section of the face. Each of these face sections, as well as, the whole face region are given as input to the SM, which is a deep CNN network without fully connected and classifier layers. Deep features from SM along with the respective input image are fed into the FEM, a shallow network with a Concatenation and Sigmoid functions. Each LAM shown in the diagram is a dedicated attention module for upper and lower face parts respectively. There are two input values given to both LAMs. The first feature vector is the output vector of FEM Sigmoid layer for upper and lower face regions, respectively, after concatenating with the corresponding input facial RoI-based features. The second input feature vector is the feature vector before the input image concatenation, which is the output from the last Convolutional layer for each face image part. Apart from focusing on local face parts, SM and FEM also processes full face image and feeds the feature vectors from SM and FEM as input to the GAM, the third attention module. The inclusion of GAM helps to preserve the relationship between the upper and lower face. At the end, the output vectors from upper and lower face LAM and GAM are concatenated and given to the DM, which is a traditional deep neural network with a classifier layer to predict the existence of the micro-expression in the input video frame.

Following subsections explain these modules in detail.

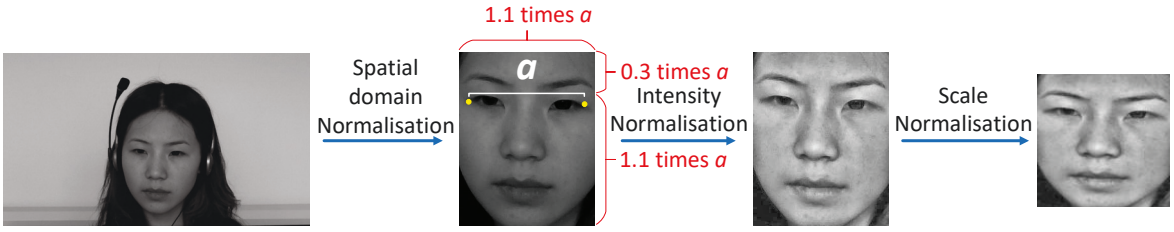


Figure 3.2: Pre-processing steps: *Spatial domain normalisation* is achieved through the difference "a" between facial feature points 37 and 46 of the active appearance model (AAM). The *intensity* and *scale* normalisations are performed subsequently [196].

3.3.1 Pre-processing

The input video frames are translated to grayscale during the pre-processing stage so that the cross-database variations between the video frames can be minimised. The descriptor involves two essential pre-processing steps: (a) data augmentation or producing synthetic samples and (b) normalisation. A series of synthesised frames are created in large numbers in the data augmentation phase, to enhance the amount of video frames, particularly for training with a deep learning model that usually requires larger dataset. Random noise is applied to the centre of the eyes and nose regions of the face using a 2D Gaussian distribution to generate synthetic frames, following the method stated in [183]. A micro-expression detection module is trained using individual frames.

Inspired by the work in [196], a set of normalisation operations are then carried out in a sequence. Initially, a region of interest (ROI) is chosen in the process of spatial normalisation for feature extraction, which excludes the insignificant areas of the video frames. This process discards the background details as well as some facial areas like ears, chin and forehead, since these regions represent no particular information regarding micro-expressions. The distance denoted as "a" in Figure 3.2, between the active appearance model (AAM) points 37 and 46 is used to crop the facial region. Secondly, using Contrast Limited Adaptive Equalization (CLAHE) [249] approach, an intensity normalisation step is implemented on every video frame to minimise the feature vector variance. One benefit of CLAHE is that the histogram segment that goes beyond the clip boundary between all histogram bins is redistributed instead of merely deleting it. A Rayleigh distribution with a cap of 0.01 and α value of 1 is chosen for this function. Thirdly, the video frames, in the scale normalisation phase, are downsized by linear interpolation to 128×128 pixels. Scale normalisation makes it possible for the same facial feature points to co-exist roughly at the same position in different video frames.

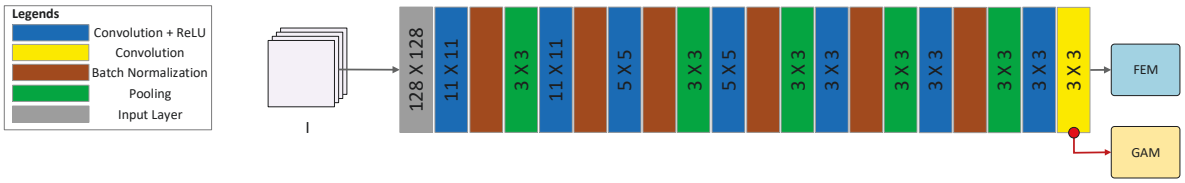


Figure 3.3: The sparse module

3.3.2 LGAttNet Components

The SM and FEM are two deep 2D-CNN components of the proposed architecture. Further, a carefully crafted attention network consisting of 2D-convolutional and sigmoid layers is placed alongside SE and FEM composed of deep CNN, while eliminating the final Softmax layer to give room for the Detection module.

3.3.2.1 Sparse Module (SM)

The image processing applications expect the input data to be represented in as few components as possible for faster processing. Sparse coding technique is useful for solving the classification problems where specific dictionaries for respective classes are built and the input is processed to classify the dictionary corresponding to the sparsest representation. In recent years, sparse coding is applied to a variety of image processing and computer vision applications, such as image classification [235, 236], image de-noising [162], compression [129], inpainting [129], object tracking [227] and other applications.

In general, SM is implemented using a deep convolutional neural network without fully connected and classification layer. In a similar way, our SM architecture is implemented using a deep convolutional neural network consisting of seven convolutional layers with one additional last layer, i.e. eight, convolutional layer to extract the attention map of the input image as depicted in Figure 3.3. The input to the SM is an image I and the output is the feature map M_{sm} processed from the input image.

$$M_{sm} = f_{sm}(I) \tag{3.1}$$

where, f_{sm} is the function of SM. The SM outputs three feature maps for three different images namely; upper face RoIs, lower face RoIs and the whole face image. The feature maps of the whole face image generated by SM are fed to FEM, and as an attention input to GAM.

3.3.2.2 Feature Enhancement Module (FEM)

The FEM is a shallow 2D convolutional neural network with two traditional convolutional blocks (Figure 3.4). The input to this FEM is the output feature vector of the sparse representation module. Similar to the SM, FEM also consists of an additional convolutional layer to process the attention mapping using the extracted features from SM and FEM together. The respective sparse representation of upper and lower face regions is again convolved to obtain a feature vector, and extracted from the third convolutional layer. The output of the last convolutional layer preserves the sparsity of the image by not losing much high end representation. This extracted feature vector is passed on to the next relevant LAM.

The FEM, further, extends its performance by incorporating concatenation and Sigmoidal functions. As the name suggests, this module enhances the collected features by using the concatenation function to integrate the feature vector extracted after the third (or last) convolutional layer of FEM and the feature mapping of the input image (which includes upper, lower and entire face image) to generate a new enhanced feature vector. The next step is to pass the enhanced feature vector to the sigmoid function. Basically, sigmoid function is used because it ranges between 0 to 1. Therefore, it is used for the models where the output is the probability value for prediction. As the probability range is only within 0 and 1, thus sigmoid is the choice for the model.

$$I_{fem} = f_{fem}(M_{sm}) \tag{3.2}$$
$$M_{fem} = S(Concat(I_{fem}, I))$$

where, the f_{fem}, $Concat$ and S are the FEM, concatenation and sigmoid functions, respectively and feature map M_{fem}.

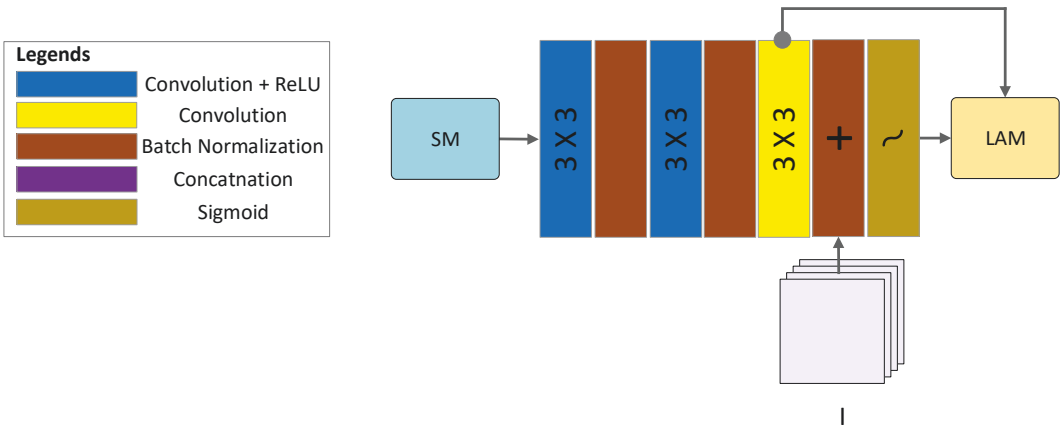


Figure 3.4: The feature enhancement module

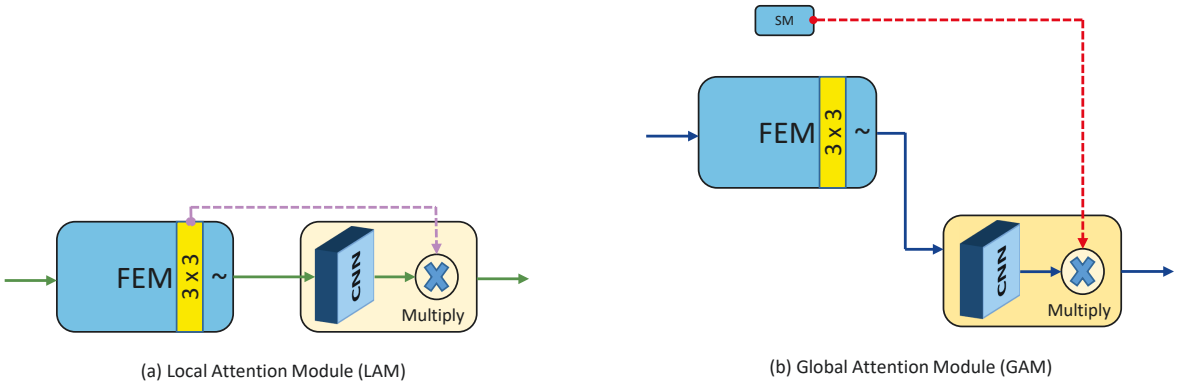


Figure 3.5: Local and Global Attention module

The prediction probability for upper and lower facial region images forms the input to the LAMs and that of the full face image is given as input to GAM.

3.3.2.3 Local Attention Module (LAM) and Global Attention Module (GAM)

The LGAttNet is built using three attention blocks: two LAMs and one GAM. The implementation of both LAM and GAM is transformed as follows.

$$A_{gam} = M_{fem} \times C(M_{sm}) \tag{3.3}$$
$$A_{lam} = I_{fem} \times C(M_{fem})$$

where C is the convolution layer function in both attention modules, and A_{gam}, A_{lam} are the attention feature maps for GAM and LAM respectively. The two LAM processes the feature maps for upper face and lower face input images respectively, whereas the GAM processes the feature vector for the full face. LAM focuses on the local facial regions while GAM considers the complete face and preserves the relationship between the upper and the lower part of the face while displaying a micro-expression. After constructing the attention feature maps, two feature level fusions are utilised, in order to get the resultant input to the detection module. The resultant feature map A_{res} is given by:

$$A_{res} = A_{gam} + \left(A_{lam}^{upper} + A_{lam}^{lower}\right) \tag{3.4}$$

where, A_{lam}^{upper} and A_{lam}^{lower} are the attention feature maps produced for upper and lower parts of the face, respectively.

The LAM and GAM are composed of two components, as can be seen from Figure 3.5; (1) 2D-CNN with one convolutional block and (2) a multiplication function. The

architecture for LAM and GAM is the same but the input given is different, and this is what makes each of the block function and process the input differently. All the three attention blocks accept two input feature vectors. The first input to both the LAM is the sigmoid function probability output of the FEM for the upper and lower facial region image respectively (solid green arrows towards LAMs in Figure 3.1). This input feature vector given to respective attention blocks is then made to pass through a 2D-CNN with one convolutional block within the attention module. The second input is the feature representation extracted from the last convolutional layer of FEM (purple dashed arrows towards LAMs in Figure 3.1). The convolved feature vector and the second input are then forwarded to the multiplication function to get a final representation of the respective upper and lower face regions.

Similarly, the sigmoid function output of FEM for full face image is passed to the Global Attention module (GAM) as a first input (solid blue arrow towards GAM in Figure 3.1). The first input is convolved when given to GAM. The second input is the features collected from the last convolutional layer of SM (dashed red arrow towards GAM in Figure 3.1). The output of CNN within GAM is then forwarded to the multiplication function along with the second input. Upon multiplication, a new feature representation is generated.

It should be noted that the second input for all the attention blocks is directly provided to the multiplication function, whereas the first input is convolved. The output vectors from all the attention blocks are added to form one vector and given to the Detection Module for the final detection of micro-expression.

3.3.2.4 Detection Module (DM)

The DM in LGAttNet consists of three fully connected (FC) layers in the size of 1024, 1024 and 512. In addition, a softmax classification layer is attached at last to perform the classification task. The estimation of micro-expression using DM is explained in Eq. 3.5.

$$\hat{y} = f_{dm}(A_{res}) \tag{3.5}$$

where, $\hat{y}$ is the prediction of an image sample.

3.3.3 Loss Function

In order to train the proposed model, the degradation function is incorporated. Generally, the binary cross entropy (BCE) performs better in closed set classification tasks. Hence,

the BEC is used in the proposed approach to estimate the classification loss, as given below.

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=0}^{N}(y \times \log \hat{y}_i) + (1-y) \times \log(1-\hat{y}_i) \tag{3.6}$$

where, y and $\hat{y}$ are the micro-expression label and the predicted value, respectively.

3.4 Experimental setup and Outcomes

The validation and efficiency of LGAttNet is verified by testing the model on some of the publicly available benchmark micro-expression databases. Apart from the model testing, the effectiveness of implementing local and global attention networks in a dual-stream pattern is demonstrated using an ablation study.

3.4.1 Datasets used

The experiments are performed on popular micro-expression databases including the Chinese Academy of Sciences Micro-Expression databases: CASME [231] and CASME II [229], Chinese Academy of Sciences Macro- and Micro-Expressions (CAS(ME)^2) [167] and Spontaneous Actions and Micro-Movements (SAMM) [29]. Details of these spontaneous micro-expression databases used in the experiments are given below.

3.4.1.1 Chinese Academy of Sciences Micro-Expression (CASME)

Introduced by the Chinese Academy of Sciences, CASME database [231], is one of the spontaneous micro-expression database widely used. CASME comprises of two subsets A and B totalling up to 195 micro-expression samples collected from 19 participants. These samples were recorded at 60 fps. These participants experience a great emotional stimulation and conceal their facial expressions. The video clips in dataset A were captured in natural light with a resolution of 1280×720 pixels. In dataset B, the video samples were recorded at 640×480 pixel resolution under LED lighting. Each sample was tagged with onset, apex and offset frames, action units (AUs) labelled and emotions correctly identified by psychologists. The database has a collection of eight micro-expression categories: contempt, disgust, fear, happiness, repression, sadness, surprise and tense.

3.4.1.2 Chinese Academy of Sciences Micro-Expression (CASME II)

As an extension to the original CASME [231] dataset, CASME II [229] was introduced. There are 247 micro-expressions newly coded with FACS captured from 26 participants under a high temporal resolution of 200 fps and a 280×340 pixels spatial resolution on the facial region to examine muscles movements in greater detail. Every video session is a short clip for several seconds which has onset, apex and offset frames marked for micro-expressions and FACS and emotion type is annotated.

3.4.1.3 Chinese Academy of Sciences Macro- and Micro-Expressions (CAS(ME)^2)

There are 87 long video samples with an average duration of 148 secs collected from 22 participants in Part A of CAS(ME)^2 database [167]. Macro and micro are the two expression types where the collected video samples can involve various macro or micro facial expressions. The authors have published an excel spreadsheet file pointing out the onset, apex and offset frame index for these expressions. Furthermore, onset and offset time for the eye blinks are also marked.

3.4.1.4 Spontaneous Actions and Micro-Movements (SAMM) in Long Videos

There are altogether 32 participants, in the SAMM database [29], recording seven video samples, each with an average duration of 35.5 secs. The first SAMM update features micro-movement sequences tagged with Action Units (AUs). The study in [28] has newly adopted objective and emotion classes for the database. The spotting task emphasises on 79 videos where every video includes one or several facial micro-movements summing up to 159 micro-movements. As the ground truth, the onset, apex, and offset frame indices of micro-movements are given wherein the micro-movement duration lasts between the onset and the offset frame. All micro-movements in this database are annotated. The identified frames can, therefore, signify micro-expressions along with other facial movements, including blinks of the eyes.

3.4.2 Experimental setup and Parameters

Model implementation is done using an open-source platform, Tensorflow. The model is trained and tested on a GPU Server with NVIDIA GeForce GTX 1080 Titan processor. LGAttNet uses SGD optimisation technique with initial learning rate of 0.001 and L2 normalisation is implemented to prevent overfitting. The input dimensions for LGAttNet

model is 128×128. The other parameter values are weight decay of 0.0005 and momentum of 0.9. The model training is executed for 100 epochs.

The results for the conducted experiments are reported using Accuracy, Area Under Curve (AUC), F1 score, Recall and Precision.

3.4.3 Outcomes and Analysis

The observations are drawn by performing analysis on four publicly available databases namely, CASME, CASME II, CAS(ME)^2 and SAMM. The inputs to the network are images or video frames from micro-expression databases. In order to confirm the effectiveness of the LGAttNet model, five metrics i.e. Accuracy, Precision, Recall, F1-score, and Area Under Curve (AUC) are chosen as evaluation metrics for binary classification.

3.4.3.1 Outcomes

The input to the network are the images or the video frames from the micro-expression databases. The evaluation of the network is conducted using the Leave-One-Subject-Out Cross-Validation (LOSOCV) technique, i.e. one subject is selected that is not used for the training process, and the network is evaluated on this unseen subject. The training and testing set consists of two classes, wherein one class has micro-expression video frames, and the other class has neutral face video frames (non-ME frames). The macro-expression samples are eliminated from our experiments. As this is a subject independent evaluation, one subject is wholly left out of the training process.

Table 3.1 demonstrates the experiment outcomes for all the performance metrics. The table depicts that the sparsely represented multi-attention micro-expression detection architecture is capable of achieving significantly high, which is in the range of 87% to 94%, detection accuracy for different databases. As can be seen, the recognition accuracies obtained on CAS(ME)^2 and SAMM datasets are lower compared to other micro-expression datasets. It is clear that the samples in SAMM datasets contain wide variations, such as included subjects from different nationalities.

Additional experiment is performed to demonstrate that the LGAttNet works equally well on a sequence of video frames for detecting micro-expression frames. In this experiment, during the testing phase, the input given to the trained LGAttNet is a series of images from a video. LGAttNet then processes the video frame-wise and classify individual frame as ME or non-ME frame, as shown Figure 3.6. The graph illustrates that the LGAttNet detects the micro-expression frames (green line in the graph) from

Table 3.1: LOSOCV Micro-Expression detection outcome using various performance metrics

Metrics \ Database	CASME	CASME II	CAS(ME)^2	SAMM
Precision	0.948	0.944	0.850	0.851
Recall	0.915	0.940	0.885	0.890
F1-score	0.931	0.942	0.867	0.870
Accuracy	0.932	0.942	0.865	0.867
TPR	0.915	0.94	0.885	0.89
FPR	0.05	0.055	0.155	0.155
AUC	0.931	0.912	0.923	0.846

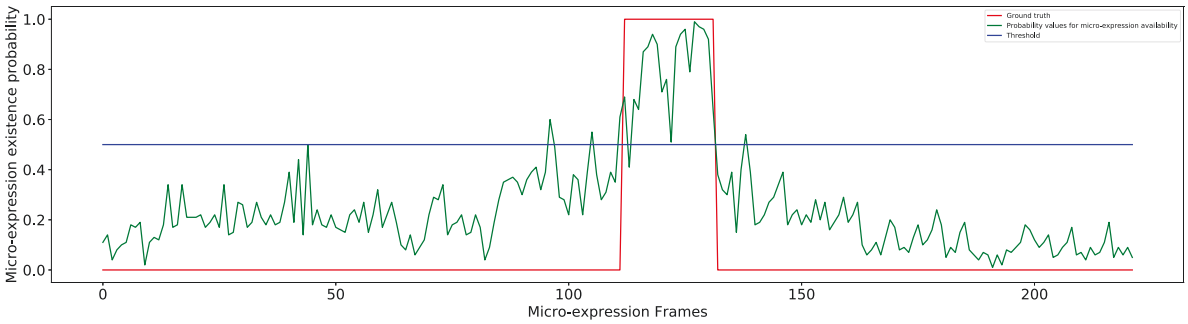


Figure 3.6: LGAttNet tested on a micro-expression sequence from CASME database. The green line indicates the generated probability values for the existence of micro-expression.

the sequence of video frames in alignment with the ground-truth (red line in the graph). Therefore, it can be seen that the LGAttNet model trained on images is capable of accurately predicting the micro-expression frames from videos.

3.4.3.2 Ablative Analysis

Table 3.2 shows the performance of ablative analysis on all the databases. For undertaking these analyses, the model is modified and evaluated by including or removing

Table 3.2: Ablative evaluation with and without (w/o) different modules of the framework

	CASME	CASME II	CAS(ME)^2	SAMM
LGAttNet	0.933	0.943	0.865	0.868
LGAttNet w/o GAM	0.892	0.915	0.812	0.822
LGAttNet w/o LAM	0.835	0.878	0.785	0.788
LGAttNet w/o GAM and LAM	0.735	0.782	0.692	0.728

each of the LAM and/or GAM component in the architecture. From the table, it can be observed that the LGAttNet performs well when it includes LAM and GAM on all the databases. The next evaluation is performed by removing the GAM module, where the results reveal a drop by 3% – 5%. The probable reason is that the network is unable to find the relation between the two individual local feature maps extracted from upper and lower face regions when the global attention module (GAM) is not present in the architecture, resulting in performance degradation.

However, it is noticeable that removing LAM from the model has a significant effect on the results. The performance is seen to deteriorate more when the LAM is removed, while the GAM is included in the system. This shows that acquiring the local level features from the facial regions assists in interpreting micro-expression in the input image. Finally, removing both GAM and LAM decreases the results by more than 20%. These results demonstrate the importance of our proposed LAM and GAM modules for correctly detecting micro-expressions.

In addition to the ablative analysis, as seen in Table 2, Figure 3.7 visualises the influence of utilising attention mechanism in LGAttNet for Disgust micro-expression. The purpose of incorporating attention mechanism in LGAttNet is to focus the attention towards specific facial regions to identify the presence of micro-movements and classify the input image as ME or non-ME frame. Similar to ablative analysis, Figure 3.7 displays the attention mapping for LGAttNet with and without LAM or GAM or both. It can be seen from Figure 3.7 (b), which is the activation map for LGAttNet without both LAM and GAM, that the activation map is scattered all over the facial region. The model was unable to highlight the specific facial regions of movements without the attention mechanisms. However, Figure 3.7 (c) is an implementation of LGAttNet with LAMs

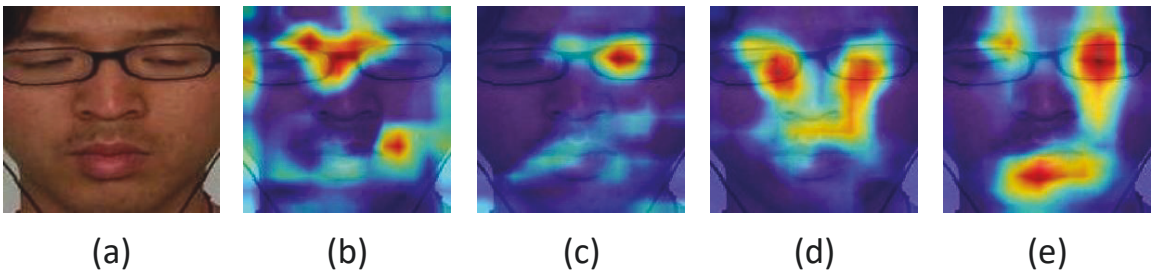


Figure 3.7: LGAttNet attention visualisation on a Disgust sample from CASME dataset. (a) Original image; (b) LGAttNet without LAM and GAM; (c) LGAttNet without GAM; (d) LGAttNet without LAM; (e) LGAttNet with LAM and GAM

and without GAM. We can observe that the LAM serves its purpose of concentrating on the specific regions of the face, which is eyes and mouth in this case. Figure 3.7 (d) is the activation map for LGAttNet with GAM and without LAMs, which highlights the central region of the face, including eyes, nose and upper lip. This indicates that unlike LAM, which processes upper and lower face individually, GAM has to process the full face. Hence, it is forced to process the facial regions which do not carry any movement and highlighting regions of non-interest. Finally, Figure 3.7 (e) is our complete model of LGAttNet with LAM and GAM. The activation map illustrates that the LAM that processes upper and lower face separately can predict the region of interest which carry a micro-expression and GAM which processes the full face correlates to the LAM features to precisely mark the facial regions with motion, eliminating any unnecessary details, classifying the input frame correctly as ME frame.

3.4.3.3 Cross-Database Analysis

The model effectiveness is also shown by conducting experiments on cross databases. The purpose of performing the cross-database evaluation is to justify that even though this architecture is a supervised learning model, it is capable of detecting micro-expressions from the images of altogether different databases. Cross-database micro-expression detection is where the training and testing samples come from two different micro-expression databases collected by different cameras or under different environments. The databases used in these experiments also has diversity in ethnic backgrounds of the participants, for instance, SAMM database has subjects from 13 different nationalities. This type of validation offers a good way to mimic the scenarios the micro-expression detection system would encounter in reality. Therefore, it is worthy to investigate this more carefully.

In the experiments, four micro-expression databases are employed. In here, training of the network is performed on one database, and randomly selected samples from the other databases are used for testing. This is repeated four times, and the respective results are presented in Tables 3.3-3.6. These results demonstrates that our proposed architecture is generic, and capable of handling cross-database. The observations of the cross-database evaluations are elaborated in Section 3.4.4.

3.4.3.4 Comparison with state-of-the-art

Table 3.7 compares the proposed technique with the existing approaches. It should be noted that the experimental configurations for the methods compared may differ.

Table 3.3: Cross-database Micro-Expression detection network trained on CASME and tested on other databases

Metrics \ Database	CASME II	CAS(ME)^2	SAMM
Precision	0.866	0.829	0.768
Recall	0.845	0.850	0.745
F1-score	0.855	0.839	0.756
Accuracy	0.857	0.837	0.760
TPR	0.845	0.850	0.745
FPR	0.130	0.175	0.225
AUC	0.861	0.802	0.715

Table 3.4: Cross-database Micro-Expression detection network trained on CASME II and tested on other databases

Metrics \ Database	CASME	CAS(ME)^2	SAMM
Precision	0.891	0.810	0.733
Recall	0.825	0.835	0.730
F1-score	0.857	0.822	0.731
Accuracy	0.862	0.820	0.732
TPR	0.825	0.835	0.730
FPR	0.100	0.195	0.73
AUC	0.891	0.781	0.705

Table 3.5: Cross-database Micro-Expression detection network trained on CAS(ME)^2 and tested on other databases

Metrics \ Database	CASME	CASME II	SAMM
Precision	0.848	0.834	0.742
Recall	0.810	0.855	0.720
F1-score	0.828	0.844	0.730
Accuracy	0.832	0.842	0.735
TPR	0.810	0.855	0.720
FPR	0.145	0.170	0.250
AUC	0.849	0.842	0.710

The results for the existing approaches are taken directly from the respective research studies. Many studies have been tested only using CASME II and SMIC repositories to detect micro-expressions. LGAttNet is also trained and tested on CAS(ME)^2 and SAMM databases in addition to the commonly used databases for micro-expression detection. It can be found from the comparison table Table 3.7 that the LGAttNet achieves considerably higher detection accuracy.

Table 3.6: Cross-database Micro-Expression detection network trained on SAMM and tested on other databases

Metrics \ Database	CASME	CASME II	CAS(ME)^2
Precision	0.668	0.678	0.715
Recall	0.685	0.645	0.680
F1-score	0.676	0.661	0.697
Accuracy	0.672	0.670	0.703
TPR	0.685	0.645	0.680
FPR	0.340	0.305	0.272
AUC	0.631	0.665	0.697

Table 3.7: Comparison with existing state-of-the-art micro-expression detection methods

Database	Method	Performance	Accuracy
CASME	Feature difference [15]	TPR=77.27%	-
	(CASME-A) LBP-χ^2 [105]	-	78.75%
	(CASME-A) LTP-ML [105]	-	77.90%
	(CASME-B) LBP-χ^2 [105]	-	82.92%
	(CASME-B) LTP-ML [105]	-	82.61%
	LGAttNet	**TPR=91.5%**	**93.2%**
CASME II	Frame difference [121]	-	81.75%
	Frame difference [14]	-	86.95%
	LBP-χ^2 [105]	-	64.08%
	LTP-ML [105]	-	65.07%
	LBP [111]	TPR=70.0% FPR=13.5%	-
	PLK+LSTM [34]	-	89.87%
	LGAttNet	**TPR=94.0%** **FPR=5.5%**	**94.2%**
CAS(ME)^2	LTP-ML [107]	F1-score=0.0055	-
	LBP [167]	AUC=0.5971	-
	LGAttNet	**F1-score=0.867** **AUC=0.923**	**86.5%**
SAMM	3D HOG-XY plane [27]	-	70.87%
	LBP-TOP -XY plane [27]	-	74.65%
	HOOF [27]	-	70.98%
	LTP-ML [107]	F1-score=0.0316	-
	PLK+LSTM (SAMM+CASME II) [34]	-	87.30%
	LGAttNet	**F1-score=0.870**	**86.7%**

TPR = True Positive Rate; FPR = False Positive Rate
[34] implements cross-database experiment (SAMM+CASME II)

3.4.4 Discussion

LGAttNet is built using attention networks which are made to focus on three different sections of an input facial image, making this as the first attempt to use attention network for local as well as global attention mapping for a facial image. Unlike some other related studies, LGAttNet also stands out for achieving profoundly high micro-expression detection accuracy from video frames. There have been some studies [121, 123] performed to recognise facial micro-expressions from single apex frames. Taking motivation from these works, a detection model is constructed to identify micro-expression images from non-micro-expression images. The observations in Table 3.1 demonstrates that the LGAttNet is capable of detecting the existence of the micro-expression in the video frames. Usually, detection of micro-expressions is performed by taking feature differences between consecutive frames or comparing the first reference frame with the rest of the frames in a video using handcraft feature descriptors. In contrast, the LGAttNet uses a deep attention network that can concentrate on the local as well as global facial regions to track the feature difference to detect micro-expressions.

It can be contemplated from the results in Tables 3.3-3.6 that the databases having participants from similar ethnic background, as in CASME, CASME II and CAS(ME)^2, display higher prediction accuracy when trained on one of these databases as compared to the other database (SAMM), which has participants from 13 different nationalities. Moreover, when trained on SAMM, the detection accuracies for CASME, CASME II and CAS(ME)^2 declines as the training database includes only three Chinese participants which are in contrary to the other databases. From this cross-database analysis, it can also be understood that the people from different ethnic backgrounds have their unique ways of hiding real emotions. Hence, it is not only the way of capturing these micro-expressions that affects the accuracies, but also the ethnicities of the participants' plays an important role.

3.5 Summary and Future Direction

In this work, a deep learning model is designed to focus on specific facial regions and establish a correlation between these regions and the whole facial area. The proposed model, LGAttNet, is a micro-expression detection model which incorporates the attention network to converge the network processing towards selected regions of the face. The LGAttNet comprises a deep and a shallow CNN supported by local and global attention networks and an Artificial Neural Network (ANN) for binary classification. The local at-

tention network processes partial facial parts, and the global attention network operates on the complete facial image. This model is an image-based supervised detection model with non-ME and ME classes.

Our model is the first to implement attention network for micro-expression detection. As compared to the available number of related state-of-the-art micro-expression detection works, LGAttNet model delivers exceedingly higher detection accuracy around more than 9%. This is because of the inclusion of the attention nets, since the micro-expression is more of a spatial feature and with the partitioning of the face, the attention nets insist the network to focus on selected facial regions. This behavior of LGAttNet can be observed from the ablative analysis, where on the removal of the local attention module (LAM), the detection accuracy is negatively affected and also on entirely removing the attention modules (LAM and GAM) the network accuracy drops significantly.

Cross-database evaluations are conducted to explain the robustness of the proposed network and to demonstrate that this model can be useful for real-time processing. We are currently working on making this model capable to implement on the video sequences considering the temporal dimension that will benefit the real-time processing industry in the near future. In the future, the model would be extended to spot micro-expressions from live-stream videos.

CHAPTER 4

EFFECTIVE FACIAL FEATURES FOR RECOGNITION

Existing research on micro-expression recognition has mainly used hand-crafted features, for example, Local Binary Pattern-Three Orthogonal Planes (LBP-TOP), Gabor filter and optical flow. Recently, Deep Convolutional neural systems have demonstrated a high degree effectiveness for difficult face recognition tasks. This chapter explores the potential usage of deep learning for recognition of micro-expressions. In this chapter, we intend to develop deep learning models which can recognise the micro-expressions from static images as well as videos. For micro-expression recognition from the static images we used a pre-trained network and perform fine-tuning using the micro-expression databases. The fine-tuned model is then used for recognition. However, to recognise micro-expressions from the video or video frames another model is developed. The technique incorporates handcrafted features and deep features. Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) is the handcraft feature which combines spatial and time analysis to encapsulate regional facet movements. Two classifiers i.e. Softmax and SVM are trained with combined feature vectors generated by LBP-TOP and CNN feature descriptors. However, to develop a reliable deep neural network extensive training sets are required with a huge number of labelled image samples. Micro-expression recognition is a challenging task owing to the repressed facial presentation and limited span, which results in the lack of training data. We also propose to generate extensive training datasets of synthetic images using data augmentation on widely used micro-expression databases. Then, these datasets are combined to tune the developed CNN-based micro-expression recogniser. The findings of the experiments show that the proposed methods,

although simple and straightforward, achieves a substantial increase in precision relative to other commonly recognised micro-expression techniques, which are trained and tested with just a few datasets.

4.1 Introduction

In recent years, identification or classification of facial micro-expression was at the forefront of computer vision. The main focus of existing research on micro-facial expression is on recognising seven universal human feelings (anger, contempt, disgust, fear, happy, sad and surprise) [40]. This recognition is often complicated because there is only a minor variation between different micro-expressions, which requires the training of a strong and profound feature extractor. Moreover, [150, 192, 193] studies have pointed out that the uneven distribution within emotional classes in the classes with lesser samples can lead to low precision.

By using transfer learning, the restriction of the information scarcity can be eliminated [159]. Patel et al. trained the CNN macro-facial expression CK+ network in this strategy. Campos et al. [18] investigates how CNN may be fine-tuned and used for prediction of visual feelings. In the light of the challenge of gathering big datasets with accurate micro-expression annotations, the majority of today's scientists concentrate on domain comprehension by analyzing the efficiency of state-of-the-art architecture that is tailored to this challenge.

The overall micro-expression recognition systems involve three principal steps: first, the facial regions of interest are selected, second, classification characteristics are set out and extracted and, third, the real micro-expression recognition is carried out using selected features and advanced algorithm for machine training. Face representations can traditionally be divided into spatial / spatio-temporal classes [191]. Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) [164], Oriented Optical Flow Histogram [186], SpatioTemporal Completed Local Quantization Pattern (STCLQP) [77] are a few existing handcrafted methods. Despite of many existing robust handcrafted feature descriptors, the advanced high-level deep learning approaches attracted huge attention from recent researches [95, 102, 190]. Recognition baselines using the handcrafted feature descriptors on various micro-expression databases were established with the original works [29, 112, 155, 167, 229]. The deep learning algorithms uses convolutional neural networks (CNN) for extracting advanced deep feature for micro-expressions [88, 159, 161].

In this chapter, we have discussed the contributions of two different experiments conducted on micro-expression images and videos respectively as given below:

1. Experiment I: Image-based facial micro-expression recognition using deep learning on small databases [192]

    * We develop a CNN architecture that achieves satisfactory recognition accuracy on micro-expression images.
    * We also present a novel way of increasing the number of samples for training CNN model by combining the two widely used databases CASME and CASME II.

2. Experiment II: Manifold Feature Integration for micro-expression recognition [193][4]

    * Model which accomplishes concatenation of handcrafted feature with deep CNN feature and conducts experiments to include publicly available CASME, CASME II, CAS(ME)^2, SAMM and SMIC (HS, NIR, VIS), and CASME+2 micro-expression datasets to confirm model efficiency and resulting uniformity on wide range of datasets. In addition, we have trained SVM classifier with its four kernels: Linear, Polynomial, Radial Basis and Sigmoid together with Softmax and observed that SVM has improved the precision of micro-expression recognition for some databases than Softmax.
    * This is the first work to include and test the framework on all the publicly available micro-expression datasets. The framework performs equally well with all the micro-expression datasets, regardless of the data collection resolution and frame rate or the ethnicity, age or gender of the individuals.
    * Implementing data augmentation is a means of resolving the data imbalance issue through the generation of synthetic data for all categories to enhance the amount of training specimens in general. The method employs horizontal flipping, doubling the training set. A slightly higher dataset is used to fine-tune the pre-trained VGGFace model achieving analogous results.

The remainder of the chapter is organised into four sections. Section 4.2 discusses the fundamental pipeline of micro-expression recognition process. The section also

[4]Madhumita A. Takalkar, Min Xu, Zenon Chaczko, *Manifold Feature Integration for Micro-Expression Recognition*, Multimedia Systems (*Under Review*)

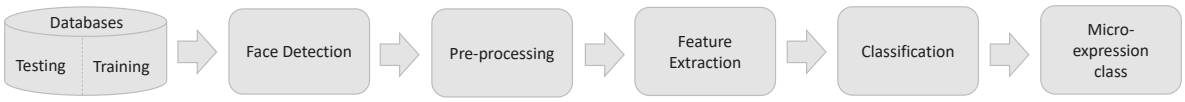


Figure 4.1: A general block diagram of micro-expression recognition system.

explains the three standard components of both the discussed models: Face detection and face registration; Data augmentation and CNN fine-tuning. Data augmentation and CNN fine-tuning are a part of the Training process, whereas face detection and registration are mandatory for Training and Testing phases. Following these steps are the main processing models for image-based recognition and video sequence based recognition which are discussed and explained in detail in Section 4.3. Section 4.4 illustrates the experimental setup, parameters and outcomes for both the experiments I and II. The proposed models are compared with state-of-the-art approaches to justify the contributions and explained with the help of performance metrics primarily using accuracy and confusion matrix along with ablative analysis for experiment II. Based on the observations, Section 4.5 discusses the proposed models and difficulties during processing. Lastly, Section 4.6 summarises the chapter.

4.2 Micro-expression Recognition Pipeline

The conventional pipeline consists of four stages: 1) face detection; 2) pre-processing; 3) feature extraction; and 4) classification. Figure 4.1 shows the basic block diagram of the micro-expression recognition.

The process of recognition can be divided into two phases as training and testing and subdivided into different phases. The first step in the training phase is the pre-processing of data. Data augmentation of training samples and fine-tuning of CNN are the two extra steps taken. The augmented data is then provided to fine-tune the pre-trained CNN model. The obtained features are given to classifiers for training. In the testing phase, the input is pre-processed and forwarded to the feature extraction block. These extracted features are given to the trained classifiers for classification of micro-expression classes in the test dataset. The detailed explanation of each block is discussed below.

4.2.1 Face detection and Face registration

The preliminary steps in the pre-processing data for micro-expression recognition are face detection and registration. The face detection stage, for our experiments, focuses

on frontal human face detection as all the databases includes videos on front faces. A sliding window detection system uses the DLib toolkit, which is based on Histogram of Oriented Gradients (HOG) and Linear SVM. It also offers pre-trained models for the detection of facial points.

After the face is detected and cropped, face alignment algorithm is applied. The process includes 1) analysis of the geometric facial structure in video frames, and 2) the approval of a translation, scaling and rotational alignment of the facial region. A technique for the face alignment is used to achieve a normalised rotation, translation, and scale representation of the face based on facial landmarks (especially the eye areas). Alignment of the face is a form of "data norming". The dataset is usually normalised before a facial recogniser is trained.

The input is a collection of facial points (input co-ordinates) with the objective of warping and translating the image into a output coordinate space. Every face in an entire dataset should be centrally aligned in a way that allows the eyes to be placed on a horizontal line (i.e. the face is rotated to straighten the eyes along the same y-coordinates). All the images should be scaled to roughly the same facial proportion (RoI) to maintain consistency in the training and testing samples.

Affine transformation accomplishes all the above mentioned tasks for registering the facial region. It determines the components of the transformation matrix. Locating the (x-y)- coordinates of the eyes are the key components used in the facial alignment algorithm. Facial points of interest tend to work better than Haar cascades as we evaluate the eye location precisely (instead of just a bounding box).

The input frames are pre-processed using the above mentioned steps of centering and aligning the facial region. Then the aligned frames are cropped to a size of 224×224 size. These cropped images are grayscaled before performing feature extraction.

4.2.2 Data Augmentation

The absence of large training datasets is a crucial bottleneck that keeps the utilisation of profound (deep) learning techniques in such cases, as the models will overfit drastically when utilising small training datasets. To address this issue, a large number of strategies have been proposed: fine-tuning models trained from other large public datasets (e.g. ImageNet [32]), using the big synthetic training datasets explored by some authors [73, 88, 110].

The results are substantially affected by the extent of datasets in the field of deep learning and thus data augmentation is frequently used to extend the training set. There

are two key groups of current data augmentation techniques: (a) a relatively universal and computationally inexpensive geometric transformation and (b) task-specific or guided techniques that can produce synthetic samples from particular labels [35].

The first set of data augmentation is invariably used in image classification to create more image information via label-preserving linear transformations (translations, rotation, scaling, flipping, horizontal shearing) such as Affine [22], elastic deformations [182], patch extraction and modification of the intensities of the RGB channels [95]. The second group proposes more complicated manually-specified augmentation strategies.

There are two critical points of interest of utilising synthetic data: (a) one can produce the same number of training samples as required, and (b) it permits explicit control over the unwanted factors. Data augmentation is a technique that is commonly used to reduce the scarcity problem. It is a set of label-preserving transforms that introduce some new instances without collecting the new data. In this, the existing training images are transformed without affecting the semantic class label. Examples of such transformations are horizontal/vertical mirroring [113], cropping, small rotations, etc. Flipping and mirroring images vertically or horizontally producing two samples of each is a commonly used data augmentation technique for face recognition.

Some training information that cover different circumstances is needed in order to better classify unseen information. Facial micro-expressions databases such as CASME, CASME II, CAS(ME)^2, SAMM and SMIC do, however, contain only a couple of hundred sequences. As there are many parameters in a typical deep network, training with fewer samples could cause the system to overfit. To overcome the overfitting problem, various data augmentation techniques are implemented. In current work, all image sequences in the training datasets are horizontally flipped, generating a mirror image of the face. Only one technique of horizontal flipping is implemented, which will merely double the training set and not increase the samples significantly. The purpose behind keeping the training samples small is that we will fine-tune the pre-trained network on small datasets and still deliver improved outcomes.

4.2.3 CNN Fine-tuning

Lack of samples for training reduces CNN-based micro-expression recognition approach performance. This problem may be handled partially by an increase in the amount of data that could overfit. Fine-tuning is therefore used to draw expression-associated functionalities from facial gray images by referring to the deeper neural network, which has achieved great success in comparable tasks. Fine-tuning helps researchers train

Table 4.1: VGGFace architecture

Conv1	Conv2	Conv3	Conv4	Conv5	Full6	Full7	Full8
conv1_1(64*3*3) relu1_1	conv2_1(128*3*3) relu2_1	conv3_1(256*3*3) relu3_1	conv4_1(512*3*3) relu4_1	conv5_1(512*3*3) relu5_1	FC6 (4096)	FC7 (4096)	Softmax
conv1_2(64*3*3) relu1_2 pool1	conv2_2(128*3*3) relu2_2 pool2	conv3_2(256*3*3) relu3_2	conv4_2(512*3*3) relu4_2	conv5_2(512*3*3) relu5_2			
		conv3_3(256*3*3) relu3_3 pool3	conv4_3(512*3*3) relu4_3 pool4	conv5_3(512x3x3) relu5_3 pool5			

neural networks with considerably less time if the conditions are met. It is one approach to transfer learning, and it is prevalent in computer vision and Natural Language Processing (NLP).

The pre-trained model used in the experiments to compute the deep micro-expression feature of the face is the VGGFace network, trained on 2.6M images of 2.6k individuals [157]. Table 4.1 describes the layers of the VGGFace pre-trained model used for fine-tuning by modifying the $Full8$ Softmax layer with the number of class labels in the respective databases. CNN VGG-Face descriptor is calculated on the basis of the VGG-Very Deep-16 CNN architecture mentioned in [157] and is assessed on Labeled Faces in the Wild [74] and the YouTube Faces [223] dataset.

Fine-tuning is usually achieved by freezing the weights of all layers of neural networks except for the penultimate layer. Usually, the last layer (Softmax) is replaced with another one of our choice (depending on the number of outputs we require for the new problem). However, for the two experiments that we have conducted in this chapter, we have considered different number of databases. For Experiment I, the training and testing of the model is carried out on images from CASME, CASME II and an additional database that we formed by aggregating CASME and CASME II, which is referred to as CASME+2 [192]. These databases consists of images labelled with one of five emotion categories: disgust, fear, happiness, sadness, and surprise. We have also taken into consideration the sixth category as 'Neutral'. The given images are divided into two different sets which are training and testing sets.

For data augmentation, mirrored images were generated by flipping images horizontally. The training set comprises of 80% of the total images in the synthetic database and remaining 20% of the images are further divided as the testing set (10%) and validation set (10%). The training process generated three fine-tuned model versions of VGGFace pre-trained model.

However, for Manifold Feature Integration that is Experiment II we have included data ranging from three to six micro-expression categories from eight different micro-

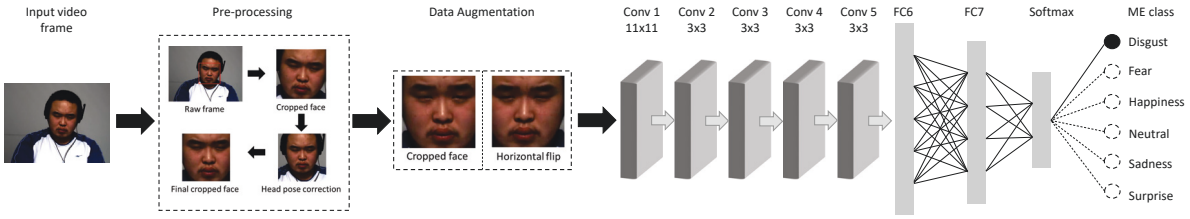


Figure 4.2: The proposed CNN architecture for image-based micro-expression recognition

expression databases such as CASME, CASME II, CASME+2, CAS(ME)^2, SAMM, SMIC (HS, NIR, VIS), so the last Softmax layer is replaced with respective (3,5 or 6) node. Each database has data distributed in different micro-expression classes. Details about the publicly available databases is discussed in Chapter 2. Hence, the VGGFace (pre-trained) network is fine-tuned on each database individually generating eight fine-tuned models.

4.3 Proposed Methods

This section discusses in detail the proposed feature extraction methods used by training and testing phases for both image-based as well as manifold feature integration experiments.

4.3.1 Image-based facial micro-expression recognition

CNN is a biologically-inspired model. The input layer receives normalised images with identical size. The convolutional layer will process a set of units in a small neighbourhood (local receptive field) in the input layer and creates a feature map. Rectified Linear Unit (ReLU) is a non-linear operation. Each feature map has only one convolutional kernel. The CNN design can effectively save computation time and allow particular feature stand out in a feature map. Typically there is more than one feature map in a convolutional layer which involves several features in the layer. To make the feature invariant to the geometrical shift and distortion, a pooling layer is followed by the convolutional layer can subsample the feature maps. Max pooling function is used for subsampling. The first convolutional layer and the pooling layer would obtain low-level details of the image, while their stack would allow for the extraction of high-level features.

The output layer acts as an input to the fully connected layer that uses a Softmax activation function in the output layer. The purpose of the fully connected layer is to use these features for classifying the input image into respective classes depending on the training dataset.

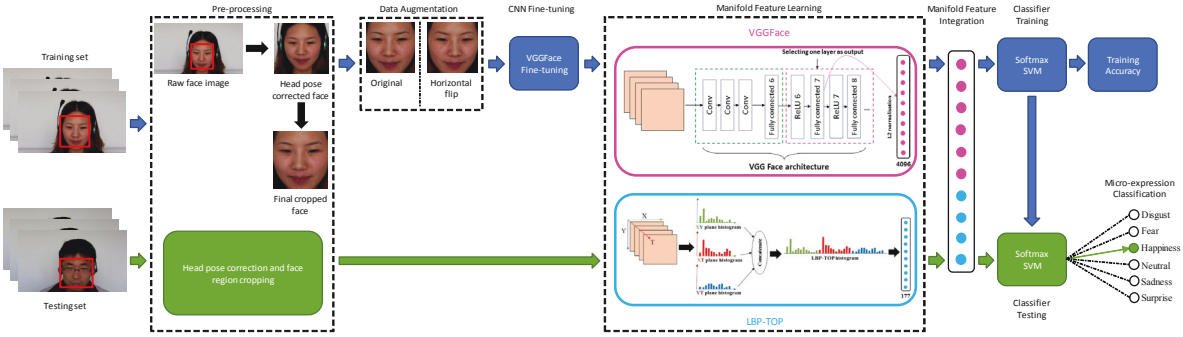


Figure 4.3: The comprehensive structure of Manifold Feature Integration model for micro-expression recognition

Putting it all together, the Convolutional + Pooling layers act as Feature Extractors while Fully Connected layer acts as a Classifier.

The CNN architecture for image-based micro-expression recognition system is depicted in Figure 4.2. This model is fine-tuned individually on each of the databases used. The three fine-tuned VGGFace models are utilised as feature extractor for any subjective face image by operating the image through the whole network, then extracting the output of the fully connected layer FC7. The extracted feature is exceedingly discriminative, minimal, and interoperable encoding of the input image. Once the features are acquired from FC7 layer of the fine-tuned VGG-Face CNN, they are utilised for training and testing subjective Softmax classifier.

4.3.2 Manifold Feature Integration Model

When recognising facial micro-expression, the motion is a cue, and it can produce a powerful countermeasure in conjunction with texture. A spatio-temporal representation, which incorporates facial aspect and dynamics, is regarded for defining the face micro-expression for classification. The proposed manifold feature integration model for recognition of micro-expression is presented in this section and describes a further outline of the process to generate synthetic training data for network. The model suggested is a multiple feature integrating method that is trained and tested using two classification methods (Figure 4.3). The Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) and the deep Convolutional Neural Network (CNN) are two feature extraction techniques. The handcraft descriptor LBP-TOP extracts the block-based spatio-temporal feature that is required to detect facial micro-expression. The CNN, a deep neural network, collects the dense local information from the region of interest (RoI). The mentioned both feature descriptors work in parallel to each other. Finally, the two feature vectors are integrated

to enhance the micro-expression recognition performance.

4.3.2.1 Manifold Feature Learning

LBP-based spatio-temporal representation performs convincingly well in modelling facial movements, recognition of facial expression and recognition of dynamic texture [165]. An LBP textural analysis operator characterised as an invariant gray-scale texture measurement obtained from normal texture definition in a local neighborhood is presented by Ojala et al. [153, 154]. This computation is expressed as [216]:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^P \tag{4.1}$$

where,

$$x_p = x_c + R_X cos(2\pi p/P), y_p = y_c + R_Y sin(2\pi p/P)$$

are the points of neighborhood, and step function:

$$s(x) = \begin{cases} 0 & if\ x < 0 \\ 1 & if\ x \geq 0 \end{cases}$$

g_c is the gray value center pixel, P reflects the total number of neighboring points, while g_p is the gray value of P equally spaced neighbouring pixels on a circle of radius R at this center pixel. An extension to three-dimensional space of LBP is referred to as LBP-TOP. LBP-TOP comprises of temporal texture (video sequence) as a value in set (X, Y, T) space where the space coordinates are marked by X and Y and time coordinates are referred to as T. The histogram containing the texture features of the image is computed on the basis of the LBP patterns derived from all the pixels in the image. A video is shown as a series of XY planes in axis T, YT planes in axis X and XT planes in axis Y, respectively. Bilinear interpolation estimates the values of neighbours that do not exactly fall on pixels. The radii in axes X, Y and T and the number of neighbouring points in the XY, XT and YT planes may differ, marked as R_X, R_Y and R_T, P_{XY}, P_{XT} and P_{YT}. Using micro-expression as an example, the XY plane contains spatial information that involves both the identification and appearance of the face, while the XT and YT planes contain information on pixel shifts in vertical and horizontal directions with time relative to the motion of the face. Three co-occurrence-based statistics obtained separately from three orthogonal planes are therefore used for the purposes of obtaining this information. As it is not appropriate for dynamic textures to set the radius in time to be equivalent

to radius in the space axis, we have different space and time radius parameters. The parameter values used in our calculations for R_X, R_Y and R_T are 1, 1 and 2 respectively and values for neighbouring points [P_{XY} P_{XT} P_{YT}] are [8 8 8].

The changes of the neighbouring points around the centre pixel are also taken into account with the time as the facial motion direction is uncertain. Taking into account the movement of the facial region, the face image is divided into a number of non-overlapping blocks. Each block's LBP-TOP histogram is computed to form one histogram and feature vector of size 3×2^P. The features were merged to reflect the appearance and motion of the series of the facial expression derived from each block size. Given that micro-expression has partial facial motion, the low-level features obtained from multiple local regions are crucial for recognition.

Local dense feature descriptors such as Convolutional Neural Networks (CNN) are deep neural networks utilised principally to categorise images (e.g. labelling images), grouping by similarity (photo data mining), and executing object recognition within scene images. The ideas of receptive field and weight sharing are applied by CNN. These ideas are causing the proliferation of data through the layers to be computed by convolution, and the quantity of trainable parameters is being diminished. A feature map is generated by convolving a signal with a filter map that contains the shared weights.

For the handcrafted feature, it is observed that the features from the local region are more crucial for the recognition of micro-expression owing to the brief length and partial facial motions. The handcrafted feature used is therefore the block-based LBP-TOP. The block-based LBP-TOP covers the facial motion on the spatial and temporal levels.

A CNN model is trained (fine-tuned) with eight databases separately for the deep local feature. Later, the pre-trained fine-tuned VGGFace CNN models are used to calculate the face's deep micro-expression feature. Five convolutional blocks and three fully connected layers form the network. Each of the pre-processed and selected facial region with micro-expression annotation is supplied as a CNN input during the training (fine-tuning) , and the weights of the FC7 fully connected layer are saved as the facial features (4096 dimensions).

4.3.2.2 Manifold Feature Integration

Integration seeks to accommodate and fuse into a common feature the advantages of handcrafted and deep features. The fusion at feature level occurs by merging deep CNN with handcrafted LBP-TOP features. The handcrafted LBP-TOP features of dimension (177×1) and the deep CNN features of dimension (4096×1) which are saved separately

into separate text files are fetched by the concatenation function to integrate LBP-TOP and CNN features and form a flattened feature vector of dimension (4273×1). The integration is just to attach the deep feature values to LBP-TOP.

Overall, with the integration of two existing methods, LBP-TOP and CNN, our framework is integrating spatial and temporal features extracted from LBP-TOP with dense spatial features from CNN. The concatenation of LBP spatial and CNN spatial features will give the model more comprehensive learning means about micro-expressions. The low-level temporal features from Three Orthogonal Planes (TOP) will provide the model about the transition or shift of the spatial features between consecutive frames that enables to process the micro-expression within a video. The 4273 dimensional vector is the shape of the training and testing samples.

4.3.2.3 Classification

The categorisation of expressions based on the selected features input is usually referred to as Classification. Classification typically involves training and testing, where the training phase prepares the classifier to distinguish micro-expression according to features and labels provided, while testing verifies the classifier's effectiveness. A number of supervised classification techniques including Support Vector Machine (SVM) [9, 81], Multiple Kernel Learning (MKL) [164], k-Nearest Neighbour (k-NN) [64], Random Forest (RF) [164], and Softmax in CNN [159] were applied to recognise micro-expression. The fully connected layer [192], the last layer in CNN, functions as classifier in the high-level approaches for micro-expression recognition. The fully connected layer is composed of conventional multilayer perceptron used in classifying high-level features. These deep features are derived from the convolutional and max-pooling layer in distinct classes with the initiation of Softmax. As this layer is initiated by Softmax, it is described as a Softmax classifier.

A classification system has two significant parts, one **score function**, which maps the raw data to class scores, and other **loss function** which assesses the agreement between the estimated results and labels of ground truth [84]. For the experiments, SVM loss (hinge loss) and Cross-entropy (used for Softmax classifier) are used. Suppose the $x_i \in R^D$ training dataset each labelled with y_i. Here $i = 1...N$ and $y_i \in 1...K$. This implies that there are N instances (all with D dimensionalities) of K discrete classes. A linear mapping function can be written as:

$$f(x_i, W, b) = Wx_i + b \tag{4.2}$$

This is estimated by flattening all the pixels of the image x_i in a single column vector $[D \times 1]$. The parameters of this function are W matrix (of size $[K \times D]$), which are frequently called weights, and the b vector (of size $[K \times 1]$) called a bias vector. Without interacting with the current information x_i, the parameters of the biased vector impact output results.

The features are combined into a single $[4273 \times 1]$ column during the experiment, W is $[6 \times 4273]$, $[5 \times 4273]$ and $[3 \times 4273]$ and b is $[6 \times 1]$, $[5 \times 1]$ and $[3 \times 1]$ (6 for CASME, CASME II, CASME+2; 5 for CAS(ME)^2, SAMM and 3 for SMIC-E). So 4273 numbers are supplied into the function (features extracted) and 6 or 5 or 3 number is given as the outcome (as the class scores).

The SVM loss is combined so that for each sample image the SVM "wants" the true class to achieve an approximately fixed margin Δ above the false classes. While hinge loss is quite popular, Deep Learning and Convolutional Neural Networks are more likely to use cross-entropy and Softmax classifiers. This is because Softmax classifier outputs the probabilities for each class label whereas the hinge loss returns the margin.

The new concatenated feature contains more abundant information. The classifiers used are, Softmax and Support Vector Machines (SVM), for micro-expression recognition. SVM uses four kernels-linear, polynomial, radial basis (RBF) and sigmoid-to test the efficiency of the CNN model. Traditionally, SVM and Softmax are analogous. Put differently, the Softmax classifier continuously improves its score: the right class always has elevated probabilities and the wrong classes are often less likely, with the loss constantly improving. The SVM is nevertheless persuaded when the margins are filtered and the precise results are not regulated beyond this restriction. Naturally, this can be considered a feature.

Hence, for validating the comparability and effectiveness, both the classifiers are implemented in the proposed model.

4.4 Experimental setup and Outcomes

4.4.1 Databases

Adequate quantity of data for the training of CNN without any underfitting or overfitting is a key element for working with the deep learning networks. In this case, micro-expressions do not have many samples that allow a deep network to be trained from scratch.

Table 4.2: Experimental Micro-expression databases and emotion categories

Database		Micro-expression classes
CASME		6 (Disgust, Fear, Happy, Neutral, Sad, Surprise)
CASME II		
CASME+2		
CAS(ME)^2		5 (Disgust, Fear, Happy, Sad, Surprise)
SAMM		
SMIC	HS	3 (Negative, Positive, Surprise)
	NIR	
	VIS	

Table 4.2 gives details about the number of emotion classes used in analysis. Six emotion classes from CASME and CASME II are selected, which are common to both the datasets and also for the combined dataset. The micro-expression class Neutral is added to CASME, CASME II and CASME+2 class since there are a few unlabelled videos without expressions. Similar micro-expression categories from CAS(ME)^2 and SAMM datasets are also selected except for Neutral as these datasets do not have any unlabelled videos. As SMIC dataset has three micro-expression types, all the classes are considered for the experiments.

The Experiment I uses CASME, CASME II and CASME+2 databases with six emotion classes whereas the Experiment II uses all the databases mentioned above, along with the combined CASME+2 database to assess the potency of the proposed micro-expression recognition approach.

4.4.2 Experimental Setup

4.4.2.1 Experiment I

The micro-expression recognition model verifies the effectiveness by conducting experiments on the CASME, CASME II and CASME+2 datasets. All the images in the databases are pre-processed and flipped vertically to increase the number of samples. The new synthetic database is then divided into three groups as Training, Testing and Validation. Each image has been categorised as: 0 = Disgust, 1 = Fear, 2 = Happiness, 3 = Neutral, 4 = Sadness, and 5 = Surprise.

We implemented the deep convolutional neural networks based on the Caffe [82] (a fast open framework for deep learning and computer vision) and took 10-12 hours to train this network. The models were trained for 100,000 iterations on CASME and 41,000 iterations on CASME II and CASME+2. The initial learning rate is changed from 0.001 to 0.0001 when the training iteration reaches 10,000. The layer parameters of

the network are modified at each round of iterations in model training on the basis of the loss. We obtain a trained model when the maximum training iterations are reached, which is basically the parameter of all the filters. The model is then saved in order to use it to predict a micro-expression from images. The input is given from the Validation sets which are the raw face images collected from the original databases. For each experiment, a corresponding Validation set is used depending on the Training database.

4.4.2.2 Experiment II

The aim is to obtain an LBP-TOP flattened feature histogram of the dimensionality of 177. The VGGFace pre-trained model is fine-tuned using all the samples in the train set from all the eight micro-expression databases individually. The fine-tuned CNN extracts a feature vector from Fully Connected (FC7) layer of dimension 4096. A concatenation operation on LBP-TOP (177×1) and CNN (4096×1) features is performed to flatten the two separate feature vectors into one feature vector of 4273 dimensionality. CNN model is also built using Caffe deep learning toolbox.

Data augmentation technique is used to double the amount of sample information in all training sets. Each of the eight databases is split into a Training set (80%), and Testing set (20%) for the assessment of the suggested framework. The hyper-parameter "*max_iter*" value, which defines the maximum number of iterations to be performed, differs considering the number of training set samples. Since there are fewer training samples in some databases, the number of iterations is also reduced. The initial learning rate '*base_lr*' hyper-parameter value is 0.001 for training and ranges from 0.001 to 0.00001 during testing with Softmax classifier.

The models for both the experiments are trained and tested on NVIDIA(R) GeForce(R) GTX 1060 with 6GB GDDR5. This hardware configuration boosted the fine-tuning process by reducing the time taken for fine-tuning as the code is run in GPU mode. The databases with a sizable number of training samples (CASME+2 and SMIC-HS) took around 96 hours due to the higher number of iterations it had to run. Whereas the other databases (CASME, CASME II, CAS(ME)^2, SMIC-NIR, SMIC-VIS) have less training samples which took less than 48 hours to finish the fine-tuning.

4.4.3 Evaluation Results

The intension of this section is to compare proposed methodology with alternative progressive algorithms in facial micro-expression recognition. The comparison is done

on the basis of the databases used in the experiments. The experimental setup and conditions for the existing methods are as mentioned in the respective research.

4.4.3.1 Experiment I

The Experiment I recognition accuracy results for the three databases used are summarised in Table 4.3. From the table, we can observe that the recognition accuracy improves as the number of training samples increases.

Table 4.4 lists the recognition accuracy of using our method and of the state-of-the-art methods in CASME dataset and CASME II dataset respectively. Most of the existing methods are video based, which more or less take advantage of the temporal information from the video. Our method is image based, which applies CNN on image frames extracted from videos. The tables testify that proposed CNN method exhibits satisfying micro-expression recognition accuracy. These results also demonstrate that image based micro-expression recognition delivers identical results as video based approaches.

Due to slightly greater number of samples in CASME II as compared to CASME dataset, the researchers in [88] opted to demonstrate deep learning results on video clips from CASME II dataset. In our research, we have applied data augmentation technique to increase the number of samples. Therefore, we could use both CASME and CASME II datasets to showcase the effectiveness of proposed CNN method.

The studies [60, 88, 127, 210, 215] have considered the temporal factor from the video for recognition of micro-expressions which have contributed an additional feature in the calculations. In case of our method, we tried to eliminate the temporal factor and simply

Table 4.3: Micro-expression recognition accuracy with different databases

Database	CASME	CASME II	CASME+2
Accuracy	74.25%	75.57%	78.02%

Table 4.4: Accuracy correlation with existing advanced approaches on public databases

Database	Method	Accuracy
CASME	LBP-TOP+ELM [60]	73.82%
	MDMO+SVM [127]	68.86%
	LBP-TOP+SVM [210]	61.85%
	Experiment I	**74.25%**
CASME II	LBP-TOP+SVM [215]	75.3%
	MDMO+SVM [127]	67.30%
	CNN+LSTM [88]	60.98%
	Experiment I	**75.57%**

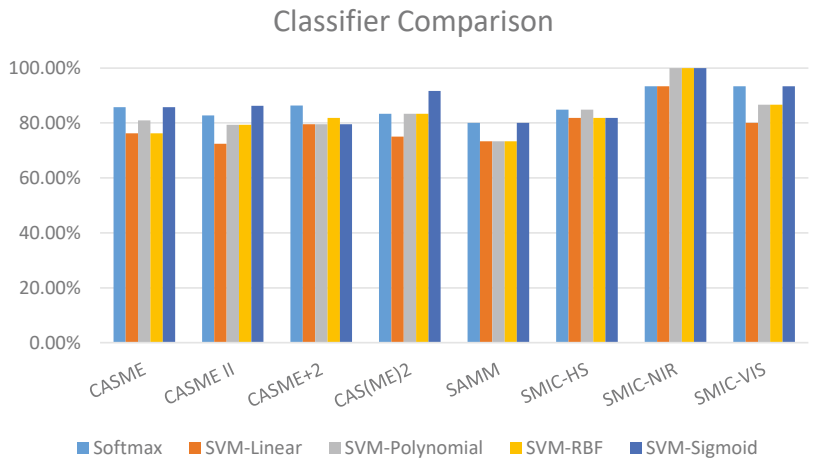


Figure 4.4: Accuracy comparison graph of Softmax and SVM kernels

exhibit the image based micro-expression recognition approach.

4.4.3.2 Experiment II

For this experiment we use all the databases to fine-tune the VGGFace model and train two classifiers. Observations in Table 4.5 shows that both the classifiers perform equally well, but sometimes SVM takes over the Softmax classifier by 4-7%.

Figure 4.7 presents a graphical demonstration of Table 4.5 to visualise the recognition accuracy results of both Softmax and SVM classifiers. The attempt to implement two classifiers, Softmax and SVM, is to verify the accuracy of our approach and to identify a more suitable classifier for our method.

Table 4.6 testifies that the suggested manifold feature integrating method outper-

Table 4.5: Evaluation results of the Manifold Feature Integration

Database		Recognition Accuracy (%)				
		Softmax	SVM			
			Linear	Polynomial	RBF	Sigmoid
CASME		**85.71%**	76.19%	80.95%	76.19%	**85.71%**
CASME II		82.75%	72.41%	79.31%	79.31%	**86.21%**
CASME+2		**86.36%**	79.55%	79.55%	81.82%	79.55%
CAS(ME)^2		83.33%	75.00%	83.33%	83.33%	**91.67%**
SAMM		**80.00%**	73.33%	73.33%	73.33%	**80.00%**
SMIC	HS	**84.84%**	81.82%	**84.84%**	81.82%	81.82%
	NIR	93.33%	93.33%	**100%**	**100%**	**100%**
	VIS	**93.33%**	80.00%	86.67%	86.67%	**93.33%**

forms cutting-edge techniques. The experimental observations of the proposed approach applied on the seven publicly available databases are considerably better than most modern techniques. Notably, the efficiency of the proposed approach is not only goes beyond the present deep learning techniques but is outstanding to handcraft based techniques. These outcomes reflect the validity of the suggested manifold feature learning and integration micro-expression recognition approach.

Table 4.6: Comparison of recognition accuracy on eight databases compared with our approach

Database	Methods	Recognition Accuracy (%)
CASME	LBP-TOP+SVM [210]	61.9%
	MDMO+SVM [127]	68.9%
	FHOFO+ LSVM [64]	71.6%
	LBP-TOP+CNN+Softmax	**85.7%**
	LBP-TOP+CNN+SVM (Sigmoid)	**85.7%**
CASME II	CNN+SVM [159]	47.3%
	MMFL+SVM [65]	59.8%
	CNN+LSTM [88]	61.0%
	FHOFO+LSVM [64]	64.0%
	Joint feature+Multi-task [71]	66.2%
	HIGO+LSVM [111]	67.2%
	MDMO+SVM [127]	67.4%
	Firefly+ISO-FLANN [1]	68.7%
	EVM+LBP-TOP+SVM [215]	75.3%
	SME+SVM [147]	85.0%
	LBP-TOP+CNN+Softmax	**82.8%**
	LBP-TOP+CNN+SVM (Sigmoid)	**86.2%**
CASME I/II OR CASME+2	DSTCNN+SVM [161]	66.7%
	LBP-TOP+CNN+Softmax	**86.4%**
	LBP-TOP+CNN+SVM (RBF)	**86.2%**
CAS(ME)^2	DRMF+Bi-WOOF+SVM [122]	59.3%
	MicroExpSTCNN [173]	87.8%
	LBP-TOP+CNN+Softmax	**83.3%**
	LBP-TOP+CNN+SVM (Sigmoid)	**91.7%**

Continued on next page

Table 4.6 – continued from previous page

Database	Methods	Recognition Accuracy (%)
SAMM	HOG3D+SMO-SVM [28]	63.9%
	LBP-TOP+SMO-SVM [28]	81.9%
	LBP-TOP+CNN+Softmax	**80.0%**
	LBP-TOP+CNN+SVM (Sigmoid)	**80.0%**
SMIC-HS	CNN+SVM [159]	53.6%
	SME+SVM [147]	58.7%
	AC-GAN+Bi-WOOF [119]	61.8%
	SAGAN+Bi-WOOF [119]	62.2%
	DRMF+Bi-WOOF+SVM [122]	62.2%
	Joint feature+Multi-task [71]	65.1%
	MicroExpSTCNN [173]	68.8%
	LBP-TOP+CNN+Softmax	**84.8%**
	LBP-TOP+CNN+SVM (Polynomial)	**84.8%**
SMIC-NIR	SME+SVM [147]	44.1%
	HIGO+LSVM [111]	67.6%
	LBP-TOP+CNN+Softmax	**93.3%**
	LBP-TOP+CNN+SVM	**100.0%**
SMIC-VIS	SME+SVM [147]	45.9%
	CNN+SVM [159]	56.3%
	HIGO+LSVM [111]	81.7%
	LBP-TOP+CNN+Softmax	**93.3%**
	LBP-TOP+CNN+SVM (Sigmoid)	**93.3%**

4.4.4 Performance Metrics

The evaluation of the experimental results of proposed approach, apart from accuracy, other performance evaluation metrics such as a confusion matrix, F1 score, Precision and Recall are also considered. The confusion matrix is considered one amongst the pre-eminent spontaneous and most comfortable parameter for identifying the true positives of the model. It is generally used for classifying the output in two or more categories. We plotted confusion matrix of each database for Experiment I and II based on the probability calculated by the classifiers used for each experiment, respectively, placed side-by-side. Figures 4.5 (a)-(c) shows the confusion matrix for Experiment I databases and Figures 4.6 (a-p) represents the confusion matrix for Experiment II databases.

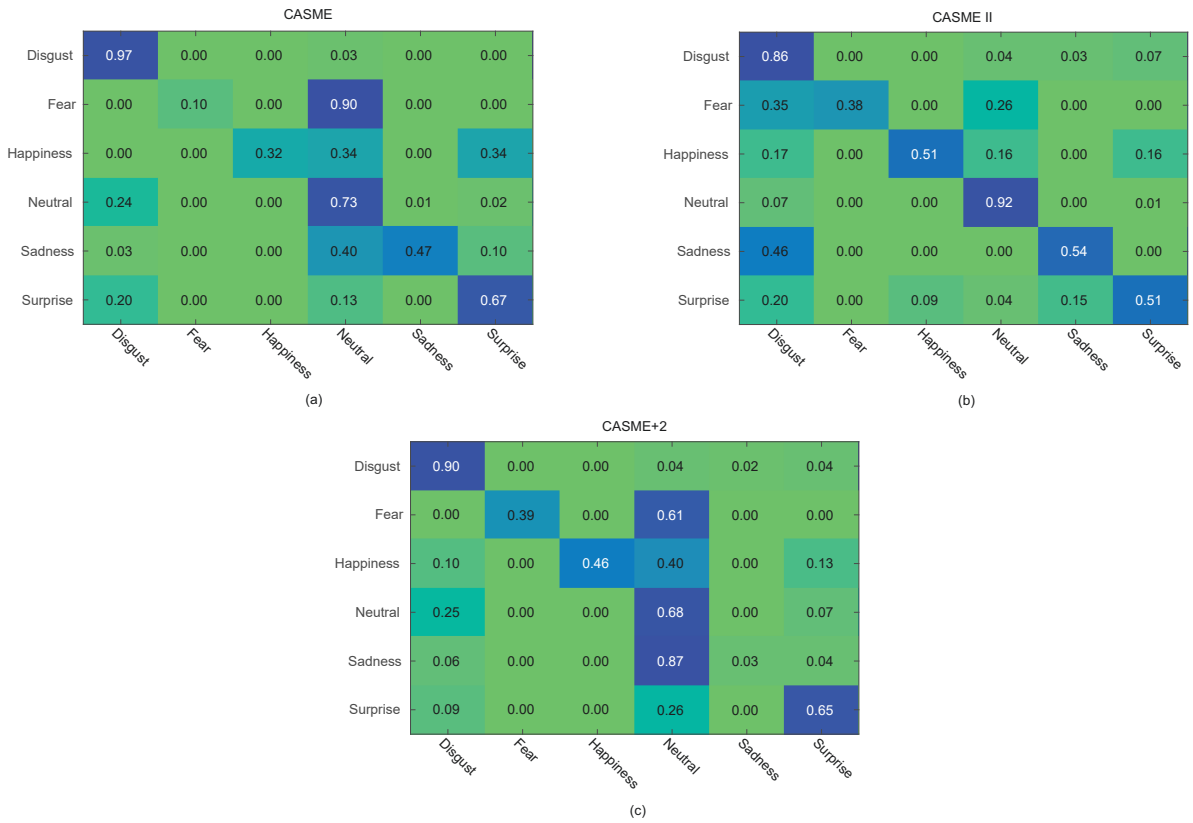


Figure 4.5: (a)-(c) Confusion matrices based on probabilities predicted by Softmax classifier

Accuracy is the number of right predictions produced by the model, over all sorts of predictions in classification problems. Tables 4.3 and Table 4.5 gives the accuracy numbers for projections from Softmax as well as Softmax and SVM (all kernels) classifiers for Experiment I and Experiment II respectively.

Along with accuracy, precision, recall and F1 score are reliable performance evaluation metrics.

Precision is a degree that tells us what extent of positive identifications are actually correct. It is also called positive predictive values.

Recall manages to determine the portion of actual positives that were correctly identified. In other words, recall (also known as sensitivity) is the portion of significant occurrences that have been recovered over the entire extent of relevant instances.

F1 score measures a test's accuracy. The harmonic mean between accuracy and recall is the F1 score. The F1 scoring range is [0,1] [144].

An understanding and measure of relevance are the basis for precision and recall; and F1 score shows how accurate and robust our classification is (Figure 4.7 (a-c)).

For Experiment II we took a step further and performed additional analysis and report precision, recall and F1 scores. Accuracy is an extremely important measure, but only with symmetrical datasets (false negative and false positive counts are close), and false negatives and false positives also carry the same costs. If false positive and false negative costs differ, F1 will be a saviour here. F1 is preferable if the class distribution is uneven, as-like in our case where the micro-expression classes have varying number of samples. Precision shows how certain our model is of the true positives while recall is how certain the model is that no positives are missing.

We chose to calculate precision and recall (Figure 4.7 (a-c)) as we wanted to be more sure with our true positives to eliminate any false negatives and also F1 score because we wanted to cover all true negatives to avoid false alarms and did not want any false positives. Results from all these performance measures have reinforced the results of our classification and convinced us that our proposed system provides significant results in comparison with state-of-the-art approaches.

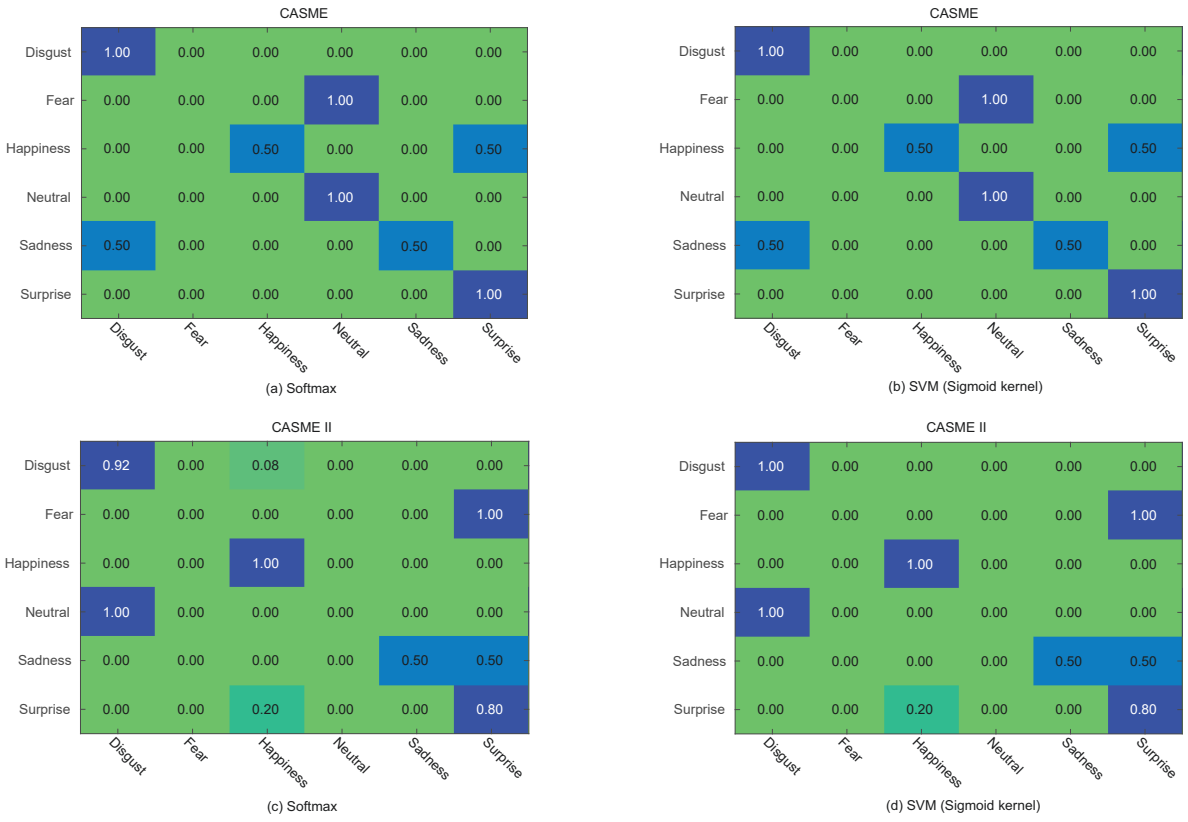
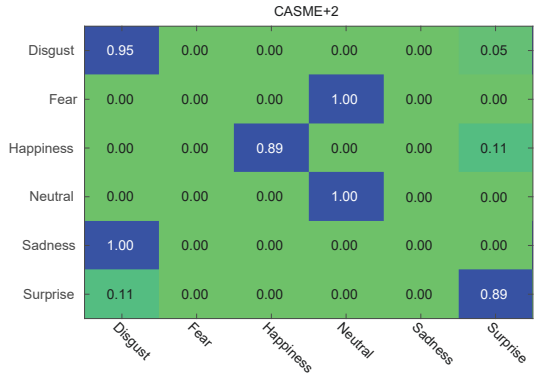
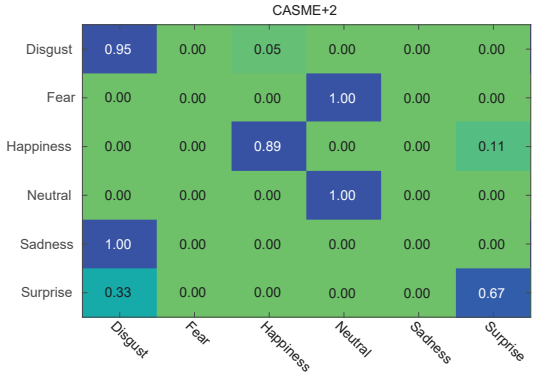


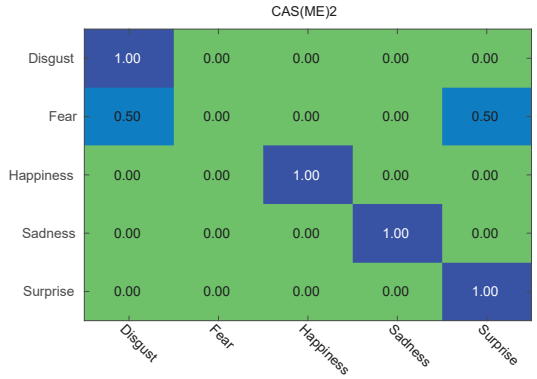
Figure 4.6: (a)-(d) Confusion matrices based on probabilities predicted by Softmax and SVM classifier



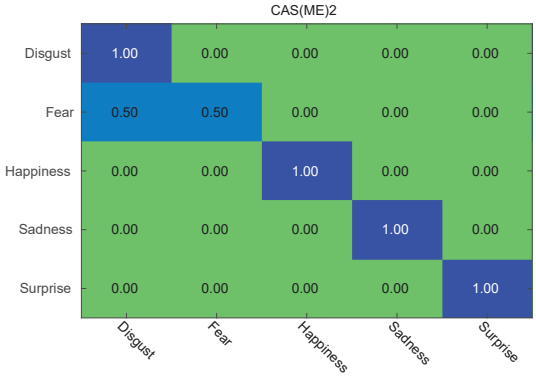
(e) Softmax



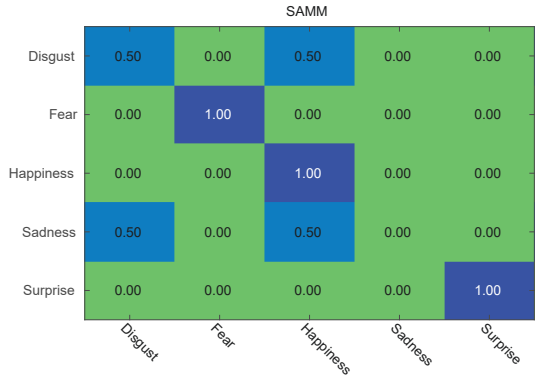
(f) SVM (RBF kernel)



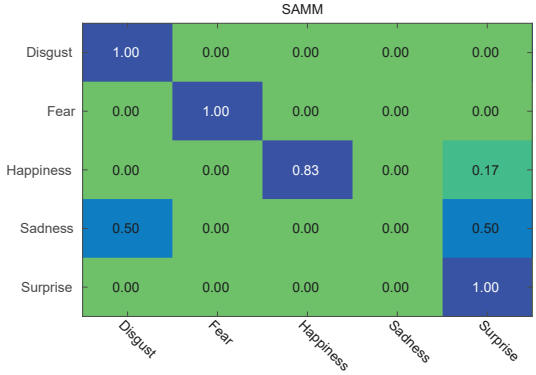
(g) Softmax



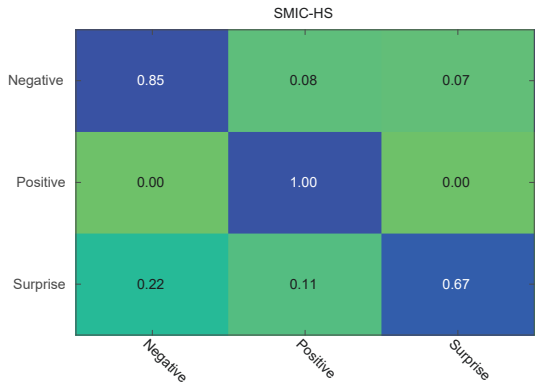
(h) SVM (Sigmoid kernel)



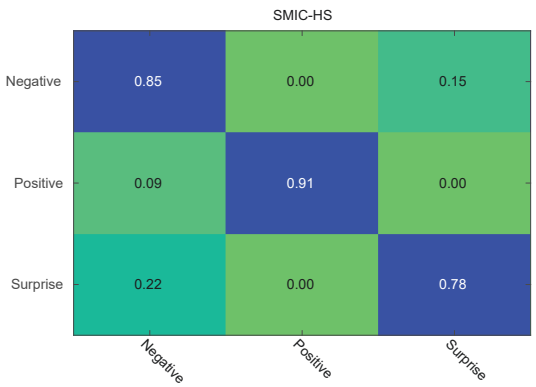
(i) Softmax



(j) SVM (Sigmoid kernel)



(k) Softmax



(l) SVM (Polynomial kernel)

Figure 4.6: (e)-(l) Confusion matrices based on probabilities predicted by Softmax and SVM classifier

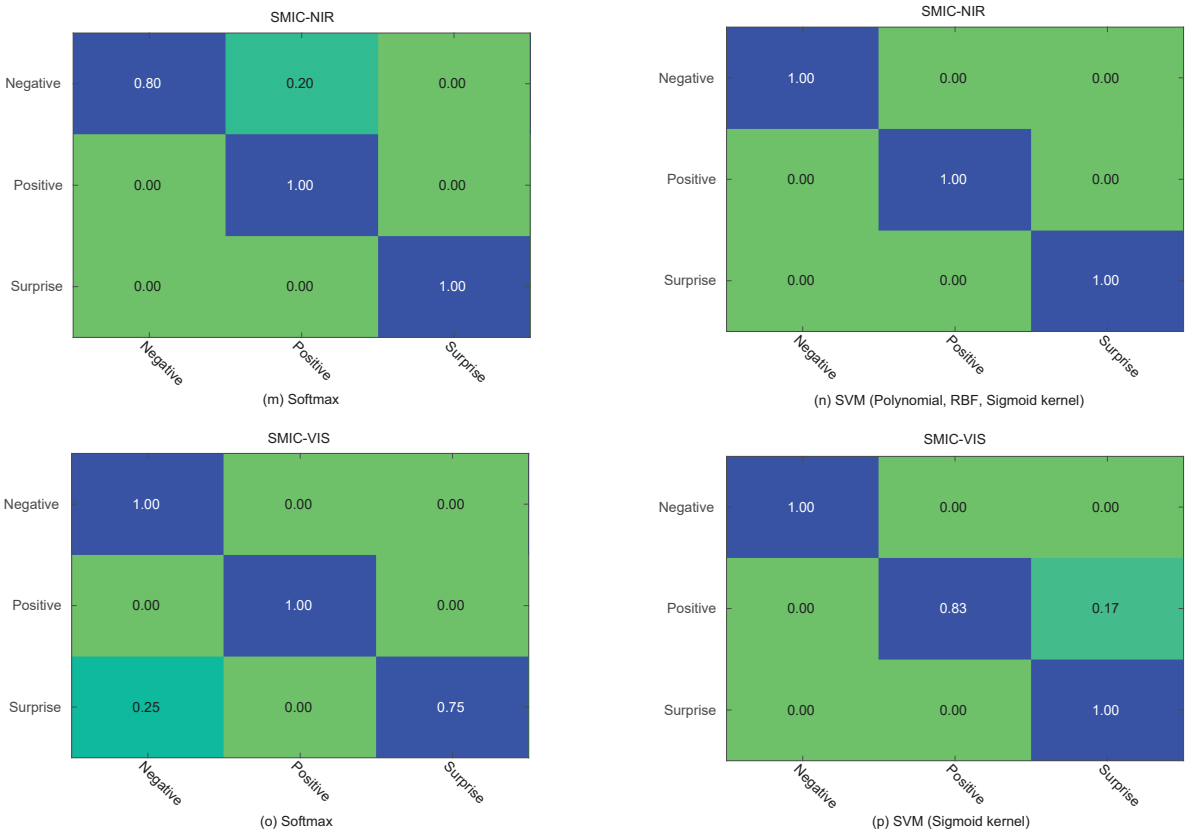
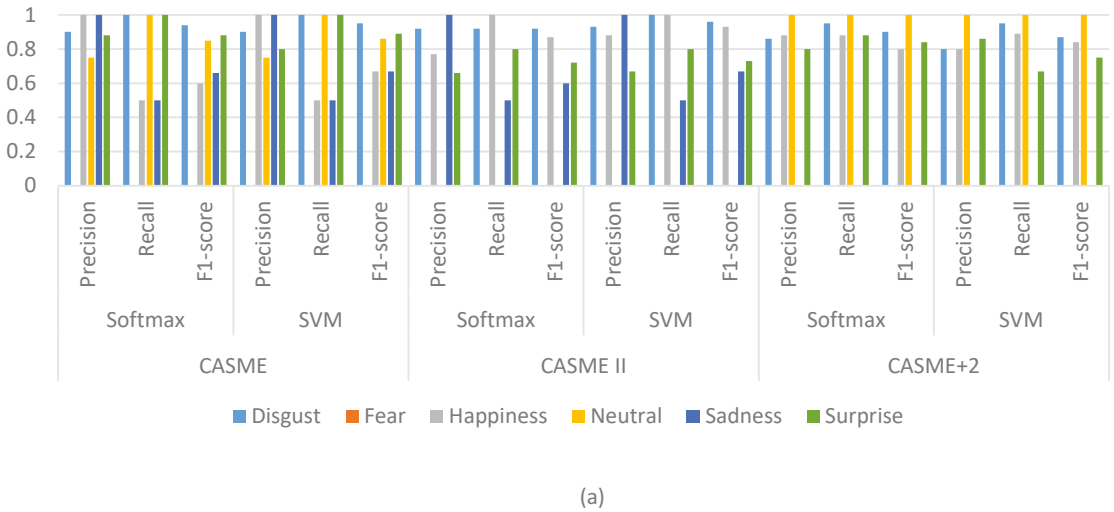
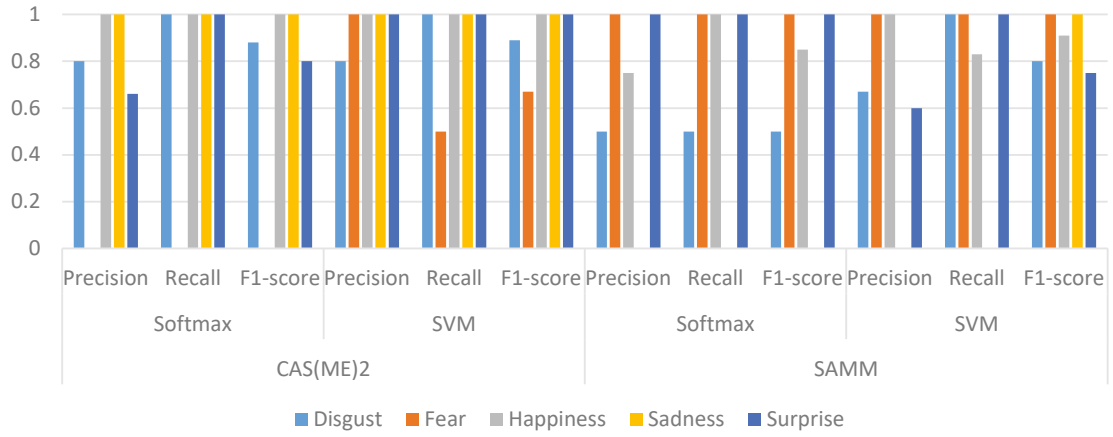


Figure 4.6: (m)-(p) Confusion matrices based on probabilities predicted by Softmax and SVM classifier

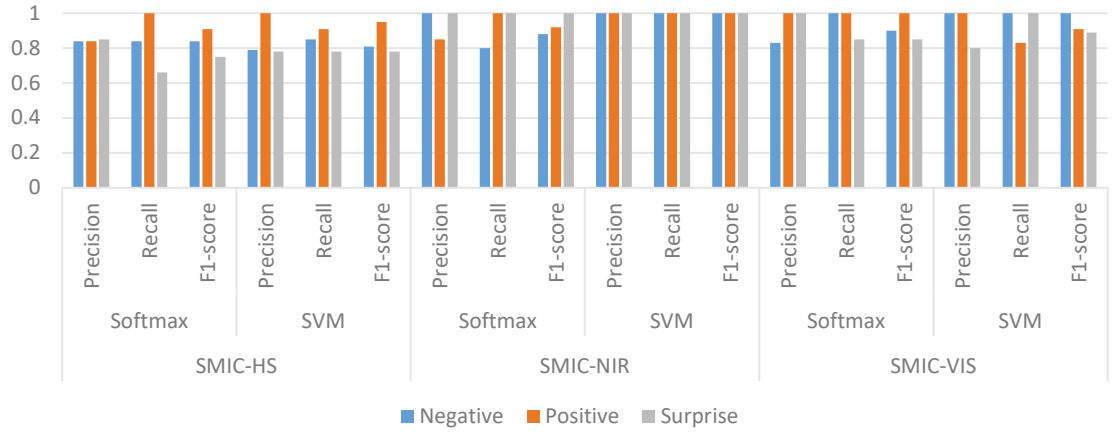


(a)

Figure 4.7: (a) Figure of Merit for experimental datasets



(b)



(c)

Figure 4.7: (b)-(c) Figures of Merit for experimental datasets

4.4.5 Ablative Analysis

4.4.5.1 Comparison with original features

Manifold Feature Integration model (Experiment II) is a model that utilises multiple features for calculating a feature vector used to train and test the classifiers. In addition to comparing the proposed framework with existing methods, we performed classification checks on LBP-TOP feature using 8×8 block partitions, CNN feature and joint feature individually. Table 4.7 demonstrates that in all repositories, our joint functionality delivers consistently better performance than the original features LBP-TOP and CNN. The explanation for the original features not being individually effective than the joint features is the number and type of features extracted for classifiers' learning. LBP-

Table 4.7: Recognition accuracy of original features

Database	Softmax			SVM		
	CNN	LBP-TOP	CNN+ LBP-TOP	CNN	LBP-TOP	CNN+ LBP-TOP
CASME	81.0%	42.8%	85.7%	71.4%	47.6%	85.7%
CASME II	75.9%	44.8%	82.7%	75.8%	51.7%	86.2%
CASME+2	77.3%	45.7%	86.4%	75.0%	48.9%	81.8%
CAS(ME)^2	75.0%	25.0%	83.3%	75.0%	33.3%	91.7%
SAMM	66.7%	40.0%	80.0%	66.7%	41.6%	80.0%
SMIC-HS	78.8%	48.4%	84.8%	75.7%	69.6%	81.8%
SMIC-NIR	80.0%	33.3%	93.3%	73.3%	53.3%	100.0%
SMIC-VIS	73.3%	53.3%	93.3%	80.0%	40.0%	93.3%

TOP has 59 containers and three orthogonal planes with a 177-dimensional feature vector. Given a limited facial region with a micro-expression, extraction of LBP-TOP features alone may be of little significance. In order to have a thorough feature vector we used CNN, which shows better results than LBP-TOP but is a spatial extractor that overlooks the time factor. The analysis shows that the deep features can strengthen the discriminating capability of the low-level handcrafted features leading to a better recognition efficiency.

4.4.5.2 Significance of data augmentation

As discussed under Section 4.2.2, data augmentation is an alternative means of reducing model overfitting, where we only increase the number of training data using information in our training data. An increasingly common and recognised technique for augmentation is the combination of affinity transformations, such as shifting, zooming in/out, rotating, flipping, distortion, or shading with hue. In this approach, augmented data is generated before the classifier training. The classification of micro-expression is impeded by lack of data. Table 4.8 supports the usefulness of the data augmentation methodology by way of affine transformation to boost number of training samples and recognition accuracy.

4.5 Discussions

In addition to width, height and time, Wang et al. [210] use the color as a tensor of fourth-order for deriving LBP-TOP features and SVM as classifier. The findings indicate that Tensor Independent Color Space (TICS) offers useful information than

Table 4.8: Comparing proposed model with and without data augmentation (DA)

Database	Softmax		SVM	
	DA	w/o DA	DA	w/o DA
CASME	85.7%	71.4%	85.7%	66.6%
CASME II	82.7%	68.9%	86.2%	58.6%
CASME+2	86.4%	69.3%	81.8%	57.1%
CAS(ME)^2	83.3%	58.1%	91.7%	50.0%
SAMM	80.0%	53.3%	80.0%	60.0%
SMIC-HS	84.8%	63.6%	81.8%	69.6%
SMIC-NIR	93.3%	73.3%	100.0%	60.0%
SMIC-VIS	93.3%	66.6%	93.3%	73.3%

RGB and grayscale. Certain studies like [64, 77] indicate that LBP-TOP alone may not be appropriate to produce reliable performance. Table 4.7 shows that the accuracy is low for LBP-TOP.

In the study by Patel et al. [159], they attempted to use deep features passed from pre-trained ImageNet models. The researchers found that it is not feasible to fine-tune the network with micro-expression datasets and chosen a feature selection approach. Also, several works such as Kim et al. [88], Peng et al. [161] investigated the utilised deep neural networks through encoding space and time features learned from comparatively shallow network architecture than those of the ImageNet challenge. The model we have used is VGGFace, which is pre-trained on a massive face image repository, contrary to the pre-trained models used in other research approaches. Using VGGFace lets us take the lead because the CNN model already knows facial characteristics, and then we fine-tune VGGFace on the increased micro-expression data.

4.5.1 Difficulties with certain expressions and databases

Confusion matrix observations demonstrate that the model projections for some classes of micro-expression tend to differ according to the training scheme and adaptive architecture scoring lower accuracy, as these classes turns out to be “harder” to train. In case of Image-based model, the classes such as “Fear”, “Happiness”, and “Sadness” appear to be the classes that shows inconsistency in the classification results. Whereas, in Manifold Feature Integration model, the classes for instance “Fear” and “Sadness” from CASME, CASME II and CASME+2 datasets, “Fear” from CAS(ME)^2, “Disgust” and “Sadness” from SAMM, and “Surprise” from SMIC exhibits similar behaviour.

The observed inconsistency might be due to two reasons as discussed. First, the

mentioned micro-expression categories contains very scanty training data samples as related to other categories in the respective datasets, making it difficult for the network to train for recognising them. Second, these micro-expressions can be extremely subtle and even the specialists find it difficult to comply with their true annotations [221].

We suspect that the inherent difficulty in assigning labels to some of the samples may have caused them to be "mislabelled", thereby affecting the models that were trained on them. We would also highlight here that some of the models were unable to predict a sufficient number of samples for label "Fear" correctly. The reason for this could be an imbalance in the training datasets. The imbalance in the number of training samples for each class of micro-expressions most likely caused our models to overfit the micro-expressions with more samples (e.g. "Disgust") at the expense of this class.

We attempted to keep the Training (80%) and Testing (20%) set distribution as person-independent in most scenarios possible. However, for instance, CAS(ME)^2, micro-expression class "Sadness" has only one sample, or CASME where only one person has "Neutral" sample videos. In such situations, the same person can be in Train as well as Test set, as can be seen from the confusion matrix. Moreover, not all subjects has all categories of micro-expression in any specified database. Some subjects have either two or sometimes just one micro-expression class in the SMIC database. As a consequence, person-independent validation becomes challenging and unjust.

4.6 Summary

This chapter presents and examines two modern methods for extraction of features for the implementation of micro-expression recognition.

The primary contribution for the Experiment I lies in fine-tuning a CNN model using small datasets to recognise micro-expressions from images. It is suggested that if we were to exploit deep neural networks such as CNN for facial micro-expression recognition to achieve the significant gains seen in other domains, then having bigger datasets is crucial. This is where we implanted the idea of combining the two databases CASME and CASME II to form a larger database. We demonstrated through our experiments that image-based micro-expression recognition could also yield acceptable accuracy as compared to image-base facial expression recognition.

A thorough study and verification of the combination between handcrafted and deep features is carried out in Experiment II to enhance the accuracy of recognition for micro-expression. This research proves to be the first work to train and test the proposed

manifold feature integration model on all seven publicly available datasets and one combined dataset and achieve acceptable micro-expression recognition accuracy. It is observed that proposed micro-expression recognition approach achieves appreciable results in comparison to the state-of-the-art techniques and networks.

CHAPTER 5

GEME: DUAL-STREAM MULTI-TASK GENDER-BASED MICRO-EXPRESSION RECOGNITION

Recognition of micro-expressions remains a topic of concern considering its brief span and low intensity. This issue is addressed through convolutional neural networks (CNNs) by developing multi-task learning (MTL) method to effectively leverage a side task: gender detection. A dual-stream multi-task framework called GEME is introduced that recognises micro-expressions by incorporating unique gender characteristics and subsequently improves the micro-expression recognition accuracy. This research aims to examine how gender differences influence the way micro-expressions are displayed. The current study proves that selecting relevant features of micro-expressions distinctive to the gender and added to the micro-expression features improves the micro-expression recognition accuracy. This network learns gender-specific features and micro-expression features and adds them together to learn the combination of shared and task-specific representations. A multi-class focal loss is used to mitigate the class imbalance issue by down-weighing the easy samples and concentrate more on misclassified samples. The Class-Balanced (CB) focal loss is also implemented for a better class balancing during Leave-One-Subject-Out (LOSO) validations where CB loss re-balances and re-weights the loss. The experimental results on four widely used databases demonstrate the improved performance of the proposed network and achieve comparable results with the state-of-the-art methods.

5.1 Introduction

Throughout recent decades, micro-expressions have gained growing publicity. A facial micro-expression is a stifled movement of the face that only persists quite briefly (i.e. 40 milliseconds) [42]. They are either a consequence of deliberate or involuntary manipulation of expressions. It is rather challenging for humans to identify a micro-expression with naked eyes. Micro-expressions may be crucial, because they relate closely to genuine emotions, to demonstrate how people seek to dissimulate emotions, particularly in contexts of great concern, like health treatments, national security and criminal investigations [44].

Evidence suggests that it is unlikely for human facial muscles to sufficiently expand in 0.5 secs to render a discernible facial expression [139, 230]. Thus, it is not straightforward for humans to reliably identify and classify micro-expressions with little or no professional knowledge. Ekman [41] has built a Micro-Expression Training Tool (METT), where people involved can acquire professional skills of seven elusive facial expressions to detect and recognise micro-expressions. Despite undertaking METT's training [51], the recognition efficiency remained deficient. Furthermore, human understanding of emotions is easily influenced by the human's perception, rendering the outcomes vary for different subjects on different occasions.

Analysing human activities is a crucial and intriguing open issue in the field of video analysis. Lately, a significant number of reviews are conducted to understand human behaviour on some specific aspects of the larger problems namely, pedestrian detection [103], human detection for driver-assistance systems [56]. Reviews on action recognition approaches [233] extending to human-object interactions and group activities [3] are also studied under human behaviour analysis. Similar to the visible actions, the detection of the emotional state is necessary to understand the human behaviour which proves to be a critical aspect in various applications such as social, medical and behavioural science. The non-verbal cues displayed in the conscious awareness like facial expressions [10, 37, 49, 93, 109, 146] and valence and arousal [91, 146, 237] of an expression depicts the mental state of a person and can be applied in evaluating the emotional impairment in neuropsychiatric disorders. The studies conducted by Thuseethan et al. uses action unit (AU) intensities to predict the intensity of the basic facial emotions [198], revealing real emotions of a person by detecting micro-expression intensity changes [197] and later extending the work to estimate the continuous pain intensity [195]. The expressions related to facial muscle movements can be distinguished

from momentary facial appearances using the Facial Action Coding System [46]. Some methodologies presented in [128, 149, 174] have been proposed to detect the occurrence of the action units on facial images.

Past research has also analysed supervised and unsupervised pre-training to enhance generalisation, some suggested turning the initial one-task problem into a new Multi-Task Learning (MTL) issue. "MTL enhances generalisation by consolidating the domain-specific information in the related task training signals" [19]. Over the years, multi-task learning has demonstrated its importance in several areas [92, 174]. Current MTL approaches are focused on an architecture template with shared CNN layers at the bottom and top layers for specific tasks [90, 145, 171, 172]. Nonetheless, it is often heuristic to choose the sharing design as it is hard to determine which layers to share. Multi-task learning commonly used in tandem with ConvNets in computer vision to co-model related tasks, such as pose prediction and action recognition [124], face detection and facial landmark detection [171], auxiliary tasks in detection [115], relevant classes for image classification [171] and so on. These approaches typically share certain features (ConvNet layers) among tasks, including some task-specific features.

Lately, there have been many efforts to establish computer vision approaches for recognition of micro-expressions. Micro-expression study has motivated some researchers to assess individual difference in Emotion Recognition Ability (ERA) that differs through gender, ethnicity, culture and psychiatric status [140]. Thereby, the effect of various ethnic groups, gender [194] and cultural communities must be addressed in order to achieve a complex and reliable facial micro-expression recognition system.

There are many existing neuropsychological studies [16, 33, 63, 94, 96, 142, 204] that give pieces of evidence of the gender difference in facial behaviours. The literature [16, 17, 33, 63, 94, 96, 204] concludes that women are more expressive and good conveyor of non-verbal communication than men. The evidence in [48, 66] suggests that women try more to conceal the expression of anger. Overall, from these researches, it is clear that women express facial muscle movements very often as compared to men, and mainly display more positive valence actions. The study conducted in [142] is by far the most extensive study of gender differences in facial behaviour which puts forward the observations that women express anger related actions less and fear and sadness related actions more than men. Moreover, the research performed by Hu et al. [70] using the computer vision approach illustrates that gender difference also affects micro-expression recognition performance.

Given the amount of research in Psychology and growing interest in Computer Vision

regarding influence of gender on expressiveness of emotions and emotion recognition, respectively, the aforementioned research is motivated to back the psychological study with the technological observations. Hence, the attention is concentrated on using the unique gender features from the input data to achieve higher micro-expression classification accuracy. It is evident that people of different gender typically have different ways to express micro-expression.

Thus, the contributions of this work are as listed [6].

* This research evaluates to demonstrate that the influence of including gender features with micro-expression features is quite significant and that gender affects the way a person expresses emotions. The dual stream model uses one stream to learn gender features and infuse these specific gender features with the second stream micro-expression features.

* A thorough search of the relevant literature yielded that this study is the first to use deep learning techniques for proposing and incorporating gender features by combining it with the micro-expression features for recognition. In this predictive network, multi-task learning approach for processing facial images which breaks down the high dimensional "gender" features and low dimensional "micro-expression" features in order to determine the impact of gender on human emotions.

* Comparable micro-expression recognition accuracy is achieved on databases used for experiments. The results are compared with state-of-the-art methods to illustrate the robustness and efficiency of the proposed multi-task learning GEME model.

* Experiments by using only single micro-expression recognition stream are also performed to indicate the impact of the gender features on the process of micro-expression recognition. It is observed that the micro-expression recognition performance using single stream is reduced as compared to the dual stream GEME model. The results are also comparable with state-of-the-art methods, and it is observed that the single emotion-stream of the proposed network also performs slightly better than the existing studies.

The remainder of the chapter has been distributed in various sections as mentioned below. Section 5.2 addresses several existing works on micro-expression recognition

[6]Xuan Nie, Madhumita A. Takalkar, Mengyang Duan, Haimin Zhang, Min Xu, *GEME: dual-stream multi-task GEnder-based Micro-Expression recognition*, Neurocomputing. (*Under Review*)

and multi-task learning. Section 5.3 outlines the proposed framework for recognising micro-expression using gender properties. Next, in Section 5.4, the experimental setup and outcomes for the GEME approach are analysed. Eventually, Section 5.5 draws conclusions.

5.2 Related Research

Micro-Expression Recognition (MER) has become a prominent research topic which inspired the development of comprehensively successful approaches. There are two key elements involved in traditional MER approaches: the extraction of facial features that seek to extract relevant information for defining micro-expression from facial videos, and the classification of micro-expression by building a classification system based on the features collected in the first step of MER method. Facial feature extraction has drawn growing research focus. Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) and its variants are some of the extensively used feature extraction methods practised in video-based MER as well as in specific computer vision applications [152, 243]. Researchers in studies [79, 111] further notes that the video's time dynamics can boost MER efficiency as they can accurately reflect the movement through a series of micro-expression frames. In some experiments, to obtain motion-based spatio-temporal information from micro-expressions, optical flow (OF) methods are used [64]. Various classifiers used for micro-expression classification such as support vector machine (SVM), relaxed K-SVD, and group sparse learning (GSL) are machine learning-based. Zong et al. [248], notably, suggested a kernelised GSL to promote acquiring a set of weights from hierarchical spatio-temporal descriptors, which might help pick essential segments from different facial segments. Zheng et al. [245] introduced a relaxed K-SVD which learned to discern various micro-expressions with a sparse dictionary by reducing the sparse coefficient variance.

Research teams also reviewed deep learning approaches to deal with the MER issue in recent years. Yang et al. [232], for instance, suggested, MERTA, a Long Short-Term Memory (LSTM) deep learning approach for MER where three attention networks are integrated with three VGGNets to extract spatio-temporal information and predict micro-expression class. The inputs to MERTA are enhanced using optical flow and optical strain, adopted from [87], and are given to the two VGGNet variants. Xia et al., in [226], suggested a spatio-temporal based enhanced version of RNNs to collectively consider both spatial as well as temporal patterns from micro-expression samples to distinguish

micro-expressions. Reddy et al. launched a recent study [173] which introduced two 3D-CNN based methods (MicroExpSTCNN and MicroExpFuseNet) for identifying micro-expressions through simultaneous extraction of spatial and temporal details using a 3D convolution operation to micro-expression videos.

Similarly, for self-learning feature extraction, Zhi et al. [246] proposed 3D convolutional neural networks (3D-CNN) model to illustrate facial micro-expressions. Khor et al. [86] recommended a lightweight dual-stream shallow network (DSSN) as a combination of condensed CNNs with different input characteristics in MER process. Recently, transfer learning has become one of the popular approach because of the small datasets. Lui et al. [126] introduces a neural micro-expression recogniser which implements an optical flow on onset and apex frames to detect the face motion. Next, a part-based average pooling model is applied to obtain discriminative information from the input source. Finally, to overcome the lack of training set, they proposed to transfer domain knowledge from macro-expression recognition tasks to micro-expression by adopting two domain adaptation methods, such as adversarial training and expression magnification and reduction (EMR). Furthermore, Sun et al. [187] proposed a novel knowledge transfer approach that transfer knowledge from action unit for micro-expression recognition. The network follows a teacher-student correlative framework where a pre-trained deep teacher network transfers knowledge to a shallow student network.

In several vision applications, deep CNNs have been very useful. The finding indicates that a salient hint includes small shifting of facial landmarks (e.g., eye-widening) when the subtle expressions reflect the emotional state. In order to include this prior information as an inductive bias for deep network design, multi-task learning is preferred. Multi-task learning is possible in numerous ways, such as joint learning, learning to learn, and learning with an auxiliary task are just a few terms included. Moreover, multi-task learning is often indirectly implemented with no reference, fine-tuning or transfer learning [187, 205, 246] are good examples. In general, when more than one loss function is optimised, it essentially becomes multi-task learning, unlike single-task learning. Although only one loss is optimised, as in a usual scenario, there is a possibility that an auxiliary function will further boost the primary task.

Multi-task learning is a paradigm which simultaneously learns a task with other associated tasks for classification problem. Multi-task learning's penalty term is an effective method to conduct the feature selection and model estimation problem. Various studies have demonstrated this structure experimentally to be beneficial [65, 70, 72]. He et al. [65] introduced a multi-task mid-level feature selection system, for micro-

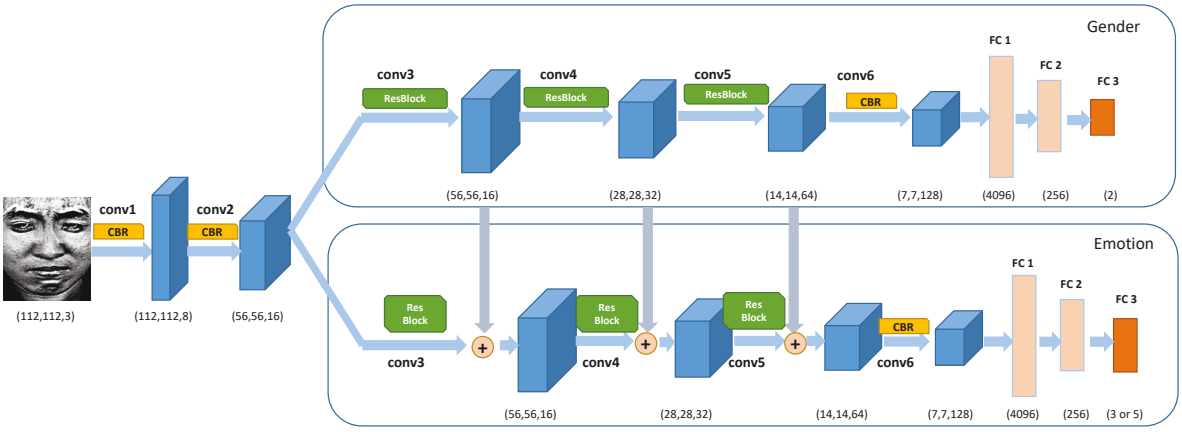


Figure 5.1: The comprehensive structure of GEME model for micro-expression recognition

expression recognition issue, by learning a series of class-specific feature mappings of the derived low-level LBP-TOP features. Hu et al. [70] proposed expansion of the local gabor binary pattern from three orthogonal planes (LGBP-TOP) employing pyramid histogram of centralised gabor binary pattern from three orthogonal planes. The gender-specific sparse multi-task learning system with adaptive regularisation concept is introduced to choose the appropriate micro-expression features to learn a compact subset of pyramid CGBP-TOP feature for micro-expression classification of different genders.

5.3 GEnder-based Micro-Expression (GEME) model description

Gender detection is ancillary to the primary function of recognising micro-expressions. In GEME, we are training the network to learn a pattern of how men and women express themselves using the facial muscles and predict the micro-expression shown in the input sequence. The input is provided to the network in the form of a dynamic image. The concept of dynamic imaging is followed [202] in the proposed approach, which brings together the subtle and unintentional motion of micro-expression image sequences in a single image. The dual-stream GEME processes these dynamic images that uses dynamic-aware features to depict the micro-expression class.

5.3.1 Pre-processing: Dynamic Imaging

Micro-expressions, in their essence, are quick and short-lived. Therefore, they are only seen in a few frames. The dynamic imaging approach is adopted to obtain these temporary shifts from the video.

A dynamic image is a regular RGB image representing the temporal and spatial details of an entire video sequence in one single instance. The method that Verma et al. [202] implements for producing precisely the same type of image has been applied in pre-processing the input video sequence for the proposed approach.

Video is referred to as a ranking method for generating a dynamic image through its frames $F_1, F_2, ...F_T$ i.e. $\sigma(F_T) \in R^d$ where $\sigma(F_T)$ reflects the RGB feature vector derived from each F_T frame. Eq. (5.1) measures the time-average φ_t of the available feature vector.

$$\varphi_t = \frac{1}{t} \sum_{T=1}^{t} \sigma(F_T) \tag{5.1}$$

Later, by using Eq. (5.2), the ranking method determines a value correlated with time t.

$$\psi(t \mid d) = < d, \varphi_t > \tag{5.2}$$

where, $d \in R^d$ illustrates a vector that evaluates the frame score in a video [184]. At time l, higher ranks are allocated to the frames i.e. $(l > t) \Rightarrow \psi(l \mid d) > \psi(t \mid d)$. Lastly, d

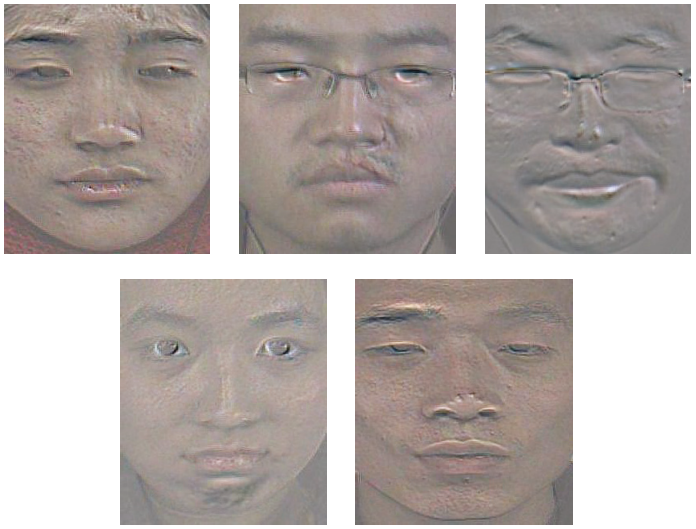


Figure 5.2: Dynamic images

is calculated by applying RankSVM [184] as described in Eq. (5.3) and Eq. (5.4).

$$d^* = \eta(F_1, F_2, ..., F_T; \sigma) = arg\ min(E(D)) \tag{5.3}$$

$$E(D) = \frac{\delta}{2} \parallel d \parallel^2 + \frac{2}{T(T-1)} \times \Sigma_{l>t} max\{0, 1 - \psi(l \mid d) + \psi(t \mid d)\} \tag{5.4}$$

Eq. (5.4) combines the solution of two key functions: one is quadratic regularisation function often implemented in SVMs, and hinge-loss soft-computing is the second function that indicates the number of pairs ($l > t$) incorrectly ranked by the rank function. Eq. (5.3), although, describes a function $\eta(F_1, F_2, ..., F_T; \sigma)$ that translates video frames into a single vector d^*. d^*, consequently, provides adequate details to rank all frame sequences in the video, it has combined details regarding all the frames and is often utilised as a video descriptor. It is seen from the resulting dynamic images that both uniform and non-uniform details are effectively retained within a single frame as shown in Figure 5.2. For detecting micro-expressions, the non-uniform differences serve a significant part. The input to GEME, for training and testing, are these dynamic images.

5.3.2 GEME Framework

GEME is a dual-stream architecture build to recognise micro-expressions from the dynamic images based on the gender feature. The proposed architecture of GEME is seen in Figure 5.1.

GEME consists of CBR blocks which is a Convolution, Batch Normalisation and ReLU function pipeline. The input dynamic image of dimensions $112 \times 112 \times 3$ is fed to the first CBR block. This first CBR block feature map is then given to the second CBR block. After passing the feature maps through second CBR block, the model is divided into two streams. One stream recognises the gender and the second stream recognises the micro-expression of the input dynamic images. There are two Fully Connected (FC1 and FC2) layers towards the end of each stream with an FC3 layer as the classifier to classify gender and micro-expression classes respective to each stream. The description of the various blocks of the model is given below.

5.3.2.1 CBR Block

The model begins with two CBR blocks. The functions within this block are Convolution (Conv), Batch normalisation (BN) and Rectified Linear Unit (ReLU) as shown in the Figure 5.3.

Convolution layer implements the convolution mechanism to the input image for the extraction of features and transfers them to the next layer. The layer comprises of a collection of neurons with acceptable weights and biases. The neuron weights are adjusted based on the activation map while adding any new feature. In GEME, the first CBR Conv layer imposes 8 filters of size 1×1 which are carried by another CBR Conv layer instituting 3×3 sized 16 filters. The feature map from second CBR is passed as input to the ResBlock. The final Conv block conv6 is also a CBR block which gets its input from ResBlock and passes its output to the FC layer.

Batch normalisation (BN) is a method applied to standardise inputs to a network, particularly for activations of a prior layer or specifically for inputs. In some instances, it expedites training by reducing the epochs in half or better and offers generalisation, thereby decreasing errors in generalisation. The purpose of BN is to ensure the distribution of activation values remain consistent during training. Using BN before the non-linearity is considered, in order to match the first and the second moments resulting in a stable distribution.

ReLU is an activation mechanism that appears and behaves similar to a linear function, but instead is a non-linear feature enabling complex relationships in the data to be learned. It is said to be sparsely activated as the function outputs 0 if the input is negative; however, for any positive input x, the output is the same value. It can then be translated as $f(x) = max(0,x)$. Since it converges the network quiet rapidly, it is also computationally competent. Another component named ResNet BasicBlock or ResBlock is introduced after CBR block.

5.3.2.2 ResBlock

ResNet utilises “skip-connections”- layers which do not do anything at the beginning. Such identical layers are skipped during training, and the activation functions from the previous layers are reused. This scales down the network to only a few layers, thereby speeding up the learning cycle. The identical layers extend and assist the system to

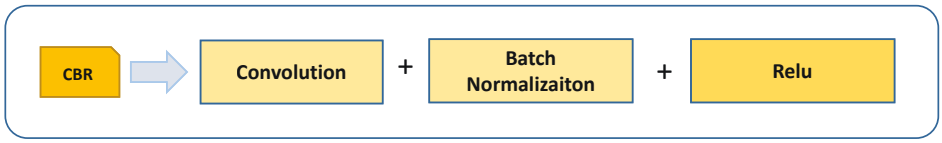


Figure 5.3: CBR Pipeline

examine more feature area when the network retrains. The ResNet's principal asset is its capacity to familiarise with various datasets and issues in a better way.

There are two variations of the skip connections as Identity Shortcut and Projection Shortcut. The identity shortcut bypasses the amount of data to the addition operator. The projection shortcut conducts a convolution process insuring the volumes remain the same size during addition operation. The implementation in GEME includes the projection shortcut that results in down-sampling by increasing the stride to 2. To maintain the time complexity for each process, the number of filters is duplicated ($56 \times 16 = 28 \times 32$).

The reason the ResNet blocks are incorporated in GEME framework is its ability to generalise well to different datasets and problems, which, in this case, is more necessary as the network is trained to be capable of learning gender as well as micro-expression properties. The structure contains three ResNet BasicBlocks each with 16, 32 and 64 filters of identical size 3×3 where a projection shortcut connection is between each pair of 3×3 filters as shown in the Figure 5.4.

5.3.2.3 Summation Function

In GEME, this layer adds the output feature maps of Convolutional blocks from conv3 to conv5 of Gender stream with the output feature maps of Convolutional blocks conv3 to conv5 of Emotion stream respectively feature-by-feature and passes the newly formed summation feature to the next Conv block of Emotion stream as can be seen in Figure 5.1. For example, feature maps from conv3 block of Gender stream is added with the feature map from conv3 of Emotion stream element-wise in the same channel and position and this fused feature map is fed to the next conv4 block of Emotion stream. This is repeated

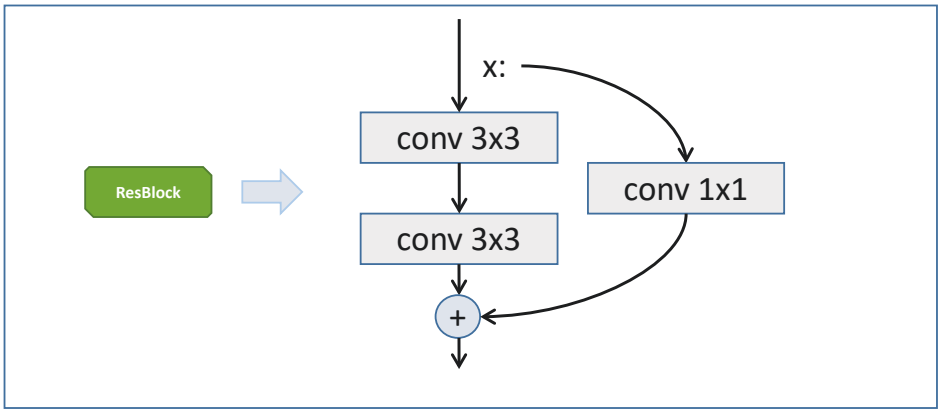


Figure 5.4: ResNet BasicBlock

for Conv blocks 3, 4, and 5. The purpose of adding gender features with emotion features is to highlight the importance of the gender features for improving micro-expression recognition accuracy. The unique gender characteristics are utilised for recognising unique micro-expression features related to particular gender. The input to conv6 CBR block is the addition of output of conv5 Gender features and output of conv5 Emotion features.

5.3.2.4 Fully Connected Layers (FC)

Within this layer, analogous to the multi-perceptron neural networks, the previous layer activations are completely linked to the neurons. The neuron activation is determined by applying a bias offset as matrix multiplication. The fully connected input layer (FC1) "flattens" the output into a single vector for the next layer. The next fully connected layer (FC2) uses the feature analysis to determine the accurate label by using weights. Eventually, the fully connected output layer (FC3), a classifier node, provides absolute probabilities for each label.

Dropout layer has been introduced to enhance the sensitivity of neurons for particular weight and to address the issue of training data overfitting.

5.3.2.5 Losses

The framework has two streams performing two different tasks; therefore, there are two different losses implemented. For gender detection, Cross-entropy loss is implemented, whereas, for micro-expression recognition using gender features, Focal loss is chosen [8, 54, 98, 116, 125]. The Class-Balanced (CB) Focal loss and CB Sigmoid Loss [25] are applied for the LOSO CV evaluation.

Gender Loss

Cross-entropy Loss: Cross-entropy loss or log loss calculates the efficiency of the classification model with a likelihood value between 0 and 1. The standard cross-entropy loss for classification can be written as [116]

$$CE(p,y) = \begin{cases} -log(p) & if\ y = 1 \\ -log(1-p) & otherwise \end{cases} \tag{5.5}$$

where $y \in \{\pm 1\}$ determines the ground-truth class and $p \in [0,1]$ is the predicted probability of the model with $y = 1$ as the class label. More concisely, $CE(p_t) = -log(p_t)$ where

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases} \tag{5.6}$$

As the Gender stream of the proposed network is bound to get the predictions between 'male' and 'female' which is binary, hence, cross-entropy loss is implied here.

Emotion Loss

Focal Loss: A balanced dataset has equally distributed target labels in the multi-class classification. However, when the number of samples in one category is exceedingly higher than the other, such datasets are regarded as imbalanced datasets. Such an imbalanced dataset results in a weakly trained model. Relevant data augmentation strategies for image classification are applied in order to build synthetic data for under-represented classes.

Lin et al. [116] introduced a focal loss function to address the problem of class imbalance for dense object detection, a binary classification problem, by re-formulating the cross-entropy (CE) loss. Although, the CE can be extended to deal multi-class classification issues. The discrete CE version expression is [125]

$$H(p,y) = -\sum_{i=1}^{n} y_i\ log\ p_i \tag{5.7}$$

where n corresponds to the number of all possible distinct distribution bins. Similar to Eqs. (5.5) and (5.6), y_i is the ground-truth probability, and p_i is the prediction probability. All the y_i's are zero except one and all the p_i's will be non-zero according to the definition of sigmoid or Softmax function and $log(0)$ cannot be calculated.

The emphasis of the focal loss is to train on an inadequate array of hard instances, avoiding large amounts of easy negatives from disrupting the network during training. The multi-class focal loss addresses the class imbalance by down-weighting easy samples in order to reduce their contribution to the total loss despite their large number, particularly, aiming its attention on minority samples for training.

An approach adopted for solving the class imbalance issue is to include a modulating factor $\alpha(1-p_t)^{\gamma}$ to multi-class CE loss. The multi-class focal loss can be defined as in Eq. (5.8) [125]

$$FL(p_t) = -\sum_{i=1}^{C} \alpha_t\ (1-p_t)^{\gamma}\ log(p_t) \tag{5.8}$$

where C is the total number of classes and i is the class number.

This is typically a reverse class frequency or is interpreted as cross-validation hyperparameter set. The focusing parameter $\gamma \geq 0$ is tuned for down weighing the easy samples, and the balancing variant $\alpha_t \geq 0$ denotes the weight for each class which is utilised to magnify the significance of the minority class. It can be observed that the modulating weight factor $\alpha(1-p_t)^{\gamma}$ with CE loss is reliant on the p_t value. The weight is small if p_t is large and the weight is large when p_t is small.

The multi-class Focal loss mentioned in Eq. 5.8 is implemented for 5-fold CV experiments.

Class-Balanced Focal Loss: Cui et al. [25] developed a re-weighting method that re-balances the loss using the effective number of samples within each class, resulting in a class-balanced loss. The effective number of samples is defined as "the sample volume, and it can be determined using a simple formulation as $(1-\beta^n)/(1-\beta)$, where n is the number of samples and $\beta \in [0,1)$ is a hyperparameter" [25].

A weighting factor α_i, which is inversely proportional to the effective number of samples for class i, is adopted to balance the loss: $\alpha_i \propto 1/E_{n_i}$. In order to render the overall loss approximately in the same range while implementing α_i, α_i is normalised such that $\sum_{i=1}^{C} \alpha_i = C$.

Conventionally, a weighting factor $(1-\beta)/(1-\beta^{n_i})$ is applied to the loss function provided a sample of class i containing a total of n_i samples, and hyperparameter $\beta \in [0,1)$. The class-balanced (CB) loss is represented as [25]:

$$CB(\mathbf{p}, y) = \frac{1}{E_{n_y}} \mathcal{L}(\mathbf{p}, y) = \frac{1-\beta}{1-\beta^{n_y}} \mathcal{L}(\mathbf{p}, y) \tag{5.9}$$

where n_y is the number of samples in the ground-truth class y, prediction probabilities of the model denoted as $\mathbf{p}$; the loss is reported as $\mathcal{L}(\mathbf{p}, y)$ and effective number of samples for class y is E_{n_y}.

Class-balanced loss can be applied regardless of the loss function preference. For LOSO CV, the focal loss is used with class-balanced loss, i.e. CB Focal loss. The original focal loss is as written in Eq. 5.8, and the CB Focal loss can be presented as shown in Eq. 5.10 [25].

$$CB_{focal}(\mathbf{z}, y) = \frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1-p_i)^{\gamma} \, log(p_i) \tag{5.10}$$

where $\mathbf{z}$ is the predicted output for all classes, C is the total number of classes, and p_i is the probability distribution over all classes.

It can be observed from Eq. 5.8 and 5.10 that the original focal loss uses α as a balancing variant. This implies that an α-balanced focal loss is equivalent to class-balanced focal loss when $\alpha_t = (1-\beta)/(1-\beta^{n_y})$. The class-balance term can thus be regarded as an explicit means of determining α_t in focal loss depending on the effective number of samples.

The Class-Balanced Focal loss mentioned in Eq. 5.10 performed better for LOSO CV experiments on CASME II, SAMM and Combined 3DB databases.

Class-Balanced Sigmoid Loss: Unlike softmax, class-probabilities determined by sigmoid rule suggest that each class is distinct rather than mutually exclusive. By utilising the sigmoid function, the multi-class visual recognition is interpreted as several binary classification activities in which network,Äôs output node executes a one-vs-all classification to estimate the likelihood of the target class over the remainder of classes. Sigmoid potentially has two benefits for real-world datasets compared to softmax : (1) Sigmoid believes that the classes are not mutually exclusive and are very much in line with the real-world results where a number of classes may be identical particularly with vast number of fine-grained groups. (2) As each class is distinct with its own predictive model, sigmoid integrates single-label classification with multi-label prediction. It is a desirable attribute as the real-world data typically have more than one semantic label.

From Eqs. 5.6, 5.9 and [**?**], the sigmoid cross-entropy loss can be denoted as in Eq. 5.11

$$\begin{aligned} CE_{sigmoid}(\mathbf{z}, y) &= -\sum_{i=1}^{C} log(sigmoid(z_i)) \\ &= -\sum_{i=1}^{C} log\left(\frac{1}{1+exp\ (-z_i)}\right) \end{aligned} \tag{5.11}$$

Hence, similar to CB Focal Loss, CB Sigmoid cross-entropy loss can be represented as in Eq. 5.12

$$CB_{sigmoid}(\mathbf{z}, y) = \frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} log\left(\frac{1}{1+exp\ (-z_i)}\right) \tag{5.12}$$

The Class-Balanced Sigmoid loss mentioned in Eq. 5.12 is used for LOSO CV on SMIC database.

Total Loss

As mentioned earlier, GEME uses two loss functions, hence the total loss for the GEME network can be calculated as in Eq. (5.13)

$$Loss_{Total} = 1.0 \times Loss_{Gender} + \lambda \times Loss_{Emotion} \tag{5.13}$$

where $Loss_{Gender}$ is the Cross-entropy loss of Gender stream, $Loss_{Emotion}$, for Emotion stream, is the multi-class Focal loss for 5-fold CV, CB Focal loss for LOSO CV on CASME II, SAMM and Combined 3DB and CB Sigmoid loss for SMIC database. λ represents the ratio of $Loss_{Gender}$ in the
$Loss_{Total}$. The best results are achieved with $\lambda = 0.5$ for individual datasets and $\lambda = 0.7$ for Combined 3DB.

The hyperparameter settings for all the loss functions used are explained in Section 4.2.

5.4 Experimental setup and Outcomes

This section provides information about the different databases used and various parameter settings for performing the experiments.

5.4.1 Databases

Analyses are performed on standard micro-expression repositories including the Chinese Academy of Sciences Micro-Expression database's updated version, i.e. CASME II [229], Spontaneous MIcro-expression Corpus (SMIC) [112] and Spontaneous Actions and Micro-Movements (SAMM) [29]. More information is provided below on the four spontaneous micro-expression databases used in this study.

Yan et al. from the Institute of Psychology, Chinese Academy of Science have compiled CASME II [229]. The collection comprises of 246 micro-expression samples which gave a high spatial and temporal resolution of 200 fps out of 26 participants, including 14 females and 22 males, with an average facial area of approximately 280×340 pixels. The database was collected in a sophisticated laboratory conditions with appropriate lighting preventing unregulated illumination. The samples are divided into five major categories, including disgust (63), happiness (32), repression (27), others (99) and surprise (25). Besides, there are also smaller categories in the dataset such as sadness (7) and fear (2), with the corresponding number of micro-expression samples in the brackets.

Li et al. collected a Spontaneous Micro-expression Corpus (SMIC) [112] at the University of Oulu. The database is comprised of three sets: High speed (HS), Near-InfraRed

(NIR) and normal VISual camera (VIS). The HS set consists of 164 ME clips from 16 participants (7 female and 9 male) captured at 100 fps and 280×340 face image resolution whereas NIR and VIS both are captured at 25 fps from 8 participants giving 71 micro-expression sequences with a resolution of 640×480. The sequences are classified under positive, negative and surprise emotion classes. The HS dataset is chosen for experiments in the current analysis with positive (51), negative (70) and surprise (43) samples solely due to the number of samples as compared to NIR and VIS.

Davison et al. [29] indicated that micro-expression repositories, today, have been deprived of ethnic diversity. Spontaneous Actions and Micro-Movements (SAMM) database was therefore developed. The SAMM database is recorded using a grey-scale sensor with an average facial size of 650×960 at 200 frames per second and 2040×1088 pixels resolution in regulated lights to avoid distortion when recording at a high rate. Participants representing 13 ethnicities from 32 subjects and even a gender split with 17 male and 16 female have been reported in this database. SAMM is reported as the first high-resolution collection of 159 spontaneously triggered micro-movements with the most extensive demographic heterogeneity. These samples are categorised into eight micro-expression classes. Contrary to the prior datasets, participants first had to fill in a questionnaire before experiments were performed. The experiment instructor shows videos that are appropriate to the responses given. All the videos captured are FACS coded with limited focus on emotional annotations. It should be noted that there are only a few micro-expression classes in SAMM with more than ten samples for experimentation. The classes includes anger (57), contempt (12), happiness (26), surprise (15), and others (26), with the corresponding number of micro-expressions in the database mentioned in brackets.

5.4.2 Setup and Parameters

The GEME model implementation is done using the open-source platform, Pytorch [158]. The GPU used for performing the computations is NVIDIA GeForce GTX 1080 with 8GB memory. The training and testing images are of size 112×112. During training, SGD optimisation technique is used with an initial learning rate of 0.001 and L2 normalisation is implemented to prevent overfitting. The other parameter values, such as weight decay and momentum, are set to 0.001 and 0.9, respectively. The batch size is 64, and the model training is executed for 100 epochs for individual databases where the learning rate is reduced to half of the previous learning rate at the 40^{th} and 70^{th} epoch, respectively. While the model is trained for 60 epochs on combined database by reducing the learning

rate to half of the previous learning rate at the 40^{th} and 50^{th} epoch, respectively. The dropout value during training is set to 0.3 for 5-fold CV experiments and 0.5 for LOSO CV evaluation.

The databases used for the analysis are CASME II, SMIC and SAMM. The validation schemes implemented to evaluate the GEME model are 5-fold cross-validation and Leave-One-Subject-Out (LOSO). In 5-fold cross-validation (CV), each database is divided into five parts where four parts are used for training and the fifth part is given as the test set. The 5-fold CV approach is motivated from the study presented by Verma et al. [202] with a better distribution of the samples in train and test sets. The database distribution for 5-fold training and testing is done in an $80:20$ (Training:Testing) ratio.

The second validation technique is the most widely used Leave-One-Subject-Out (LOSO) cross-validation technique where one subject from the training database is held out for testing and the process is iterated for the number of subjects in the dataset. An additional evaluation is performed where the datasets CASME II, SAMM and SMIC are combined to address the issue of lack of dataset availability. This evaluation is called as the Composite Database Evaluation (CDE) using LOSO CV [176].

Focal Loss:

The micro-expression databases are split into 3 and 5 emotion classes. Moreover, there are noticeable differences in the number of samples in different categories of each database as depicted in Figure 5.5. It is difficult to classify some of the expression categories such as surprise (25) from CASME II, or contempt (12) from SAMM, in contrast to the categories others (99) from CASME II and negative (71) from SMIC which are more easily classified. The focal loss balancing parameter value α is set depending on the number of samples in each database, and focusing parameter γ value is set to 2. The following Eq. (5.14) defines the formulation for α in the experiments.

$$\alpha_i = \frac{max(\{S_i : i \in L\})}{S_i} \tag{5.14}$$

where,

$$L = \{i : i \text{ is a label in respective micro-expression dataset}\},$$

and S_i is the number of samples in label i of the dataset. For instance, let's calculate α for CASME II database. Set of labels is,
$L = \{Disgust, Happiness, Repression, Surprise, Others\}$ and set of number of samples

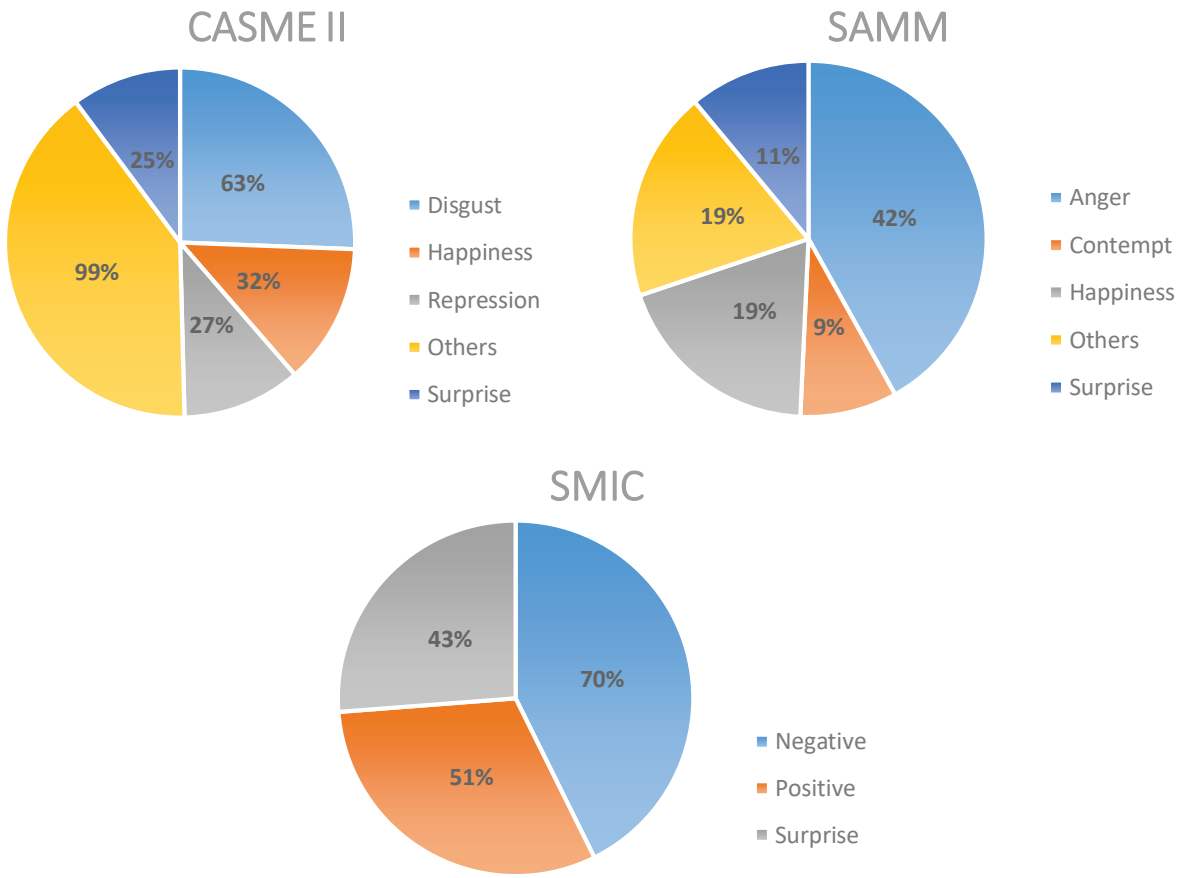


Figure 5.5: Pie charts to illustrate the class imbalance problem for the four databases used

corresponding to each label is $S_i = \{63, 32, 27, 25, 99\}$. Therefore,

$$max(\{S_i : i \in L\}) = max(63, 32, 27, 25, 99) = 99$$

and,

$$\alpha_{Disgust} = 99/63 = 1.57$$

Similarly, α value can be calculated for each label in CASME II dataset.

However, the elements of set L would differ based on the label in the database used and the number of samples S corresponding to each label in that database.

CB Focal Loss and CB Sigmoid Loss:

There are two different losses calculated during the LOSO CV; the parameter values for CB Focal loss are $\beta = 0.9999$ and $\gamma = 2$ and CB Sigmoid loss with $\beta = 0.99$ as per the settings in [25].

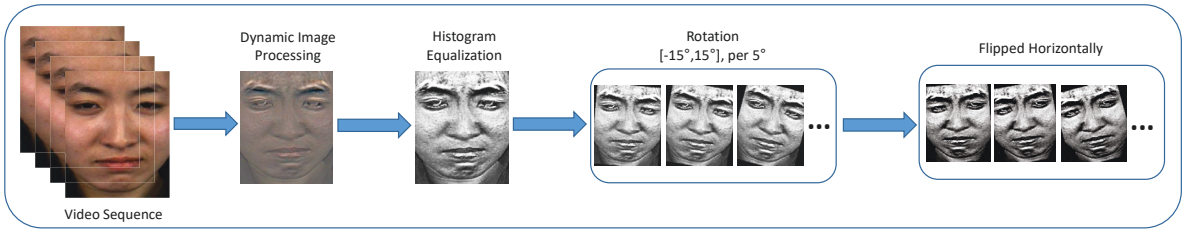


Figure 5.6: Data augmentation process

Data Augmentation:

To overcome the difficulty of scarce data and prevent overfitting, data augmentation is implemented in the experiments. Each of the newly generated dynamic images is enhanced by performing histogram equalisation and rotated between $[-15°, 15°]$ with 5° increment. Then all the generated images are horizontally flipped. The data augmentation process is illustrated in Figure 5.6.

5.4.3 Performance Metrics

There is a clear imbalance in the class distribution of composite database, i.e. negative:positive:surprise classes are distributed in 3:1.3:1 ratio (Figure 5.5). The outcome is recorded using three appropriate measures in order to manage these class imbalances better.

5.4.3.1 Accuracy (Acc)

The recognition accuracy for both the validation schemes is calculated using the following formula [202].

$$Recog.\ Acc. = \frac{Total\ no.\ of\ correctly\ predicted\ samples}{Total\ no.\ of\ samples} \times 100 \tag{5.15}$$

5.4.3.2 Unweighted F1-score (UF1)

The metric, also known as macro-averaged F1-score, is an ideal choice to offer equal consideration to rare categories in imbalanced multi-class environments. In this calculation, all the True Positives (TP), False Positives (FP) and False Negatives (FN) for each class c (of C classes) over all LOSO k-folds are initially obtained, and later relevant F1

scores are determined. The average of F1 score per class is measured to evaluate the unweighted F1-score (UF1), $F1_c$:

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{5.16}$$

$$UF1 = \frac{F1_c}{C} \tag{5.17}$$

5.4.3.3 Unweighted Average Recall (UAR)

This measurement is also called as balanced accuracy and is considered as a better measure than the standard Accuracy metric (or Weighted Average Recall) which could predict the more significant categories correctly. Similarly, the accuracy values per class are determined first, before being averaged by the number of classes:

$$UAR = \frac{1}{C} \sum_c \frac{TP_c}{n_c} \tag{5.18}$$

where n_c denotes the number of samples of the c^{th} class. UF1 and UAR offer a fair estimation of whether an approach can anticipate that a method might only function effectively for specific categories.

5.4.4 GEME Multi-task outcome comparison with state-of-the-art methods

GEME Multi-task approach provides valuable justification for using the gender features combined with micro-expression features. Tables 5.1-5.3 compare GEME model with the state-of-the-art approaches on the selected databases and similar micro-expression classes. The published results are taken for the compared approaches taking into consideration the databases used and the number of micro-expression classes.

It is observed that most of the recognition approaches use CASME II due to long video sequences, and also that CASME II and SMIC are introduced earlier than SAMM. Also, as the comparison criteria, here, is depending on the number of classes used for classification, many existing studies could not be considered for comparison. The recognition results using 5-fold CV scheme are observed to achieve higher accuracy than some of the existing approaches.

Considering CASME II, SAMM and SMIC results when compared to the state-of-the-art methods, GEME achieves comparable results and better than most of the existing

results justifying the importance of including gender features, and also that person of different gender expresses emotions uniquely and differently. Combining the micro-expression features extracted from a dynamic image along with unique gender features produces a comprehensive description of the facial features of a person enabling the network to learn thoroughly.

Another factor contributing to the improved accuracy is the use of the loss function. Rigorous experiments were performed to determine the suitable loss function for each database during the LOSO CV scheme. The CB Focal loss function delivered higher results for CASME II and SAMM databases while CB Sigmoid loss achieved better accuracy for SMIC database.

Class wise recognition performance of CASME II on single-task as well as multi-task frameworks can be seen in Figure **??** (a) and (b) respectively. It is observed that GEME shows improvement in the micro-expression recognition accuracy by combining gender feature maps. Figure 5.7 (c) and (d) illustrates the confusion matrix for SAMM database, where Figure 5.7 (c) shows signs of overfitting using single-task framework due to the small data samples. However, GEME, in LOSO validation, does not seem to alleviate the overfitting problem completely. The confusion matrix (Figure 5.7 (d)) demonstrates

Table 5.1: Comparing the recognition accuracy of GEME with modern approaches on CASME II database

Ref. No.	Method	Classes	Accuracy (%)	F1-score
[228]	FDM	5	41.96%	0.4700
[87]	ELRCN	5	52.44%	0.5000
[108]	OF+CNN	5	56.94%	N\A
[88]	CNN+LSTM	5	60.98%	N\A
[141]	TIM+DCNN+SVM	5	64.90%	N\A
[248]	Hierarchical STLBP-IP + KGSL	5	65.18%	0.6254
[70]	Pyramid CGBP-TOP	5	65.8%	N\A
[246]	3D-CNNs (with transfer learning)	5	65.90%	N\A
[111]	EVM+HIGO	5	67.21%	N\A
[86]	DSSN	5	70.78%	0.7297
[86]	SSSN	5	71.19%	0.7151
[185]	TSCNN-I	5	74.05%	0.7327
[185]	TSCNN-II	5	80.97%	0.8070
–	**GEME (MTL; LOSO)**	**5**	**75.20%**	**0.7354**
–	**GEME (MTL; 5 fold)**	**5**	**77.24%**	**0.7528**

Table 5.2: Comparing the recognition accuracy of GEME with modern methods on SAMM database

Ref. No.	Method	Classes	Accuracy (%)	F1-score
[243]	LBP-TOP	5	34.56%	0.2892
[67]	LBP-SIP	5	36.03%	0.3133
[111]	HIGO-TOP	5	41.18%	0.3920
[86]	SSSN	5	56.62%	0.4513
[86]	DSSN	5	57.35%	0.4644
[185]	TSCNN-I	5	63.53%	0.6065
[185]	TSCNN-II	5	71.76%	0.6942
–	**GEME (MTL; LOSO)**	**5**	**52.21%**	**0.4433**
–	**GEME (MTL; 5 fold)**	**5**	**65.44%**	**0.5467**

Table 5.3: Comparing GEME recognition accuracy with state-of-the-art methods on SMIC database

Ref. No.	Method	Classes	Accuracy (%)	F1-score
[228]	FDM	3	54.88%	0.538
[248]	Hierarchical STLBP-IP + KGSL	3	60.37%	0.6125
[70]	Pyramid CGBP-TOP	3	59.4%	N\A
[86]	SSSN	3	63.41%	0.6329
[86]	DSSN	3	63.41%	0.6462
[200]	CapsuleNet	3	58.00%	0.5900
[185]	TSCNN-I	3	72.74%	0.7236
-	**GEME (MTL; LOSO)**	**3**	**64.63%**	**0.6158**
-	**GEME (MTL; 5 fold)**	**3**	**65.24%**	**0.6431**

a slight increase in recognition accuracy for “Anger”, “Happiness” and “Others” classes. Due to the fact that SAMM consists of participants from diverse cultural backgrounds and class-wise gender imbalance affects the recognition accuracy. Figures 5.7 (e) and (f) depicts the confusion matrics for SMIC database. Two out of the three classes shows improved results for GEME compared with single-task.

5.4.5 Composite Database Evaluation (CDE)

The composite database evaluation (CDE) is another approach followed for ablative experiments. Based on the instructions given in MEGC 2019 challenge [176], the samples of the three widely used spontaneous databases CASME II, SAMM and SMIC are combined to form a single composite database with generalised emotion classes. Training

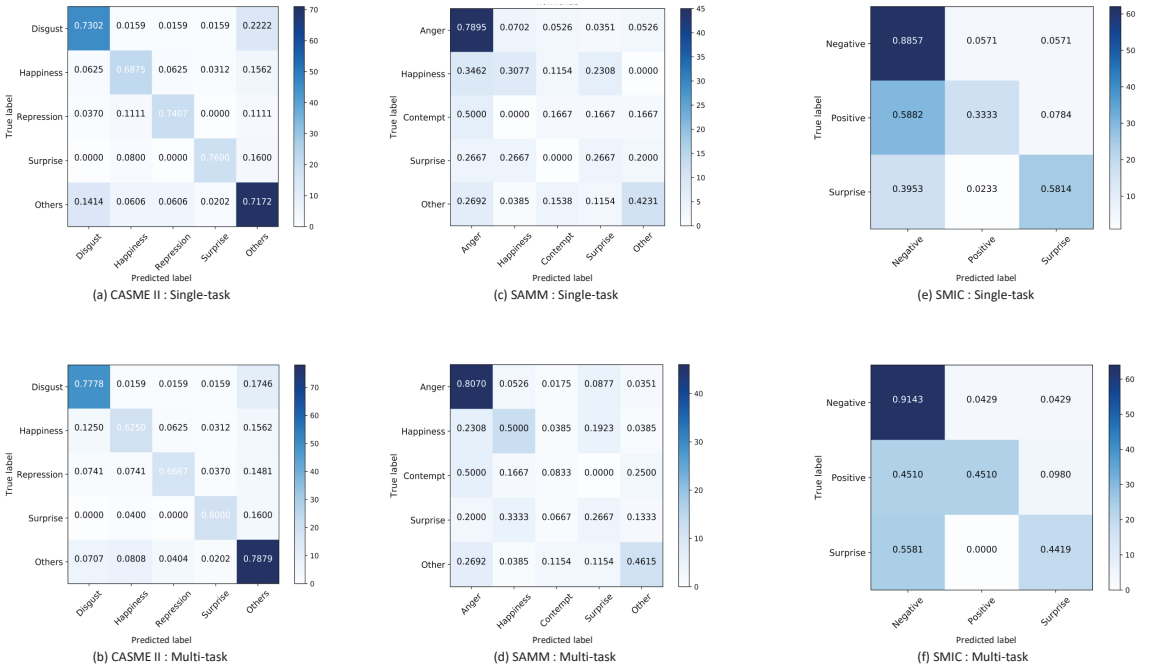


Figure 5.7: Confusion matrix using LOSO validation scheme for single-task network and multi-task GEME network on individual databases

and Test set distribution is determined by LOSO CV method. To enable the three databases to be used together, the emotion classes are combined to reduce and generalise the categories. The elementary classes are mapped to three distinct classes Negative, Positive and Surprise. The categories combined to form these three classes are:

* Negative: Repression, Anger, Contempt, Disgust, fear, Sadness
* Positive: Happiness
* Surprise: Surprise

The emotion category “Others” from CASME II and SAMM are discarded because the samples have unrelated and non-specific emotions. Table 5.4 shows a summary of the number of samples for all three databases. A total of 68 subjects from three databases (24 from CASME II, 28 from SAMM and 16 from SMIC) are repeatedly evaluated for 68 times by holding out samples from one test subject and training the network on remaining samples from 67 subjects. The CDE approach depicts real-time circumstances with increased number of participants from variable environmental settings such as illumination with diverse identity backgrounds, including ethnicity, gender, and the emotional intensity in a single recognition model. The LOSO CV also ensures a subject-independent evaluation. Table 5.5 presents compiled outcomes using UF1 and UAR against some handcrafted

Table 5.4: Sample distribution after combining three databases CASME II, SAMM and SMIC into three classes for CDE

Emotion class	CASME II	SAMM	SMIC	3DB-combined
Negative	32	26	51	109
Positive	88	92	70	250
Surprise	25	15	43	83
TOTAL	145	133	164	442

Table 5.5: Combined 3DB LOSO CV performance compared against various baseline and recent methods

Method	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [243]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
Bi-WOOF [122]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	20.5139
OFF-ApexNet [55]	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
Quang et al. [200]	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
Zhou et al. [247]	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
Liong et al. [118]	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
Liu et al. [126]	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152
GEME (Single task)	0.7395	0.7500	0.6288	0.6570	0.8401	0.8508	0.6868	0.6541
GEME (Multi-task)	0.7221	0.7303	0.6038	0.6387	0.8831	0.8790	0.5843	0.5455

benchmark methods as well as some latest micro-expression recognition approaches based on CNN. It is clearly visible from the results that in comparison to CASME II, SAMM and SMIC databases persist to be more challenging datasets with the use of CB focal loss function. Perhaps the factors affecting the performance could be: the low resolution and slow frame rate used during the collection of SMIC while the diversity in age and nationality of participants in the SAMM dataset.

Figure 5.8 (a-h) demonstrates the confusion matrix for 3 combined databases and validation on individual database from the combined 3DB. The Figure 5.8 (a) and (b) is the evaluation of the combined three databases using single-task stream and multi-task GEME. The results for “Negative” class significantly improved in the multi-task system whereas the results decreased for “Positive” and “Surprise” classes. This could be the effect of overfitting. However, the results for CASME II database of Combined 3DB (Figures 5.8 (c) and (d)) shows slight improvement for multi-task approach as compared to single-task. Figures 5.8 (e) and (f) are results for SAMM database and Figures 5.8 (g) and (h) are results for SMIC database for which the single-task performs slightly better

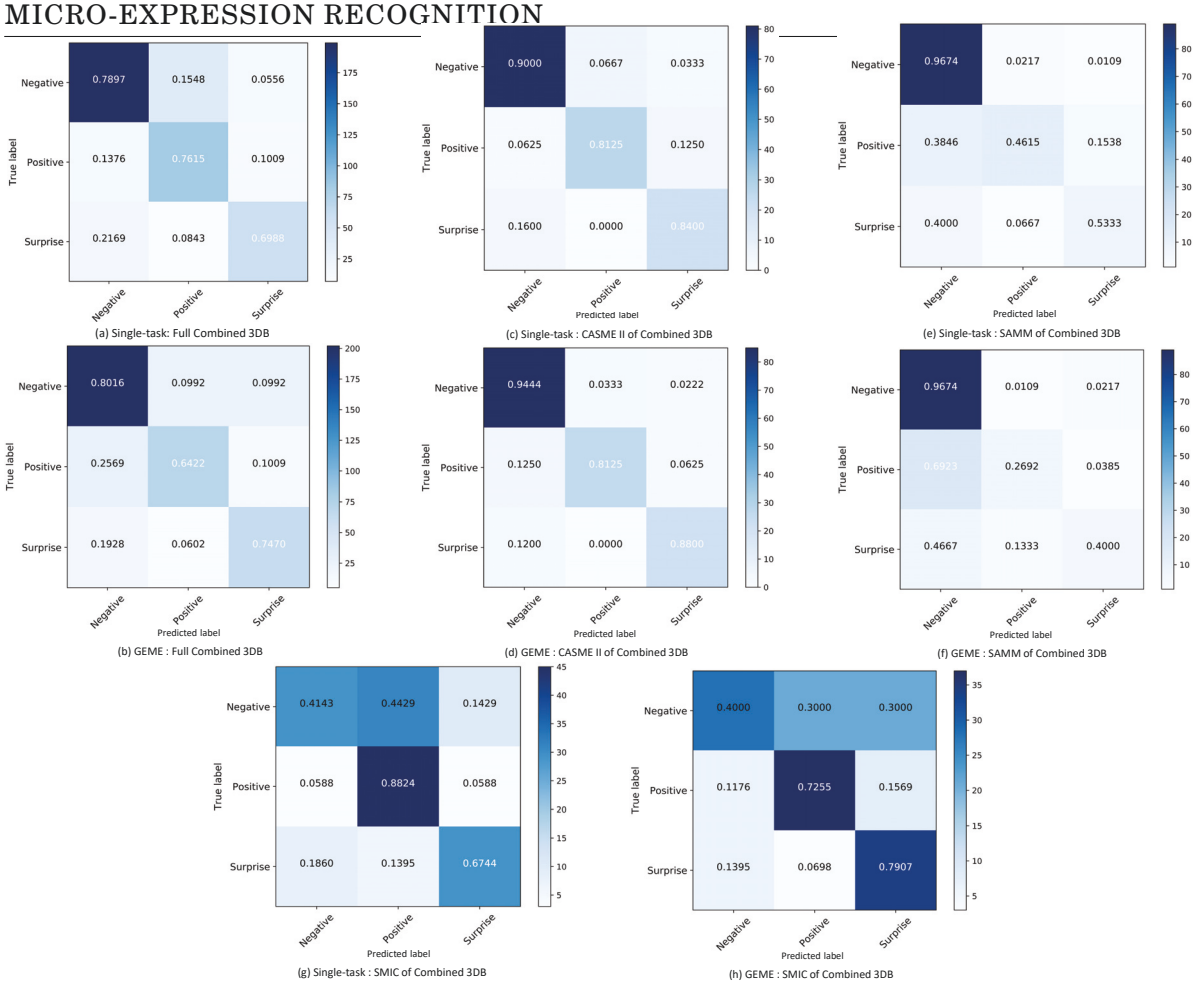


Figure 5.8: Confusion matrix using LOSO validation scheme for single-task network and multi-task GEME network on Combine 3DB and individual databases

than the multi-task approach. The probable reason for reduced classification results is the number of samples available in each of the micro-expression category as well as gender-wise sample distribution.

5.4.6 Ablative Analysis

The ablative study is performed by removing the Gender stream from the GEME framework and performing only micro-expression recognition. The single micro-expression recognition stream is as shown in the Figure 5.9.

The task is labelled as *Single-task* in Tables 5.6 and 5.7, and *Multi-task* is using both Gender and Micro-expression streams. The Single-task evaluation is also done using 5-fold and LOSO cross-validation. The original Focal loss achieved acceptable results for 5-fold evaluation whereas the LOSO CV performance was improved by using

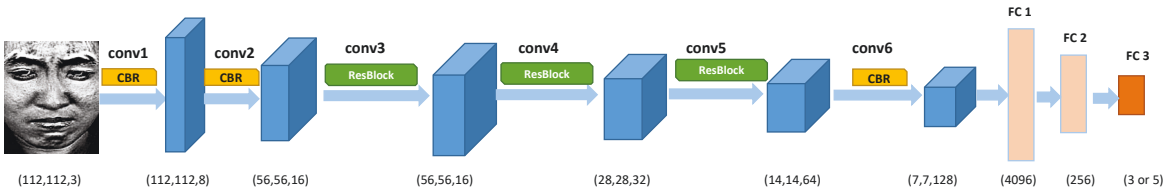


Figure 5.9: Single-task Micro-expression recognition network

Table 5.6: Performance metrics for Single-task and Multi-task learning using 5-fold validation approach

	Single-Task			Multi-Task		
	Acc.(%)	UF1	UAR	Acc.(%)	UF1	UAR
CASME II	72.76%	0.7073	0.6841	77.24%	0.7528	0.7312
SMIC	64.63%	0.6376	0.6297	65.24%	0.6431	0.6375
SAMM	59.56%	0.4659	0.4640	65.44%	0.5467	0.5414

Table 5.7: Performance metrics for Single-task and Multi-task learning using LOSO validation approach

	Single-Task			Multi-Task		
	Acc.(%)	UF1	UAR	Acc.(%)	UF1	UAR
CASME II	72.36%	0.7255	0.7271	75.20%	0.7354	0.7315
SMIC	63.41%	0.6055	0.6001	64.63%	0.6158	0.6023
SAMM	51.47%	0.3962	0.3907	55.88%	0.4538	0.4635

Class-Balanced (CB) Focal loss for CASME II and SAMM databases and CB Sigmoid loss for SMIC database.

Tables 5.6 and 5.7 reports the performance of the single-task and multi-task GEME model for 5-fold and LOSO cross validation, respectively, on the databases used. The outcomes presented in these tables makes it more evident that the inclusion of gender features trains the network to learn unique features and as a consequence improves the recognition accuracy. It can also be observed from Tables 5.6 and 5.7 that the single-task analysis attains slightly higher recognition accuracy as compared to the state-of-the-art methods except for a few methods. Therefore, in this work, a new micro-expression recognition single-task model is also proposed.

5.4.7 Qualitative Analysis

The examples of prediction results of the proposed single-task and multi-task (GEME) networks with the ground-truth of CASME II database are as presented in Figure 5.10.

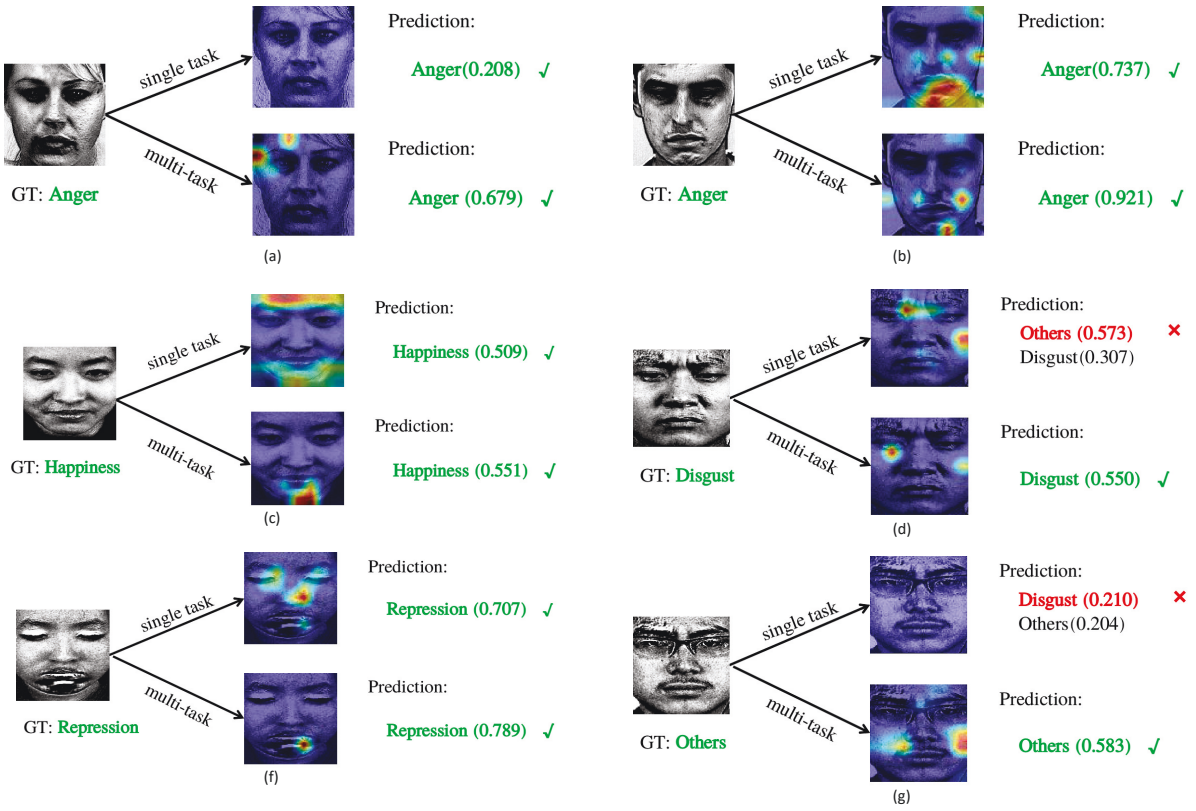


Figure 5.10: Examples of recognition performance. A check mark (✓) represents a correct prediction result, while a cross mark (×) represents an incorrect prediction result. The value in the parentheses is the confidence of the prediction.

In order to highlight the difference between single-task network and GEME network's attention to micro-expression features, Grad-CAM [177] is implemented to visualise the activation map of the last convolutional layer.

The GEME network fuses gender features with micro-expression features and the network learns better about the facial properties distinct to male and female face. Hence, by using Grad-CAM to return the activation intensity of the final Conv block in the form of heatmap visualisation, as demonstrated in Figure 5.10, it is obvious that GEME is capable of highlighting appropriate facial regions of significant facial muscle movement in the video giving accurate predictions about the micro-expression classes.

5.5 Summary

The research presented in this chapter aims in validating and confirming the significance of incorporating gender characteristics in the process of micro-expression recognition.

In this work, a dual-stream network is designed to perform multiple tasks of gender recognition and micro-expression recognition. The input given to the network is generated as a dynamic image which is a regular RGB image representing spatial and temporal details of a video in a single image frame.

The extracted image features are fed to the two separate streams of the network. The Gender stream extracts gender-related features and is added to the micro-expression related features extracted from the similar block of the parallel Emotion stream. They are then given to the next layer in the Emotion stream. The classifier layer FC3 classifies the input dynamic image in one of the micro-expression class.

The evaluation results demonstrate that the inclusion of gender features adds an additional depth of detail to the micro-expression feature, which is unique to the person and the micro-expression being displayed. As compared to the state-of-the-art approaches, the proposed GEME model demonstrates an acceptable micro-expression recognition accuracy and testifies the influence of gender on micro-expressions. It was found that the databases with different gender ratios did impact the experimental results to some extent. Similarly, the single-task model for micro-expression recognition also delivers slightly better results when compared with the state-of-the-art methods. The preceding research is believed to provide a new perspective in capturing discriminative human characteristics for classification of captured micro-expressions.

Even though GEME is performing well with comparable results, there are some limitations observed. First, the proposed model is not an end-to-end learning network since the pre-processing module processes the raw input external to the network. The pre-processes dynamic image is given as the input to GEME. As a part of the future works, the pre-processing unit will be remodelled to embed it within the network to form an end-to-end learning network. Second, during the experiments and testing phase, it is encountered that even if the overall female to male ratio of the database is somewhat balanced, the emotion category-wise gender ratio is imbalanced which can also be seen from the results. Hence, balancing the gender ratio for each emotion class is considered as the future work for this study. Transfer learning is one of the possible solutions to mitigate the gender imbalance where the network can be trained on either facial expression databases or specific gender annotated databases.

Apart from gender features, based on the psychological studies [38, 50, 80, 179, 222], age and cultural background or ethnicity of the subject also alters human emotions. Further extending the current study to include the age and nationality of the participants can mimic the real-world scenario.

The challenge in including the age and ethnicity features for micro-expression recognition is the availability of databases with diversity in age and nationality. The mean age of the participants is around 22 years for most of the databases and 34 years for some databases. Moreover, the diversity in the ethnic backgrounds of the participants is minimal with only one database SAMM to include subjects from 13 nationalities. In general, to make the micro-expression recognition system more robust and suitable for real-time scenarios, the databases should involve more participants with different age groups and varied cultural backgrounds.

Part III

Conclusion

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

6.1 Synopsis

Chapter 3 concentrates on the micro-expression detection phase in the micro-expression analysis. We designed an architecture, called LGAttNet, which is a dual-stream of attention networks concentrating on different regions of the face along with the full face. Attention networks are capable of collecting the local-level and global-level feature maps for identifying the existence of micro-expressions in the image from particular facial regions of interest as well as the full face. This is an image-based supervised micro-expression detection network. LGAttNet outperforms the state-of-the-art results in detecting the micro-expression in the input image. Upon additional experimentation it is observed that LGAttNet is capable of accurately detecting micro-expression frames from a sequence of video frames.

However, in Chapter 4, we proposed two approaches for micro-expression recognition from images and video sequences. Our first approach proposes fine-tuning a CNN framework to recognise the micro-expression from the image. This method demonstrated that, unlike facial macro-expression recognition from images, image-based micro-expression recognition could also yield acceptable accuracy. The second approach discussed in this chapter involves using handcrafted and deep features in an early fusion method framework. Fusing handcrafted LBP-TOP features with CNN deep features enables the network to learn comprehensive features of the input, giving an appropriate prediction of the micro-expression class. Experiments show that both methods can obtain better

and comparable results than traditional features.

Further, Chapter 5 encourages an analysis on the influence of gender features on micro-expressions. In this chapter, we explore the factor that affects the micro-expression recognition accuracy. We have developed a multi-task network, named as GEME, to detect the gender of the input subject and fuse gender features with micro-expression features to enhance micro-expression feature description benefiting the recognition accuracy. From the observations, it can be inferred that gender influences the way humans exhibit micro-expressions and that incorporating gender features in micro-expression recognition process improves the recognition accuracy compared to the state-of-the-art approaches.

6.2 Open and Unsettled Issues

There remain specific unresolved issues.

The datasets involved in the studies have young and healthy student or teacher participants with no criminal background, hindering the database from examining real-life deception, high-stake scenario or medical attention. A database for working with illumination based research requiring diverse illumination settings is hard to find. Furthermore, research correlated with occlusion is essential since partial occlusion often occurs in the real world. Notwithstanding occlusions from sunglasses, facial hairs, hands, scarves, etc., the method requires to be competent enough to recognise micro-expressions. Typically speaking, establishing a database to fulfil everyone's requirements is a challenging task.

We anticipate accepting new and additional micro-expression repositories which are publicly accessible comprising more data samples, data captured in actual deceit circumstances different occlusion situations, illumination, etc.

Among the problems mentioned above, the following are some available concerns:

* All the samples gathered for the databases are only front face view of the participant. Considering this, can micro-expression be identified and recognised from side facial profiles?

* The tests are performed in regulated settings, and the studied approached might not be germane in a natural setting with obstacles such as irregular lighting, noise, etc. Micro-expression datasets with an emphasis on real-world scenarios are required.

* Disparities in the distribution and low sample sizes exacerbate the performance of the algorithms being evaluated. For high-level analysis, a dataset containing a broad amount of samples for balanced classes is vital.

* The existing datasets are focused solely on facial expressions. Body language micro-expressions may assist in boosting precision because it has been revealed that the resemblance in body language can be cross-cultured.

6.3 Future Directions

Although feasible solutions have been proposed in the field of micro-expression spotting and recognition, there might still be a couple of possible research directions that we can suggest. This thesis can contribute to some of the potential future work ideas.

Micro-expression spotting using Attention Nets. In this thesis, we deal with detecting the presence of the micro-expression from static micro-expression frames. As a part of future work, attention networks will be investigated for micro-expression spotting from the videos. Implementing 3D CNN to replace 2D CNN to continue working with Attention networks that concentrate on the Local and Global facial features can be applied to spot micro-expressions. It would be beneficial to identify the onset, apex and offset frames of the micro-expression videos.

Dimensionality reduction and late fusion. Currently, the network processes a very high dimensional feature vector generated after the concatenation of the handcrafted and deep features. This affects the processing speed and accuracy of the model. The inclusion of a dimensionality reduction method to reduce the dimensionality of the feature vector to store specific and relevant features and remove irrelevant information can be considered in the future work. It would also be intriguing to investigate late fusion of handcrafted and deep features for micro-expression recognition results.

Age and Cultural background influence. We incorporated gender features for the micro-expression recognition system. Based on the justification that gender influences micro-expressions, it is logically inferrable that age and ethnicity as a factor may also have some influence on the way the micro-expressions are decoded. Investigating the effects of age and cultural background on micro-expressions would be included as future works.

BIBLIOGRAPHY

[1] M. N. A. AADIT, M. T. MAHIN, AND S. N. JUTHI, *Spontaneous micro-expression recognition using optimal firefly algorithm coupled with iso-flann classification*, in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, 2017, pp. 714–717.

[2] S. AGARWAL AND D. P. MUKHERJEE, *Facial expression recognition through adaptive learning of local motion descriptor*, Multimedia Tools and Applications, 76 (2017), pp. 1073–1099.

[3] J. K. AGGARWAL AND M. S. RYOO, *Human activity analysis: A review*, ACM Computing Surveys (CSUR), 43 (2011), pp. 1–43.

[4] N. AIFANTI, C. PAPACHRISTOU, AND A. DELOPOULOS, *The mug facial expression database*, in 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, IEEE, 2010, pp. 1–4.

[5] B. ALLAERT, I. M. BILASCO, AND C. DJERABA, *Consistent optical flow maps for full and micro facial expression recognition*, 2017.

[6] N. ASLA, J. DE PAÚL, AND A. PÉREZ-ALBÉNIZ, *Emotion recognition in fathers and mothers at high-risk for child physical abuse*, Child abuse & neglect, 35 (2011), pp. 712–721.

[7] A. ASTHANA, S. ZAFEIRIOU, S. CHENG, AND M. PANTIC, *Robust discriminative response map fitting with constrained local models*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3444–3451.

[8] Z. BAO, S. YOU, L. GU, AND Z. YANG, *Single-image facial expression recognition using deep 3d re-centralization*, in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[9] A. BEN-HUR AND J. WESTON, *A user's guide to support vector machines*, in Data mining techniques for the life sciences, Springer, 2010, pp. 223–239.

[10] C. F. BENITEZ-QUIROZ, R. SRINIVASAN, Q. FENG, Y. WANG, AND A. M. MARTINEZ, *Emotionet challenge: Recognition of facial expressions of emotion in the wild*, arXiv preprint arXiv:1703.01210, (2017).

[11] V. BETTADAPURA, *Face expression recognition and analysis: the state of the art*, arXiv preprint arXiv:1203.6722, (2012).

[12] A. BLACKWELL, *A gentle introduction to random forests, ensembles, and performance metrics in a commercial system*, Citizennet, (2012).

[13] D. BORZA, R. DANESCU, R. ITU, AND A. DARABANT, *High-speed video system for micro-expression detection and recognition*, Sensors, 17 (2017), p. 2913.

[14] D. BORZA, R. ITU, AND R. DANESCU, *Real-time micro-expression detection from high speed cameras*, in 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2017, pp. 357–361.

[15] D. BORZA, R. ITU, AND R. DANESCU, *Micro expression detection and recognition from high speed cameras using convolutional neural networks*, in VISIGRAPP, 2018.

[16] N. J. BRITON AND J. A. HALL, *Beliefs about female and male nonverbal communication*, Sex Roles, 32 (1995), pp. 79–90.

[17] R. BUCK, R. E. MILLER, AND W. F. CAUL, *Sex, personality, and physiological variables in the communication of affect via facial expression.*, Journal of personality and social psychology, 30 (1974), p. 587.

[18] V. CAMPOS, B. JOU, AND X. GIRO-I NIETO, *From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction*, Image and Vision Computing, 65 (2017), pp. 15–22.

[19] R. CARUANA, *Multitask learning*, Machine learning, 28 (1997), pp. 41–75.

[20] R. CHAUDHRY, A. RAVICHANDRAN, G. D. HAGER, AND R. VIDAL, *Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions*, 2009 IEEE Conference on Computer Vision and Pattern Recognition, (2009), pp. 1932–1939.

[21] M. CHEN, H. T. MA, J. LI, AND H. WANG, *Emotion recognition using fixed length micro-expressions sequence and weighting method*, in 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR), IEEE, 2016, pp. 427–430.

[22] D. C. CIREŞAN, U. MEIER, J. MASCI, L. M. GAMBARDELLA, AND J. SCHMIDHUBER, *High-performance neural networks for visual object classification*, arXiv preprint arXiv:1102.0183, (2011).

[23] T. F. COOTES, G. J. EDWARDS, AND C. J. TAYLOR, *Active appearance models*, IEEE Transactions on pattern analysis and machine intelligence, 23 (2001), pp. 681–685.

[24] T. F. COOTES, C. J. TAYLOR, D. H. COOPER, AND J. GRAHAM, *Active shape models-their training and application*, Computer vision and image understanding, 61 (1995), pp. 38–59.

[25] Y. CUI, M. JIA, T.-Y. LIN, Y. SONG, AND S. BELONGIE, *Class-balanced loss based on effective number of samples*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

[26] N. DALAL AND B. TRIGGS, *Histograms of oriented gradients for human detection*, in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 886–893.

[27] A. DAVISON, W. MERGHANI, C. LANSLEY, C.-C. NG, AND M. H. YAP, *Objective micro-facial movement detection using facs-based regions and baseline evaluation*, in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 642–649.

[28] A. DAVISON, W. MERGHANI, AND M. YAP, *Objective classes for micro-facial expression recognition*, Journal of Imaging, 4 (2018), p. 119.

[29] A. K. DAVISON, C. LANSLEY, N. COSTEN, K. TAN, AND M. H. YAP, *Samm: A spontaneous micro-facial movement dataset*, IEEE transactions on affective computing, 9 (2016), pp. 116–129.

[30] A. K. DAVISON, M. H. YAP, N. COSTEN, K. TAN, C. LANSLEY, AND D. LEIGHTLEY, *Micro-facial movements: An investigation on spatio-temporal descriptors*, in European conference on computer vision, Springer, 2014, pp. 111–123.

[31] F. De la Torre, W. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, *Intraface*, in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, May 2015, pp. 1–8.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, in 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[33] U. Dimberg and L.-O. Lundquist, *Gender differences in facial reactions to facial expressions*, Biological psychology, 30 (1990), pp. 151–159.

[34] J. Ding, Z. Tian, X. Lyu, Q. Wang, B. Zou, and H. Xie, *Real-time micro-expression detection in unlabeled long videos using optical flow and lstm neural network*, in International Conference on Computer Analysis of Images and Patterns, Springer, 2019, pp. 622–634.

[35] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos, *Aga: Attribute-guided augmentation*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7455–7463.

[36] M. M. Donia, A. A. Youssif, and A. Hashad, *Spontaneous facial expression recognition based on histogram of oriented gradients descriptor.*, Computer and Information Science, 7 (2014), pp. 31–37.

[37] S. Du, Y. Tao, and A. M. Martinez, *Compound facial expressions of emotion*, Proceedings of the National Academy of Sciences, 111 (2014), pp. E1454–E1462.

[38] N. C. Ebner and M. K. Johnson, *Young and older emotional faces: are there age group differences in expression identification and memory?*, Emotion, 9 (2009), p. 329.

[39] G. J. Edwards, C. J. Taylor, and T. F. Cootes, *Interpreting face images using active appearance models*, in Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 300–305.

[40] P. Ekman, *The argument and evidence about universals in facial expressions*, Handbook of social psychophysiology, (1989), pp. 143–164.

[41] P. Ekman, *Microexpression training tool (mett)[computer software]*, University of California, San Franscico, (2002).

[42] P. EKMAN, *Lie catching and microexpressions*, The philosophy of deception, 1 (2009), p. 5.

[43] P. EKMAN AND W. V. FRIESEN, *Nonverbal leakage and clues to deception*, Psychiatry, 32 (1969), pp. 88–106.

[44] P. EKMAN AND W. V. FRIESEN, *Constants across cultures in the face and emotion.*, Journal of personality and social psychology, 17 2 (1971), pp. 124–9.

[45] P. EKMAN AND W. V. FRIESEN, *Unmasking the face : a guide to recognizing emotions from facial clues*, 1975.

[46] ——, *Facial action coding system: Investigator's guide*, Consulting Psychologists Press, 1978.

[47] J. ENDRES AND A. LAIDLAW, *Micro-expression recognition training in medical students: a pilot study*, BMC medical education, 9 (2009), p. 47.

[48] C. EVERS, A. H. FISCHER, AND A. S. MANSTEAD, *Gender and emotion regulation: A social appraisal perspective on anger*, in Emotion regulation and well-being, Springer, 2011, pp. 211–222.

[49] C. FABIAN BENITEZ-QUIROZ, R. SRINIVASAN, AND A. M. MARTINEZ, *Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5562–5570.

[50] M. FÖLSTER, U. HESS, AND K. WERHEID, *Facial age affects emotional expression decoding*, Frontiers in psychology, 5 (2014), p. 30.

[51] M. G. FRANK, C. J. MACCARIO, AND V. GOVINDARAJU, *Behavior and security. protecting airline passengers in the age of terrorism*, 2009.

[52] M. G. FRANK AND E. SVETIEVA, *Microexpressions and deception*, in Understanding facial expressions in communication, Springer, 2015, pp. 227–242.

[53] E. FRIESEN AND P. EKMAN, *Facial action coding system: a technique for the measurement of facial movement*, Consulting Psychologists Press, 1978.

[54] L. GAN, Y. ZOU, AND C. ZHANG, *Discriminative feature learning using two-stage training strategy for facial expression recognition*, in International Conference on Artificial Neural Networks, Springer, 2019, pp. 397–408.

[55] Y. GAN, S.-T. LIONG, W.-C. YAU, Y.-C. HUANG, AND L.-K. TAN, *Off-apexnet on micro-expression recognition system*, Signal Processing: Image Communication, 74 (2019), pp. 129–139.

[56] D. GERONIMO, A. M. LOPEZ, A. D. SAPPA, AND T. GRAF, *Survey of pedestrian detection for advanced driver assistance systems*, IEEE transactions on pattern analysis and machine intelligence, 32 (2009), pp. 1239–1258.

[57] R. GROSS, I. MATTHEWS, J. COHN, T. KANADE, AND S. BAKER, *Multi-pie*, Image and Vision Computing, 28 (2010), pp. 807–813.

[58] C. GUO, J. LIANG, G. ZHAN, Z. LIU, M. PIETIKÄINEN, AND L. LIU, *Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition*, IEEE Access, 7 (2019), pp. 174517–174530.

[59] Y. GUO, Y. TIAN, X. GAO, AND X. ZHANG, *Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method*, in 2014 international joint conference on neural networks (IJCNN), IEEE, 2014, pp. 3473–3479.

[60] Y. GUO, C. XUE, Y. WANG, AND M. YU, *Micro-expression recognition based on cbp-top feature with elm*, Optik, 126 (2015), pp. 4446–4451.

[61] Z. GUO, L. ZHANG, AND D. ZHANG, *A completed modeling of local binary pattern operator for texture classification*, IEEE transactions on image processing, 19 (2010), pp. 1657–1663.

[62] E. A. HAGGARD AND K. S. ISAACS, *Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy*, in Methods of research in psychotherapy, Springer, 1966, pp. 154–165.

[63] J. A. HALL AND S. D. GUNNERY, *Gender differences in nonverbal communication.*, (2013).

[64] S. HAPPY AND A. ROUTRAY, *Fuzzy histogram of optical flow orientations for micro-expression recognition*, IEEE Transactions on Affective Computing, (2017).

[65] J. HE, J.-F. HU, X. LU, AND W.-S. ZHENG, *Multi-task mid-level feature learning for micro-expression recognition*, Pattern Recognition, 66 (2017), pp. 44–52.

[66] U. HESS, R. ADAMS JR, AND R. KLECK, *Who may frown and who should smile? dominance, affiliation, and the display of happiness and anger*, Cognition & Emotion, 19 (2005), pp. 515–536.

[67] X. HONG, G. ZHAO, S. ZAFEIRIOU, M. PANTIC, AND M. PIETIKÄINEN, *Capturing correlations of local features for image representation*, Neurocomputing, 184 (2016), pp. 99–106.

[68] B. K. HORN AND B. G. SCHUNCK, *Determining optical flow*, in Techniques and Applications of Image Understanding, vol. 281, International Society for Optics and Photonics, 1981, pp. 319–331.

[69] C. HOUSE AND R. MEYER, *Preprocessing and descriptor features for facial micro-expression recognition*, IEEE transaction, (2015).

[70] C. HU, J. CHEN, X. ZUO, H. ZOU, X. W. DENG, YU-CHENG, AND SHU, *Gender-specific multi-task micro-expression recognition using pyramid cgbp-top feature*, 2019.

[71] C. HU, D. JIANG, H. ZOU, X. ZUO, AND Y. SHU, *Multi-task micro-expression recognition combining deep and handcrafted features*, in 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 946–951.

[72] G. HU, L. LIU, Y. YUAN, Z. YU, Y. HUA, Z. ZHANG, F. SHEN, L. SHAO, T. HOSPEDALES, N. ROBERTSON, ET AL., *Deep multi-task learning to recognise subtle facial expressions of mental states*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 103–119.

[73] G. HU, X. PENG, Y. YANG, T. M. HOSPEDALES, AND J. VERBEEK, *Frankenstein: Learning deep face representations using small data*, IEEE Transactions on Image Processing, 27 (2017), pp. 293–303.

[74] G. B. HUANG, M. MATTAR, T. BERG, AND E. LEARNED-MILLER, *Labeled faces in the wild: A database forstudying face recognition in unconstrained environments*, in Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.

[75] X. HUANG, S. WANG, G. ZHAO, AND M. PIETIKÄINEN, *Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection*,

2015 IEEE International Conference on Computer Vision Workshop (ICCVW), (2015), pp. 1–9.

[76] X. HUANG AND G. ZHAO, *Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern*, in 2017 International Conference on the Frontiers and Advances in Data Science (FADS), IEEE, 2017, pp. 159–164.

[77] X. HUANG, G. ZHAO, X. HONG, W. ZHENG, AND M. PIETIKÄINEN, *Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns*, Neurocomputing, 175 (2016), pp. 564–578.

[78] C. M. HURLEY, *The effects of motivation and training format on the ability to detect hidden emotions*, State University of New York at Buffalo, 2010.

[79] N. T. ISSA, V. STATHIAS, S. SCHÜRER, AND S. DAKSHANAMURTHY, *Machine and deep learning approaches for cancer drug repurposing*, in Seminars in Cancer Biology, Elsevier, 2020.

[80] R. JACK, O. GARROD, H. YU, R. CALDARA, AND P. SCHYNS, *Dynamic cultural representations of facial expressions of emotion are not universal*, Journal of Vision, 11 (2011), p. 563.

[81] B. C. JC ET AL., *A tutorial on support vector machines for pattern recognition*, Data mining and knowledge discovery, 2 (1998), pp. 121–167.

[82] Y. JIA, E. SHELHAMER, J. DONAHUE, S. KARAYEV, J. LONG, R. GIRSHICK, S. GUADARRAMA, AND T. DARRELL, *Caffe: Convolutional architecture for fast feature embedding*, in Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[83] Y. JIAO, Y. NIU, Y. ZHANG, F. LI, C. ZOU, AND G. SHI, *Facial attention based convolutional neural network for 2d+ 3d facial expression recognition*, in 2019 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2019, pp. 1–4.

[84] A. KARPATHY ET AL., *Cs231n convolutional neural networks for visual recognition*, Neural networks, 1 (2016).

[85] S. A. KHAN, A. HUSSAIN, AND M. USMAN, *Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features*, Multimedia Tools and Applications, 77 (2018), pp. 1133–1165.

[86] H.-Q. KHOR, J. SEE, S.-T. LIONG, R. C. PHAN, AND W. LIN, *Dual-stream shallow networks for facial micro-expression recognition*, in 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 36–40.

[87] H.-Q. KHOR, J. SEE, R. C. W. PHAN, AND W. LIN, *Enriched long-term recurrent convolutional network for facial micro-expression recognition*, in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 667–674.

[88] D. H. KIM, W. J. BADDAR, AND Y. M. RO, *Micro-expression recognition with expression-state constrained spatio-temporal feature representations*, in Proceedings of the 24th ACM international conference on Multimedia, ACM, 2016, pp. 382–386.

[89] I. KIM, J. H. SHIM, AND J. YANG, *Face detection*.

[90] I. KOKKINOS, *Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6129–6138.

[91] D. KOLLIAS, A. SCHULC, E. HAJIYEV, AND S. ZAFEIRIOU, *Analysing affective behavior in the first abaw 2020 competition*, arXiv preprint arXiv:2001.11409, (2020).

[92] D. KOLLIAS, V. SHARMANSKA, AND S. ZAFEIRIOU, *Face behavior\a la carte: Expressions, affect and action units in a single network*, arXiv preprint arXiv:1910.11111, (2019).

[93] D. KOLLIAS, P. TZIRAKIS, M. A. NICOLAOU, A. PAPAIOANNOU, G. ZHAO, B. SCHULLER, I. KOTSIA, AND S. ZAFEIRIOU, *Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond*, International Journal of Computer Vision, 127 (2019), pp. 907–929.

[94] M. E. KRET AND B. DE GELDER, *A review on sex differences in processing emotional signals*, Neuropsychologia, 50 (2012), pp. 1211–1221.

[95] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.

[96] E. KRUMHUBER, A. S. MANSTEAD, AND A. KAPPAS, *Temporal aspects of facial displays in person and expression perception: The effects of smile dynamics, head-tilt, and gender*, Journal of Nonverbal Behavior, 31 (2007), pp. 39–56.

[97] A. KUMAR, A. KAUR, AND M. KUMAR, *Face detection techniques: a review*, Artificial Intelligence Review, 52 (2019), pp. 927–948.

[98] Z. LAI, R. CHEN, J. JIA, AND Y. QIAN, *Real-time micro-expression recognition based on resnet and atrous convolutions*, Journal of Ambient Intelligence and Humanized Computing, (2020), pp. 1–12.

[99] S. M. LAJEVARDI AND H. R. WU, *Facial expression recognition in perceptual color space*, IEEE transactions on image processing, 21 (2012), pp. 3721–3733.

[100] A. C. LE NGO, A. JOHNSTON, R. C.-W. PHAN, AND J. SEE, *Micro-expression motion magnification: Global lagrangian vs. local eulerian approaches*, in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 650–656.

[101] A. C. LE NGO, Y.-H. OH, R. C.-W. PHAN, AND J. SEE, *Eulerian emotion magnification for subtle expression recognition*, in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 1243–1247.

[102] Y. LECUN, L. BOTTOU, Y. BENGIO, P. HAFFNER, ET AL., *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[103] K. LEE, D. SATO, S. ASAKAWA, H. KACORRI, AND C. ASAKAWA, *Pedestrian detection with wearable cameras for the blind: A two-way perspective*, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–12.

[104] C. LI, R. WANG, J. LI, AND L. FEI, *Face detection based on yolov3*, in Recent Trends in Intelligent Computing, Communication and Devices, Springer, 2020, pp. 277–284.

[105] J. LI, C. SOLADIE, AND R. SEGUIER, *Ltp-ml: micro-expression detection by recognition of local temporal pattern of facial movements*, in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 634–641.

[106] J. LI, C. SOLADIE, R. SEGUIER, J. LI, C. SOLADIE, R. SEGUIER, S.-J. WANG, M. H. YAP, R. WEBER, J. LI, ET AL., *A survey on databases for facial micro-expression analysis*, 2019.

[107] J. LI, C. SOLADIE, R. SEGUIER, S.-J. WANG, AND M. H. YAP, *Spotting micro-expressions on long videos sequences*, in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[108] Q. LI, J. YU, T. KURIHARA, AND S. ZHAN, *Micro-expression analysis by fusing deep convolutional neural network and optical flow*, in 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), IEEE, 2018, pp. 265–270.

[109] S. LI AND W. DENG, *Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition*, IEEE Transactions on Image Processing, 28 (2018), pp. 356–370.

[110] W. LI, M. LI, Z. SU, AND Z. ZHU, *A deep-learning approach to facial expression recognition with candid images*, in 2015 14th IAPR International Conference on Machine Vision Applications (MVA), IEEE, 2015, pp. 279–282.

[111] X. LI, X. HONG, A. MOILANEN, X. HUANG, T. PFISTER, G. ZHAO, AND M. PIETIKÄINEN, *Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods*, IEEE transactions on affective computing, 9 (2017), pp. 563–577.

[112] X. LI, T. PFISTER, X. HUANG, G. ZHAO, AND M. PIETIKÄINEN, *A spontaneous micro-expression database: Inducement, collection and baseline*, in 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

[113] X. LI, J. YU, AND S. ZHAN, *Spontaneous facial micro-expression detection based on deep learning*, in 2016 IEEE 13th International Conference on Signal Processing (ICSP), IEEE, 2016, pp. 1130–1134.

[114] Y. LI, X. HUANG, AND G. ZHAO, *Micro-expression action unit detection with spatio-temporal adaptive pooling*, ArXiv, abs/1907.05023 (2019).

[115] L. LIEBEL AND M. KÖRNER, *Auxiliary tasks in multi-task learning*, arXiv preprint arXiv:1805.06334, (2018).

[116] T.-Y. LIN, P. GOYAL, R. GIRSHICK, K. HE, AND P. DOLLÁR, *Focal loss for dense object detection*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[117] H. LING, J. WU, J. HUANG, J. CHEN, AND P. LI, *Attention-based convolutional neural network for deep face recognition*, Multimedia Tools and Applications, 79 (2020), pp. 5595–5616.

[118] S.-T. LIONG, Y. GAN, J. SEE, H.-Q. KHOR, AND Y.-C. HUANG, *Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition*, in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[119] S.-T. LIONG, Y. GAN, D. ZHENG, H.-X. XUA, H.-Z. ZHANG, R.-K. LYU, K.-H. LIU, ET AL., *Evaluation of the spatio-temporal features and gan for micro-expression recognition system*, arXiv preprint arXiv:1904.01748, (2019).

[120] S.-T. LIONG, J. SEE, R. C.-W. PHAN, Y.-H. OH, A. C. LE NGO, K. WONG, AND S.-W. TAN, *Spontaneous subtle expression detection and recognition based on facial strain*, Signal Processing: Image Communication, 47 (2016), pp. 170–182.

[121] S.-T. LIONG, J. SEE, K. WONG, A. C. LE NGO, Y.-H. OH, AND R. PHAN, *Automatic apex frame spotting in micro-expression database*, in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 665–669.

[122] S.-T. LIONG, J. SEE, K. WONG, AND R. C.-W. PHAN, *Less is more: Micro-expression recognition from video using apex frame*, Signal Processing: Image Communication, 62 (2018), pp. 82–92.

[123] S.-T. LIONG AND K. WONG, *Micro-expression recognition using apex frame with phase information*, in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 534–537.

[124] A.-A. LIU, N. XU, W.-Z. NIE, Y.-T. SU, AND Y.-D. ZHANG, *Multi-domain and multi-task learning for human action recognition*, IEEE Transactions on Image Processing, 28 (2018), pp. 853–867.

[125] W. LIU, L. CHEN, AND Y. CHEN, *Age classification using convolutional neural networks with the multi-class focal loss*, in IOP Conference Series: Materials Science and Engineering, vol. 428, IOP Publishing, 2018, p. 012043.

[126] Y. LIU, H. DU, L. ZHENG, AND T. GEDEON, *A neural micro-expression recognizer*, in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–4.

[127] Y.-J. LIU, J.-K. ZHANG, W.-J. YAN, S.-J. WANG, G. ZHAO, AND X. FU, *A main directional mean optical flow feature for spontaneous micro-expression recognition*, IEEE Transactions on Affective Computing, 7 (2015), pp. 299–310.

[128] Z. LIU, J. DONG, C. ZHANG, L. WANG, AND J. DANG, *Relation modeling with graph convolutional networks for facial action unit detection*, in International Conference on Multimedia Modeling, Springer, 2020, pp. 489–501.

[129] C. LU, J. SHI, AND J. JIA, *Online robust dictionary learning*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 415–422.

[130] H. LU, K. KPALMA, AND J. RONSIN, *Micro-expression detection using integral projections*, (2017).

[131] Z. LU, Z. LUO, H. ZHENG, J. CHEN, AND W. LI, *A delaunay-based temporal coding model for micro-expression recognition*, in Asian conference on computer vision, Springer, 2014, pp. 698–711.

[132] B. D. LUCAS, *Generalized image matching by the method of differences.*, (1986).

[133] P. LUCEY, J. F. COHN, T. KANADE, J. SARAGIH, Z. AMBADAR, AND I. MATTHEWS, *The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression*, in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, 2010, pp. 94–101.

[134] X. LUN, L. XIN, Y. XIUJUN, AND W. ZHILIANG, *Cognitive regulation and emotion modeling for micro-expression*, Int J Control Autom, 9 (2016), pp. 361–372.

[135] M. J. LYONS, S. AKAMATSU, M. KAMACHI, J. GYOBA, AND J. BUDYNEK, *The japanese female facial expression (jaffe) database*, in Proceedings of third international conference on automatic face and gesture recognition, 1998, pp. 14–16.

[136] P. D. MARRERO FERNANDEZ, F. A. GUERRERO PENA, T. REN, AND A. CUNHA, *Feratt: Facial expression recognition with attention net*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[137] P. J. MARSH, M. J. GREEN, T. A. RUSSELL, J. MCGUIRE, A. HARRIS, AND M. COLTHEART, *Remediation of facial emotion recognition in schizophrenia: Functional predictors, generalizability, and durability*, American Journal of Psychiatric Rehabilitation, 13 (2010), pp. 143–170.

[138] D. MATSUMOTO AND H. C. HWANG, *Microexpressions differentiate truths from lies about future malicious intent*, Frontiers in psychology, 9 (2018), p. 2545.

[139] D. MATSUMOTO AND H. S. HWANG, *Evidence for training the ability to read microexpressions of emotion*, Motivation and emotion, 35 (2011), pp. 181–191.

[140] D. MATSUMOTO, J. LEROUX, C. WILSON-COHN, J. RAROQUE, K. KOOKEN, P. EKMAN, N. YRIZARRY, S. LOEWINGER, H. UCHIDA, A. 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 (2000), pp. 179–209.

[141] V. MAYYA, R. M. PAI, AND M. M. PAI, *Combining temporal interpolation and dcnn for faster recognition of micro-expressions in video sequences*, in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2016, pp. 699–703.

[142] D. MCDUFF, E. KODRA, R. EL KALIOUBY, AND M. LAFRANCE, *A large-scale analysis of sex differences in facial expressions*, PloS one, 12 (2017).

[143] S. MINAEE AND A. ABDOLRASHIDI, *Deep-emotion: Facial expression recognition using attentional convolutional network*, arXiv preprint arXiv:1902.01019, (2019).

[144] A. MISHRA, *Metrics to evaluate your machine learning algorithm*, Towards Data Science. URL: https://towardsdatascience. com/metrics-toevaluate-your-machine-learning-algorithm-f10ba6e38234. Accessed, 12 (2018), p. 2018.

[145] I. MISRA, A. SHRIVASTAVA, A. GUPTA, AND M. HEBERT, *Cross-stitch networks for multi-task learning*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3994–4003.

[146] A. MOLLAHOSSEINI, B. HASANI, AND M. H. MAHOOR, *Affectnet: A database for facial expression, valence, and arousal computing in the wild*, IEEE Transactions on Affective Computing, 10 (2017), pp. 18–31.

[147] N. MUNA, U. D. ROSIANI, E. M. YUNIAMO, AND M. H. PUMOMO, *Subpixel subtle motion estimation of micro-expressions multiclass classification*, in 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), IEEE, 2017, pp. 325–330.

[148] S. NAG, A. K. BHUNIA, A. KONWER, AND P. P. ROY, *Facial micro-expression spotting and recognition using time contrasted feature with visual memory*, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2022–2026.

[149] S. NERELLA, A. BIHORAC, P. TIGHE, AND P. RASHIDI, *Facial action unit detection on icu data for pain assessment*, arXiv preprint arXiv:2005.02121, (2020).

[150] H.-W. NG, V. D. NGUYEN, V. VONIKAKIS, AND S. WINKLER, *Deep learning for emotion recognition on small datasets using transfer learning*, in Proceedings of the 2015 ACM on international conference on multimodal interaction, ACM, 2015, pp. 443–449.

[151] P. M. NIEDENTHAL AND M. BRAUER, *Social functionality of human emotion*, Annual review of psychology, 63 (2012), pp. 259–285.

[152] Y.-H. OH, J. SEE, A. C. LE NGO, R. C.-W. PHAN, AND V. M. BASKARAN, *A survey of automatic facial micro-expression analysis: Databases, methods, and challenges*, Frontiers in psychology, 9 (2018), p. 1128.

[153] T. OJALA, M. PIETIKÄINEN, AND D. HARWOOD, *A comparative study of texture measures with classification based on featured distributions*, Pattern recognition, 29 (1996), pp. 51–59.

[154] T. OJALA, M. PIETIKÄINEN, AND T. MÄENPÄÄ, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Transactions on Pattern Analysis & Machine Intelligence, (2002), pp. 971–987.

[155] Y. OUYANG AND N. SANG, *A facial expression recognition method by fusing multiple sparse representation based classifiers*, in International Symposium on Neural Networks, Springer, 2013, pp. 479–488.

[156] S. Y. PARK, S. H. LEE, AND Y. M. RO, *Subtle facial expression recognition using adaptive magnification of discriminative facial motion*, in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 911–914.

[157] O. M. PARKHI, A. VEDALDI, A. ZISSERMAN, ET AL., *Deep face recognition.*, in bmvc, vol. 1, 2015, p. 6.

[158] A. PASZKE, S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA, ET AL., *Pytorch: An imperative style, high-performance deep learning library*, in Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[159] D. PATEL, X. HONG, AND G. ZHAO, *Selective deep features for micro-expression recognition*, in 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2258–2263.

[160] D. PATEL, G. ZHAO, AND M. PIETIKÄINEN, *Spatiotemporal integration of optical flow vectors for micro-expression detection*, in International conference on advanced concepts for intelligent vision systems, Springer, 2015, pp. 369–380.

[161] M. PENG, C. WANG, T. CHEN, G. LIU, AND X. FU, *Dual temporal scale convolutional neural network for micro-expression recognition*, Frontiers in psychology, 8 (2017), p. 1745.

[162] Y. PENG, D. MENG, Z. XU, C. GAO, Y. YANG, AND B. ZHANG, *Decomposable nonlocal tensor dictionary learning for multispectral image denoising*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2949–2956.

[163] T. PFISTER, X. LI, G. ZHAO, AND M. PIETIKÄINEN, *Differentiating spontaneous from posed facial expressions within a generic facial expression recognition*

framework, in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 868–875.

[164] T. PFISTER, X. LI, G. ZHAO, AND M. PIETIKÄINEN, *Recognising spontaneous facial micro-expressions*, IEEE, 2011, pp. 1449–1456.

[165] M. PIETIKÄINEN, A. HADID, G. ZHAO, AND T. AHONEN, *Computer vision using local binary patterns*, vol. 40, Springer Science & Business Media, 2011.

[166] S. POLIKOVSKY, Y. KAMEDA, AND Y. OHTA, *Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor*, (2009).

[167] F. QU, S.-J. WANG, W.-J. YAN, H. LI, S. WU, AND X. FU, *Cas(me)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition*, IEEE Transactions on Affective Computing, 9 (2017), pp. 424–436.

[168] F. QU, S. YAN, J. LIANG, AND J. WANG, *Effect of short-term micro-expression training on the micro-expression recognition performance of preschool children*, in International Conference on Cognitive Systems and Signal Processing, Springer, 2018, pp. 54–62.

[169] S. RAGHAV AND H. KIRSTIE, *(can't) lie to me : Using micro expressions for user authentication*, in Symposium on Usable Privacy and Security (SOUPS), 2014.

[170] R. RANJAN, A. BANSAL, J. ZHENG, H. XU, J. GLEASON, B. LU, A. NANDURI, J.-C. CHEN, C. D. CASTILLO, AND R. CHELLAPPA, *A fast and accurate system for face detection, identification, and verification*, IEEE Transactions on Biometrics, Behavior, and Identity Science, 1 (2019), pp. 82–96.

[171] R. RANJAN, V. M. PATEL, AND R. CHELLAPPA, *Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41 (2017), pp. 121–135.

[172] R. RANJAN, S. SANKARANARAYANAN, C. D. CASTILLO, AND R. CHELLAPPA, *An all-in-one convolutional neural network for face analysis*, in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 17–24.

[173] S. P. T. REDDY, S. T. KARRI, S. R. DUBEY, AND S. MUKHERJEE, *Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks*, arXiv preprint arXiv:1904.01390, (2019).

[174] A. RUIZ, J. VAN DE WEIJER, AND X. BINEFA, *From emotions to action units with hidden and semi-hidden-task learning*, in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3703–3711.

[175] T. A. RUSSELL, E. CHU, AND M. L. PHILLIPS, *A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool*, British journal of clinical psychology, 45 (2006), pp. 579–583.

[176] J. SEE, M. H. YAP, J. LI, X. HONG, AND S.-J. WANG, *Megc 2019–the second facial micro-expressions grand challenge*, in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[177] R. R. SELVARAJU, M. COGSWELL, A. DAS, R. VEDANTAM, D. PARIKH, AND D. BATRA, *Grad-cam: Visual explanations from deep networks via gradient-based localization*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[178] H. SHAHAR AND H. HEL-OR, *Micro expression classification using facial color and deep learning methods*, in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[179] B. SHAO, L. DOUCET, AND D. R. CARUSO, *Universality versus cultural specificity of three emotion domains: Some evidence based on the cascading model of emotional intelligence*, Journal of Cross-Cultural Psychology, 46 (2015), pp. 229–251.

[180] X.-B. SHEN, Q. WU, AND X.-L. FU, *Effects of the duration of expressions on the recognition of microexpressions*, Journal of Zhejiang University Science B, 13 (2012), pp. 221–230.

[181] M. SHREVE, S. GODAVARTHY, D. GOLDGOF, AND S. SARKAR, *Macro-and micro-expression spotting in long videos using spatio-temporal strain*, in Face and Gesture 2011, IEEE, 2011, pp. 51–56.

[182] P. Y. Simard, D. Steinkraus, and J. C. Platt, *Best practices for convolutional neural networks applied to visual document analysis*, 2003, pp. 958–963.

[183] P. Y. Simard, D. Steinkraus, J. C. Platt, et al., *Best practices for convolutional neural networks applied to visual document analysis.*, in Icdar, vol. 3, 2003.

[184] A. J. Smola and B. Schölkopf, *A tutorial on support vector regression*, Statistics and computing, 14 (2004), pp. 199–222.

[185] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, *Recognizing spontaneous micro-expression using a three-stream convolutional neural network*, IEEE Access, 7 (2019), pp. 184537–184551.

[186] Y. Song, L.-P. Morency, and R. Davis, *Learning a sparse codebook of facial and body microexpressions for emotion recognition*, in Proceedings of the 15th ACM on International conference on multimodal interaction, ACM, 2013, pp. 237–244.

[187] B. Sun, S. Cao, D. Li, J. He, and L. Yu, *Dynamic micro-expression recognition using knowledge distillation*, IEEE Transactions on Affective Computing, (2020).

[188] W. Sun, H. Zhao, and Z. Jin, *A visual attention based roi detection method for facial expression recognition*, Neurocomputing, 296 (2018), pp. 12–22.

[189] X. Sun, P. Wu, and S. C. Hoi, *Face detection using deep learning: An improved faster rcnn approach*, Neurocomputing, 299 (2018), pp. 42–50.

[190] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[191] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, *A survey: facial micro-expression recognition*, Multimedia Tools and Applications, 77 (2018), pp. 19301–19325.

[192] M. A. Takalkar and M. Xu, *Image based facial micro-expression recognition using deep learning on small datasets*, in 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2017, pp. 1–7.

[193] M. A. TAKALKAR, H. ZHANG, AND M. XU, *Improving micro-expression recognition accuracy using twofold feature extraction*, in International Conference on Multimedia Modeling, Springer, 2019, pp. 652–664.

[194] P. K. C. TAY, *The adaptive value associated with expressing and perceiving angry-male and happy-female faces*, Frontiers in psychology, 6 (2015), p. 851.

[195] S. THUSEETHAN, S. RAJASEGARAR, AND J. YEARWOOD, *Deep hybrid spatiotemporal networks for continuous pain intensity estimation*, in International Conference on Neural Information Processing, Springer, 2019, pp. 449–461.

[196] ——, *Detecting micro-expression intensity changes from videos based on hybrid deep cnn*, in Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2019, pp. 387–399.

[197] ——, *Detecting micro-expression intensity changes from videos based on hybrid deep cnn*, in Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2019, pp. 387–399.

[198] ——, *Emotion intensity estimation from video frames using deep hybrid convolutional neural networks*, in 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–10.

[199] M. VALSTAR AND M. PANTIC, *Induced disgust, happiness and surprise: an addition to the mmi facial expression database*, in Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Paris, France, 2010, p. 65.

[200] N. VAN QUANG, J. CHUN, AND T. TOKUYAMA, *Capsulenet for micro-expression recognition*, in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–7.

[201] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in Advances in neural information processing systems, 2017, pp. 5998–6008.

[202] M. VERMA, S. K. VIPPARTHI, G. SINGH, AND S. MURALA, *Learnet: Dynamic imaging network for micro expression recognition*, IEEE Transactions on Image Processing, 29 (2019), pp. 1618–1627.

[203] P. VIOLA AND M. J. JONES, *Robust real-time face detection*, International journal of computer vision, 57 (2004), pp. 137–154.

[204] H. G. WALLBOTT, *Big girls don't frown, big boys don't cry,Äîgender differences of professional actors in communicating emotion via facial expression*, Journal of Nonverbal Behavior, 12 (1988), pp. 98–106.

[205] C. WANG, M. PENG, T. BI, AND T. CHEN, *Micro-attention for micro-expression recognition*, arXiv preprint arXiv:1811.02360, (2018).

[206] G. WANG, W. WANG, J. WANG, AND Y. BU, *Better deep visual attention with reinforcement learning in action recognition*, in 2017 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2017, pp. 1–4.

[207] L. WANG, X. YU, T. BOURLAI, AND D. N. METAXAS, *A coupled encoder–decoder network for joint face detection and landmark localization*, Image and Vision Computing, 87 (2019), pp. 37–46.

[208] S.-J. WANG, H.-L. CHEN, W.-J. YAN, Y.-H. CHEN, AND X. FU, *Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine*, Neural processing letters, 39 (2014), pp. 25–43.

[209] S.-J. WANG, B.-J. LI, Y.-J. LIU, W.-J. YAN, X. OU, X. HUANG, F. XU, AND X. FU, *Micro-expression recognition with small sample size by transferring long-term convolutional neural network*, Neurocomputing, 312 (2018), pp. 251–262.

[210] S.-J. WANG, W.-J. YAN, X. LI, G. ZHAO, AND X. FU, *Micro-expression recognition using dynamic textures on tensor independent color space*, in 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 4678–4683.

[211] S.-J. WANG, W.-J. YAN, X. LI, G. ZHAO, C.-G. ZHOU, X. FU, M. YANG, AND J. TAO, *Micro-expression recognition using color spaces*, IEEE Transactions on Image Processing, 24 (2015), pp. 6034–6047.

[212] S.-J. WANG, J. YANG, M.-F. SUN, X.-J. PENG, M.-M. SUN, AND C.-G. ZHOU, *Sparse tensor discriminant color space for face verification*, IEEE Transactions on Neural Networks and Learning Systems, 23 (2012), pp. 876–888.

[213] S.-J. WANG, J. YANG, N. ZHANG, AND C.-G. ZHOU, *Tensor discriminant color space for face recognition*, IEEE Transactions on Image Processing, 20 (2011), pp. 2490–2501.

[214] X. WANG, Y. WU, L. ZHU, AND Y. YANG, *Symbiotic attention with privileged information for egocentric action recognition*, arXiv preprint arXiv:2002.03137, (2020).

[215] Y. WANG, J. SEE, Y.-H. OH, R. C.-W. PHAN, Y. RAHULAMATHAVAN, H.-C. LING, S.-W. TAN, AND X. LI, *Effective recognition of facial micro-expressions with video motion magnification*, Multimedia Tools and Applications, 76 (2017), pp. 21665–21690.

[216] Y. WANG, J. SEE, R. C.-W. PHAN, AND Y.-H. OH, *Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition*, in Asian conference on computer vision, Springer, 2014, pp. 525–537.

[217] Y. WANG, J. SEE, R. C.-W. PHAN, Y.-H. OH, AND P. J. HILLS, *Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition*, PloS one, 10 (2015).

[218] G. WARREN, E. SCHERTLER, AND P. BULL, *Detecting deception from emotional and unemotional cues*, Journal of Nonverbal Behavior, 33 (2009), pp. 59–69.

[219] S. WEINBERGER, *Intent to deceive? can the science of deception detection help to catch terrorists? sharon weinberger takes a close look at the evidence for it*, Nature, 465 (2010), pp. 412–416.

[220] K. WEZOWSKI AND P. WEZOWSKI, *The micro expressions book for business*, New Vision, Antwerp, 127 (2012).

[221] S. C. WIDEN, J. A. RUSSELL, AND A. BROOKS, *Anger and disgust: Discrete or overlapping categories*, in 2004 APS Annual Convention, Boston College, Chicago, IL, 2004.

[222] A. WIERZBICKA, *Human emotions: Universal or culture-specific?*, American anthropologist, 88 (1986), pp. 584–594.

[223] L. WOLF, T. HASSNER, AND I. MAOZ, *Face recognition in unconstrained videos with matched background similarity*, IEEE, 2011.

[224] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, *Cbam: Convolutional block attention module*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[225] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, *Eulerian video magnification for revealing subtle changes in the world*, ACM transactions on graphics (TOG), 31 (2012), pp. 1–8.

[226] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, *Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions*, IEEE Transactions on Multimedia, (2019).

[227] J. Xing, J. Gao, B. Li, W. Hu, and S. Yan, *Robust object tracking with online multi-lifespan dictionary learning*, in Proceedings of the IEEE International conference on computer vision, 2013, pp. 665–672.

[228] F. Xu, J. Zhang, and J. Z. Wang, *Microexpression identification and categorization using a facial dynamics map*, IEEE Transactions on Affective Computing, 8 (2017), pp. 254–267.

[229] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, *Casme ii: An improved spontaneous micro-expression database and the baseline evaluation*, PloS one, 9 (2014), p. e86041.

[230] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, *How fast are the leaked facial expressions: The duration of micro-expressions*, Journal of Nonverbal Behavior, 37 (2013), pp. 217–230.

[231] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, *Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces*, in 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, 2013, pp. 1–7.

[232] B. Yang, J. Cheng, Y. Yang, B. Zhang, and J. Li, *Merta: micro-expression recognition with ternary attentions*, Multimedia Tools and Applications, (2019), pp. 1–16.

[233] J. Yang, W. Liu, J. Yuan, and T. Mei, *Hierarchical soft quantization for skeleton-based human action recognition*, IEEE Transactions on Multimedia, (2020).

[234] J. YANG, L. ZHANG, Y. XU, AND J.-Y. YANG, *Beyond sparsity: The role of l1-optimizer in pattern classification*, Pattern Recognition, 45 (2012), pp. 1104–1118.

[235] M. YANG, D. DAI, L. SHEN, AND L. VAN GOOL, *Latent dictionary learning for sparse representation based classification*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4138–4145.

[236] M. YANG, L. VAN GOOL, AND L. ZHANG, *Sparse variation dictionary learning for face recognition with a single training sample per person*, in Proceedings of the IEEE international conference on computer vision, 2013, pp. 689–696.

[237] S. ZAFEIRIOU, D. KOLLIAS, M. A. NICOLAOU, A. PAPAIOANNOU, G. ZHAO, AND I. KOTSIA, *Aff-wild: Valence and arousal'in-the-wild'challenge*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 34–41.

[238] E. ZAREZADEH AND M. REZAEIAN, *Micro expression recognition using the eulerian video magnification method*, BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 7 (2016), pp. 43–54.

[239] M. ZHANG, Q. FU, Y.-H. CHEN, AND X. FU, *Emotional context influences micro-expression recognition*, PloS one, 9 (2014), p. e95018.

[240] S. ZHANG, B. FENG, Z. CHEN, AND X. HUANG, *Micro-expression recognition by aggregating local spatio-temporal patterns*, in International Conference on Multimedia Modeling, Springer, 2017, pp. 638–648.

[241] X. ZHANG, L. CHEN, Z. ZHONG, H. SUI, AND X. SHEN, *The effects of the micro-expression training on empathy in patients with schizophrenia*, in International Conference on Man-Machine-Environment System Engineering, Springer, 2017, pp. 189–194.

[242] Z. ZHANG, T. CHEN, H. MENG, G. LIU, AND X. FU, *Smeconvnet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos*, IEEE Access, 6 (2018), pp. 71143–71151.

[243] G. ZHAO AND M. PIETIKAINEN, *Dynamic texture recognition using local binary patterns with an application to facial expressions*, IEEE transactions on pattern analysis and machine intelligence, 29 (2007), pp. 915–928.

[244] Y. ZHAO AND J. XU, *An improved micro-expression recognition method based on necessary morphological patches*, Symmetry, 11 (2019), p. 497.

[245] H. ZHENG, X. GENG, AND Z. YANG, *A relaxed k-svd algorithm for spontaneous micro-expression recognition*, in Pacific Rim International Conference on Artificial Intelligence, Springer, 2016, pp. 692–699.

[246] R. ZHI, H. XU, M. WAN, AND T. LI, *Combining 3d convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition*, IEICE Transactions on Information and Systems, 102 (2019), pp. 1054–1064.

[247] L. ZHOU, Q. MAO, AND L. XUE, *Dual-inception network for cross-database micro-expression recognition*, in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[248] Y. ZONG, X. HUANG, W. ZHENG, Z. CUI, AND G. ZHAO, *Learning from hierarchical spatiotemporal descriptors for micro-expression recognition*, IEEE Transactions on Multimedia, 20 (2018), pp. 3160–3172.

[249] K. ZUIDERVELD, *Contrast Limited Adaptive Histogram Equalization*, Academic Press Professional, Inc., USA, 1994, pp. 474–485.