Sudan University of Science and Technology

College of Graduate Studies

Development of Adaptive Mask for Regions-based Facial Micro-Expressions Recognition Using Objective Classes

تطوير قناع تكيفى للتعرف على تعبيرات الوجه الدقيقة المعتمدة على المناطق بإستخدام الفئات الموضوعية

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

By

Walied Ali Merghani Elzwain

Supervised by

Dr. Moi Hoon Yap

August 2020

(وَالْكَاظِمِينَ الْغَيْظَ وَالْعَافِينَ عَنِ النَّاسِ)

(آل عمران : 134)

Abstract

Facial micro-expressions are very brief spontaneous expressions that appear on the face of humans when a person either deliberately or unconsciously conceals a feeling or emotion. Unlike regular facial expressions, it is difficult to fake a micro-expression due to its subtlety and very short duration. Facial micro-expression recognition is still a challenging area with some gaps and limitations which need to fill such as the low accuracy achieved so far for the recognition, there is no investigation on the effect of frame rate and resolution changes for facial micro-expression recognition on the existing datasets, the emotion classes within the datasets are based on Action Units (AUs) and self-reports, which creating conflicts when training the machine learning method.
This research focuses on exploring the best features of descriptors and representation of facial micro-expressions for recognition. Objective classes labeling based on AUs has been introduced. In addition, the effect of resolution and frame rate for facial micro-expressions recognition have been experimented and identified.

To provide new insights into the roles of temporal and spatial settings, an investigation has been conducted into the use of different frame rates and resolutions on current benchmark datasets (SMIC and CASME II). By using Temporal Interpolation Model, SMIC has been sub-sampled (original frame rate is 100 fps) to 50 fps, and CASME II (original frame rate is 200 fps) into 100 fps and 50 fps. In addition, the resolution settings have been adjusted to three scaling factors: (100% original resolution), 75%, and 50%. Three feature types have been used to test the performance of these settings.

Emotion classes within the current dataset are based on self-reports. Instead of that, restructuring for the classes has been done around the AUs to removes the potential bias of human reporting. A list of AUs and combinations are proposed for a fair categorization of the SAMM and CASME II datasets. Categorizing in this way make datasets classes more unified. Finally, a new method for facial micro-expression recognition has been proposed. The proposed method is a region-based with an adaptive mask for facial micro-expression recognition.

Based on the most frequent Action Units on the two publicly available datasets, i.e. CASME II and SAMM, 14 ROIs are defined. The adaptive mask has been created by calculating the oriented magnitude of optical flow after Gaussian smoothing. also the problem of light condition which considers as micro-movement has been solved using a proposed method called remove random displacements which remove the random pixels caused by brightness changes or head-movements. features have been extracted from each region using Local Binary Patterns on Three orthogonal Planes (LBP-TOP).

The proposed method evaluated on two benchmark datasets: CASME II and SAMM.It performing better than state-of-the-art, achieved results up to 69.6 and 0.59 in terms of accuracy and F1-score respectively on CASMEII, and 59.7 and 0.51 on SAMM. The proposed method has tested using objective classes and achieved a higher result reach to 77.9 accuracy and 0.72 F1-score on CASME II.

المستخلص

تعبيرات الوجه الدقيقة هي تعبيرات عفوية وجيزة جدًا تظهر على وجه البشر عندما يخفي الشخص عمداً أو دون وعي شعورًا أو عاطفة. على عكس تعابير الوجه العادية ، من الصعب تزوير تعبير دقيق بسبب الحركة الدقيقة والفترة الزمنية القصيرة جدًا. لا يزال مجال التعرف على تعبيرات الوجه الدقيقة يمثل تحدياً مع وجود بعض الثغرات و النواقص التي يجب سدها مثل نسبة التعرف المنخفضة التي تم تحقيقها حتى الآن كذلك لا توجد أي دراسة في تأثير تغيير معدل الإطار ودقة الصورة في التعرف على تعبيرات الوجه الدقيقة على مجموعات البيانات الحالية ،كما تستند فئات العاطفة المستخدمة حاليا داخل مجموعات البيانات إلى التقارير الذاتية مما يخلق لبس عند تدريب طريقة التعلم الآلي.

يركز هذا البحث على استكشاف أفضل السمات لتمثيل و للتعرف على تعابير الوجه الدقيقة. تم إستخدام الفئات الموضوعية استنادًا إلى وحدات النشاط بدلاً عن التقارير الذاتية كتصنيف للعاطفة. بالإضافة إلى ذلك ، تم اختبار وتحديد تأثير التغيير في دقة الصورة ومعدل الإطارات للتعرف على تعابير الوجه الدقيقة.

لتقديم رؤية جديدة حول أدوار الإعدادات الزمنية والمكانية ، تم إجراء دراسة بإستخدام معدل إطارات و دقة صور مختلفة على مجموعات البيانات المرجعية الحالية (SMIC و CASME II). بإستخدام نموذج الاستيفاء الزمني ، تم أخذ عينات فرعية من SMIC (معدل الإطار الأصلي هو 100 إطارًا في الثانية) إلى 50 إطارًا في الثانية ، و CASME II (معدل الإطارات الأصلي هو 200 إطارًا في الثانية) إلى 100 إطارًا في الثانية و 50 إطارًا في الثانية. بالإضافة إلى ذلك ، تم ضبط إعدادات دقة الصورة على ثلاثة عوامل تحجيم: (دقة أصلية 100٪) , 75٪ و 50٪. تم استخدام ثلاثة أنواع من الميزات لاختبار أداء هذه الإعدادات.

تستند الفئات في مجموعة البيانات الحالية إلى التقارير الذاتية. بدلاً من ذلك ، تم إجراء إعادة هيكلة للفئات حول وحدات النشاط لإزالة التحيز المحتمل للتقارير البشرية. تم اقتراح قائمة بوحدات النشاط وتركيباتها من أجل تصنيف عادل لمجموعات البيانات SAMM و CASME II. يجعل التصنيف بهذه الطريقة فئات مجموعات البيانات أكثر توحيدًا.

وأخيرًا ، تم اقتراح طريقة جديدة للتعرف على تعبيرات الوجه الدقيقة. تعتمد الطريقة المقترحة على المناطق ذات الأهمية (14 منطقة) مع قناع تكيفي للتعرف على تعبيرات الوجه الدقيقة. استنادًا إلى وحدات النشاط الأكثر شيوعًا على مجموعتي البيانات CASME II و SAMM تم تحديد 14 منطقة مهمة في الوجه. تم إنشاء القناع التكيفي عن طريق حساب الحجم الموجه للتدفق البصري بعد عملية التنعيم. كما تم حل مشكلة حالة الضوء التي تعتبر حركة دقيقة باستخدام طريقة مقترحة تسمى إزالة الإزاحة العشوائية والتي تزيل البكسل العشوائي الناتج عن تغيرات السطوع أوحركات الرأس. تم استخراج الميزات من كل منطقة باستخدام الأنماط الثنائية المحلية على ثلاثة مستويات متعامدة (LBP-TOP).

تم تقييم الطريقة المقترحة على مجموعتي بيانات مرجعيتين: CASME II و SAMM ، وكان أداؤها أفضل من أحدث النتائج المحققة بنتائج وصلت إلى 69.6 و 0.59 من حيث نسبة التعرف و F1-Score على التوالي في CASMEII و 59.7 و 0.51 على SAMM . تم اختبار الطريقة المقترحة باستخدام فئات موضوعية وحققت نتيجة أعلى تصل إلى 77.9 دقة و 0.72 F1-Score في CASME II.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor, Dr. Moi Hoon Yap, for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank Dr. Adrian Keith Davison, for his insightful comments and encouragement.

My sincere thanks also go to SUST represented by Prof.Osman and all Ph.D. program staff, who provided me an opportunity for this study, and who gave access to the laboratory and research facilities. Without their precious support, it would not be possible to conduct this research.

I thank my fellow lab mates for the stimulating discussions, and for all the fun we have had in the last years. Also, I thank all my friends for their continuous asking about my progress.

My most sincere thanks go to my family, for always believing I could succeed. To my Mother and my Wife, to all my brothers and sisters.

Contents

1 Introduction and Background 1

1.1 Introduction 1

1.2 Background 1

1.2.1 Importance and applications 1

1.2.2 Why FME recognition system? 2

1.2.3 Challenges 2

1.2.4 Research direction 2

1.3 Problem statement 3

1.4 Aim and objectives 4

1.5 Thesis contributions 4

1.6 Research Methodology 4

1.7 Structure of the Thesis 5

2 Literature Review 6

2.1 Introduction 6

2.2 Facial Micro-expression Datasets 8

2.2.1 Non-spontaneous datasets 8

2.2.1.1 Polikovsky Dataset 8

2.2.1.2 USF-HD 8

2.2.1.3 YorkDDT 9

2.2.2 Spontaneous datasets 9

2.2.2.1 Chinese Academy of Sciences Micro-Expressions (CASME) 10

2.2.2.2 Spontaneous Micro-expression Corpus (SMIC) 10

2.2.2.3 Chinese Academy of Sciences Micro-Expression II (CASME II) 11

2.2.2.4 Spontaneous Actions and Micro-Movements (SAMM) . . 12

2.2.2.5 A Dataset of Spontaneous Macro-Expressions and Micro-Expressions (CAS(ME)^2) 13

2.2.3 Dataset comparison 14

2.3 Features 14

2.3.1 3D Histograms of Oriented Gradients (3DHOG) 15

2.3.2 Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) and Variations 16

2.3.3 Histogram of Oriented Optical Flow (HOOF) 24

2.3.4 Deep Learning Approaches 25

2.3.5 Other Feature Extraction Methods 27

2.4 Performance Metrics and Validation Techniques 30

2.4.1 Metrics 30

2.4.2 Validation Techniques 31

2.5 Challenges 31

2.5.1 The effect of Spatial Temporal Settings in Data Collection 32

2.5.2 Emotional Classes versus Objective Classes in Data Labelling . . . 34

2.5.3 Face Regions in Data Analysis 35

2.5.4 Deep Learning versus Conventional Approaches 37

2.5.5 Standardisation of Metrics 38

2.5.6 Real-world Implementation 38

2.6 Summary 39

3 Methodology **40**

3.1 Introduction 40

3.2 FME Recognition Pipeline 40

3.2.1 Preprocessing 40

3.2.2 Features description 41

3.2.2.1 Local Binary Pattern-Three Orthogonal Planes(LBP-TOP) 41

3.2.2.2 Histograms of Oriented Gradient on 3D (HOG3D) 43

3.2.2.3 Histogram of Oriented Optical Flow (HOOF) 44

3.2.3 Classification 45

3.2.3.1 Support vector machine(SVM) using sequential minimal optimization (SMO) 45

3.3 The Implication of Spatial Temporal Changes on Facial Micro-Expression Analysis 46

3.3.1 Overview 46

3.3.2 Frame Rate Subsampling 47

3.3.3 Resolution Down-Scale 48

3.3.4 Feature Representation and Classification 48
3.4 Objective Classes for Micro-Facial Expression Recognition 49
3.4.1 Overview 49
3.4.2 Datasets Analysis 50
3.4.2.1 CASME II 50
3.4.2.2 SAMM 51
3.4.3 Classes Restructuring 52
3.4.4 Feature Representation and Classification 53
3.4.5 Evaluating objective classes on a composite database 53
3.5 Adaptive Mask for Region-based Facial Micro-Expression Recognition . . 55
3.5.1 Overview 55
3.5.2 Facial Landmarks Detection 55
3.5.3 The Facial Action Coding System 56
3.5.4 FACS-Based Regions 57
3.5.5 Smoothing 57
3.5.6 Optical Flow 58
3.5.7 Oriented Magnitude Mask 59
3.5.8 Remove Random displacements 61
3.5.9 Feature Decriptors, Classification and Validation 62
3.6 Summary 63

4 Result and Discussion 64
4.1 Introduction 64
4.2 The Implication of Spatial Temporal Changes on Facial Micro-Expression Analysis 64
4.2.1 Comparison of the State-of-the-art Methods 65
4.2.2 Temporal Analysis 65
4.2.3 Spatial Analysis 69
4.2.4 Features Analysis 69
4.2.5 Result Summary 69
4.3 Objective Classes for Micro-Facial Expression Recognition 70
4.3.1 Evaluating objective classes on a composite database 73
4.4 Adaptive Mask for Region-based Facial Micro-Expression Recognition . . 74
4.5 Summary 76

5 Conclusion and Future work **77**

5.1 Conclusion 77

5.2 Future work 78

Bibliography **79**

List of Tables

2.1 The frequency occurance of each emotion category in the CASME dataset . 10

2.2 Type of Emotions Frequency in SMIC [83] 11

2.3 The frequency of each micro-expression class in the CASME II dataset. . . 12

2.4 Type of emotion and their frequencies in the CAS(ME)^2 dataset. 13

2.5 A Summary of non-spontaneous and spontaneous datasets. 14

2.6 Summary (Part I: 2009 - 2016) of the feature types, classifier and metrics used over the past decade for micro-expression recognition by year and authors. 17

2.7 Summary (Part II: 2016 - 2019) of the feature type, classifier and metrics used over the past decade for micro-expression recognition by year and authors. 24

2.8 A breakdown of the number of clips categorised into estimated emotion classes for the SAMM dataset. 34

2.9 Each class represents AUs that can be linked to emotion. 35

3.1 A summary of the different features of the CASME II and SAMM datasets. 51

3.2 Each class represents AUs that can be linked to emotion. 52

3.3 The total number of movements assigned to the new classes for both SAMM and CASME II. 53

3.4 The total number of movements assigned to the new classes for both SAMM and CASME II. 54

3.5 A Summary of AUs frequency on CASME II and SAMM. 59

3.6 The region number, name and its associated AUs. 60

4.1 The performance of the state-of-the-art methods on SMIC and CASME II. Note that we have only included some popular previous works that reported their results on both datasets. 65

4.2 The results of CASME II for both the 10-fold cross-validation and leave-one-subject-out for the 3DHOG, HOOF and LBP-TOP features with a varying resolutions and frame rates. 66
4.3 The results of SMIC for both the 10-fold cross-validation and leave-one-subject-out for the 3DHOG, HOOF and LBP-TOP features with varying resolutions and frame rates. 67
4.4 The sampling variation of classes distribution through the 10 folds. 70
4.5 Results on the CASME II dataset showing each feature, proposed classes, and the original classes defined in [149] for comparison. 72
4.6 Results on the SAMM dataset showing each feature and proposed classes. . 72
4.7 The results for Task B based on LOSO cross validation. LBP-TOP, 3DHOG, HOOF are the baseline methods using 5x5 blocks. Proposed is our proposed method using selective block-based feature fusion representation. . . 74
4.8 Comparison between proposed method with the state-of-the-art methods on CASME II and SAMM. 75
4.9 Results of different experiments. 75

List of Figures

2.1 Sample of HS SMIC dataset with negative expression. 11

2.2 Sample of CASME II dataset with happiness expression, the participant has been FACS coded with AU1+AU12 (Inner brow raiser+lip corner puller). 12

2.3 Sample of SAMM dataset with anger expression, the participant has been FACS coded with AU4+AU7 (Brow lowerer+lid tightener). 13

2.4 Illustration of the number of publications in micro-expressions recognition based on feature types over the past 10 years. 15

2.5 An example of different resolution by downscaling an image from CASME II dataset. From left to right: 100% (Original resolution), 75% of the original resolution, 50% of the original resolution and 25% of the original resolution. 32

2.6 The effect of image resolution on micro-expression recognition using LBP-TOP, 3DHOG and HOOF on two different evaluation method: (a) 10-fold cross validation, and (b) LOSO. 33

2.7 Illustration of face regions: (a) 5 × 5 blocks, (b) 8 × 8 blocks, (c) Delaunay triangulation, and (d) FACS-based regions. 36

3.1 LBP code calculation by using the difference of the neighbourhood pixels around the centre. 42

3.2 LBP is calculated on every block in all three planes. Each plane is then concatenated to obtain the final LBP-TOP feature histogram. 43

3.3 Build HOOF features with 4 bins 44

3.4 Down scale resolution: 100% (original resolution), 75% and 50% from the original resolution. 49

3.5 Sample frames showing Subject 11's micro-expression clip 'EP19_03f' that was coded as an AU4 in the 'others' category. 51

3.6 Sample frames showing Subject 26's micro-expression clip 'EP18_50' that was coded as an AU4 in the 'disgust' category. 52

3.7 Images split into 5x5 blocks. On the left: an example from CASME II; and on the right: an example from SAMM. The total is 25 blocks 54

3.8 Our proposed selective block-based. On the left: an example from CASME II; and on the right: an example from SAMM. The block is reduced from 25 blocks to 18 blocks. 55

3.9 Methodology of proposed method. 56

3.10 FACS AUs defined as two main groups and split into sub-groups. 57

3.11 (a)Facial landmark points on a sample subject of CASMEII (b) 14 ROIs based on the frequency of AUs occurrences. 58

3.12 Preprocessing step: (a) Before smoothing (b) after smoothing. 58

3.13 Illustrations of the magnitude calculation of optical flow in 8 orientation (surrounding images), the centre image shows the average of 8 surrounding images. 61

3.14 Optical flow mask: (a) Oriented magnitude (b) Black and White (c) Removing random displacements (d) applying mask to original image (f) regions after mask. 62

4.1 Comparison of F-Measure using LBP-TOP with varying resolution and frame-rates. The graph shows the best result is using high frame rate and high resolution when evaluated using LOSO validation on CASME II (100%, 75% and 50%) and SMIC (HS100%, HS75% and HS50%). . . . 67

4.2 Comparison of F-Measure using HOOF with varying resolution and frame-rates. The graph shows the best result is using high frame rate and lower resolution when evaluated using LOSO validation on CASME II (100%, 75% and 50%) and SMIC (HS100%, HS75% and HS50%). 68

4.3 Comparison of F-Measure using 3DHOG with varying resolution and frame-rates. The graph shows the overall best result is achieved by using low frame rate when evaluated using LOSO validation on CASME II (100%, 75% and 50%) and SMIC (HS100%, HS75% and HS50%). 68

4.4 Confusion matrix of the original CASME II classes using the LBP-TOP feature, using SMO as a classifier. 71

4.5 Confusion matrix of the proposed classes I-V on the CASME II dataset using the LBP-TOP feature and SMO as a classifier. 71

Abbreviations

3DHOG	3D Histograms of Oriented Gradients
AUs	Action Units
CAS(ME)2	A Dataset of Spontaneous Macro-Expressions and Micro-Expressions
CASME	Chinese Academy of Sciences Micro-Expressions
CNN	convolutional neural networks
FACS	Facial Action Coding System
FME	Facial Micro-Expression
HOOF	Histogram of Oriented Optical Flow
LBP-TOP	Local Binary Pattern-Three Orthogonal Planes
LOSO	Leave one subject out
OF	Optical Flow
ROI	Regions of Interest
SAMM	Spontaneous Actions and Micro-Movements
SMIC	Spontaneous Micro-expression Corpus
SMO	Sequential Minimal Optimization

SVM	Support Vector Machine
USF-HD	University of South Florida - High Defenetion
YorkDDT	York Deception Detection Test

List of Publications

1- Merghani, Walied, Adrian K. Davison, and Moi Hoon Yap. "The implication of spatial temporal changes on facial micro-expression analysis." Multimedia Tools and Applications 78.15 (2019): 21613-21628.

2- Merghani, Walied, Adrian Davison, and Moi Yap. "Facial Micro-expressions Grand Challenge 2018: evaluating spatio-temporal features for classification of objective classes." 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018.

3- Merghani, Walied, and M. Hoon Yap. "Adaptive Mask for Region-based Facial Micro-Expression Recognition." 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020.

4- Davison, Adrian K., Walied Merghani, and Moi Hoon Yap. "Objective classes for micro-facial expression recognition." Journal of Imaging 4.10 (2018): 119.

5- Davison, A., Merghani, W., Lansley, C., Ng, C. C., & Yap, M. H. (2018, May). 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) (pp. 642-649). IEEE.

Chapter 1

Introduction and Background

1.1 Introduction

This chapter introduces the research and terms described in the facial micro-expression recognition. The main terminology used throughout this thesis is defined alongside the problem statement, thesis contributions and the thesis structure.

1.2 Background

Facial Micro-expression (FME) is a kind of very rapid spontaneous facial expression [47], which appears when a person either deliberately or unconsciously conceals a feeling [30]. Haggard and Isaacs [49] were the first to describe micro expressions (also known as micro momentary expressions) in their study of psychotherapeutic interviews, followed by Ekman and Friesen [32], whose discovered the patient lied to conceal the plan to commit suicide [82]. Micro-expression can be characterized by its short duration. Some of the psychological research shows that facial FMEs is less than 0.2 seconds [60]. Till now there is no standard duration but generally, the upper limit can be up to 0.5 second [150]. As in the regular expression, there are seven universal FMEs: disgust, anger, fear, sadness, happiness, surprise, and contempt [21, 47].

1.2.1 Importance and applications

Unlike regular facial expressions, FME could not be fake [60] and might be uncontrollable to the person, therefore it may be a good cue for lie detection in interrogation. FME is universal rather than culturally defined [21], this gives it a widely used. In addition to interrogation, FME can take advantage of the medical field, which could be used to get the true feelings of psychological patients and provide appropriate treatment accordingly.

1.2.2 Why FME recognition system?

The short duration of FME makes it hard to be observed with a naked eye. Ekman developed FME training tool [60] to improve this ability. Even with this tool, the recognition accuracy has been achieved by humans just around 40% [44]. Due to these reasons, it is necessary to establish research on recognition, detection and analyzing FME.

1.2.3 Challenges

As in the majority of computer vision research, FME requires datasets to run the experiment and validate the result. Creating spontaneous dataset is a real challenge because it is not easy to induce FME and it can be done just in a specific situation when the individual tries to conceal his true feelings, this led to limit of available datasets.

To recognize or detect FME, features need to be extracted like any other recognition or detection systems. Due to the subtle movements in FME, the features should be well described and only related to micro-expressions without interfering with other subtle movements. Muscles movement in FME could not appear in both upper and lower halves of the face simultaneously. Therefore, previous work for regular expression recognition may not be suitable for FMEs [93].

1.2.4 Research direction

Research in FMEs includes mainly recognition and detection. Some research also focuses on creating new datasets that can be used to evaluate different methods and algorithms. Recognition concern on "What is this emotion?" between known universal emotions. On the other hand, detection experiments aim to distinguish between micro-expression and other facial-movements. Research in this area achieved a good result reach 92% in terms of accuracy [21]. Yan et al. [150, 154] and Li et al. [83] put the priority for creating FMEs datasets with some baseline to be used as a benchmark for the others. Eliciting micro-expression is difficult because it appears only in a certain situation when an individual conceals his true emotion [150]. Also, there still challenges in eliciting some expression under the lab environment like sadness.

This research will focus on FME recognition, which still a challenging area with some gaps and limitations need to fill such as:

1. Features need to be well described enough for better recognition rate.

2. There is no analysis of the effect of changes in video resolution and frame rate.

3. The emotion classes within the datasets are based on Action Units and self-reports, creating conflicts when training the machine learning method.
4. There is a still lack of available datasets for FMEs.
5. Hardly to differentiate between two micro-expressions for one subject.
6. The dissimilarity of the same micro-expression between two subjects due to the different facial morphology.

New insights and recommendations have been provided for advancing the micro-expression analysis research. Also, good guidelines provided for beginners and detailed challenges and recommendations for those who are already working in this area. The focuses will be on the first three issues, exploring the features descriptors for best FME recognition, analyzing the effect of changes in video resolution and frame rate for FME recognitions and classifying expressions using Action Units instead of predicted emotion.

1.3 Problem statement

The low accuracy achieved so far for the FME recognition does not meet the needs of applications, such as interrogation and medical applications. This low accuracy came from the low intensity of facial micro-movements, which is hard to get a good description and representation features for it. Most of the features representation for FME recognition extracted from the whole face although micro-movements occur in parts of the face. In addition to that, the brightness and head-movements may confuse machine learning which considered it as facial movements. There are some methods used with the average result achieved and there still room for enhancement. The datasets available for micro-facial expression have different resolutions and different frame rates. To date, there is no investigation on the effect of frame rate and resolution for micro-facial expression recognition. Currently, the emotion classes within the datasets are based on Action Units and self-reports, which leads to conflicts when training the machine learning method.
The main problem of this research is the lack of focus on the face areas that represent FME in the previous studies, in addition to the confusion that can be considered as facial muscles movements such as head movement and changes in brightness.

1.4 Aim and objectives

The primary aim of this research is to develop a method for FME recognition. To achieve this aim, the following objectives have been established: The objectives are:

1. To conduct comprehensive literature review to explore the features descriptors and representation for FME recognition.

2. To investigate and identify the effect of resolution and frame rate changes on FME recognition.

3. To classify expressions using Action Units instead of predicted emotion, to remove the potential bias of human reporting.

4. To propose a method of recognizing FME using an adaptive mask for region-based.

5. To evaluate the performance of the proposed methods against state-of-the-art algorithms.

1.5 Thesis contributions

1. New method for facial micro-expression recognition.

2. Investigate of the implication of spatial temporal changes on facial micro-expression analysis.

3. Introduces a objective classes for facial micro-expression recognition.

1.6 Research Methodology

a comprehensive literature review has been done for better known about the facial micro-expressions research area. Then the acquisition for publicly available datasets done. Preliminary studies by repeating existing algorithms also have been done. Develop the proposed method has been established by preprocessing steps which include: cropping for facial parts, converting images to gray level and smoothing. For localization the ROIs analysis on AUs done. Then a visualization for facial micro-movements has been done using optical-flow. Features have been extracted to represent FME to classify the emotions. Finally, an evaluation of the proposed method against the state-of-the-art has been done. Writing up the thesis concluded the research.

1.7 Structure of the Thesis

This thesis is split into five chapters as follow: Chapter 1; provides an overview of this thesis, introducing the work presented and outlining what to expect from the research. Chapter 2; presents fundamental knowledge and a review of the literature relating to facial micro-expression recognition. Given the nature of the field, feature representation and datasets. Chapter 3; provides the methodology for all the thesis contributions. Describe the general pipeline for FME recognition. Chapter 4; provides the contributions results and its discussion. Chapter 5; concludes this thesis with a summary of contributions, the limitations faced in the field of FME analysis and the future research direction.

Chapter 2

Literature Review

2.1 Introduction

Over the past ten years, automatic micro-expressions recognition has attracted increasing attention from researchers in psychology, computer science, security, neuroscience and other related disciplines. The aim of this chapter is to provide the insights of automatic micro-expressions analysis and recommendations for future research. There has been some of datasets released over the last decade that facilitated the rapid growth in this field. However, comparison across different datasets is difficult due to the inconsistency in experiment protocol, features used and evaluation methods. To address these issues, we review the datasets, features and the performance metrics deployed in the literature. Relevant challenges such as the spatial temporal settings during data collection, emotional classes versus objective classes in data labelling, face regions in data analysis, standardisation of metrics and the requirements for real-world implementation are discussed. We conclude by proposing some promising future directions to advancing micro-expressions research.

Facial expression research has a long history and accelerated through the 1970s. The modern theory on basic emotions by Ekman et al [27, 29, 38] has generated more research than any other in the psychology of emotion [120]. They outline 7 universal facial expressions: happy, sad, anger, fear, surprise, disgust and contempt, as the universality of emotion. When an emotional episode is triggered, there is an impulse which may induce one or more of these expressions of emotion.

Facial micro-expression (henceforth, micro-expression) analysis has become an active research area in recent years. Micro-expressions occur when a person attempts to conceal their true emotion [29, 31]. When they consciously realise that a facial expression is occurring, the person may try to suppress the facial expression because showing the emotion may not be appropriate or could be due to a cultural display rule [99]. Once the suppression has occurred, the person may mask over the original facial expression and cause a micro-expression. In a high-stakes environment, these expressions tend to become more likely as there is more risk to showing the emotion.

Micro-expressions contain a significant and effective amount of information about the true emotions which may be useful in practical applications such as security and interrogations [107, 42, 44]. It is not easy to extract this information due to the brief movements in micro-expressions, where there is a need for the features to be more descriptive. The difficulty also comes from one of the main characteristics of micro-expressions which is the short duration, with the general standard being a duration of no more than 500 ms [153]. Other definitions of speed that have been studied show micro-expressions to last less than 250 ms [28], less than 330 ms [38] and less than half a second [42]. Following Ekman and Friesen as first to define a micro-expression [32], a usual duration considered is less than 200 ms. Duration is the main feature that distinguishing micro-expressions from macro-facial expressions [122], which make it more challenging than micro-expressions in the following aspects:

1. Difficulties for human to spot micro-expressions: Humans find it difficult to spot micro-expressions consistently [42]. This is due to macro-expressions tend to be large and distinct, whereas micro-expressions are very quick and subtle muscle movements.

2. Datasets Creation: It is difficult to induce micro-expressions if compared to macro-expressions. Current available micro-expression datasets were induced in a laboratory controlled environment. Macro-expressions can be recorded by normal camera. However, the speed and subtlety of micro-expressions require high-speed camera, where this digital capture device produces more noisy data than the normal camera.

3. The history of algorithm development: Automated micro-expression recognition is relatively new (found work in 2009 [118]), when compared with facial expression recognition (found in 1990s [39, 74]).

Although both micro and macro-expressions loosely related due to the facial expression aspect, these two topics should be looked upon as different research problems. Our focus is to provide comprehensive review and new insights for micro-expressions. For review in macro-expressions, please refer to [109, 40].

This chapter introduces and surveys recent research advances of micro-expressions. We present a comprehensive review and comparison on the datasets, the state-of-the-art features for micro-expression recognition and the performance metrics. We demonstrate the potential and challenges of micro-expression analysis. This rest of the chapter is organised as follows: Section 2 provides a review on publicly available datasets. Section 3 presents the feature representation. Detailed performance metrics used in this field are shown in Section 4 and Section 5 outlines challenges and Section 6 concludes this chapter by providing future recommendations.

2.2 Facial Micro-expression Datasets

This section will compare and contrast the relevant publicly available datasets for facial micro-expressions analysis.

2.2.1 Non-spontaneous datasets

The earlier research dependent on non-spontaneous datasets. Here we present a review on the three earliest non-spontaneous datasets.

2.2.1.1 Polikovsky Dataset

One of the first micro-expression datasets was created by Polikovsky et al. [118]. The participants were 10 university students in a laboratory setting and their faces were recorded at 200fps with a resolution of 640×480. The demographic was reasonably spread but limited in number with 5 Asians, 4 Caucasians and 1 Indian student participants.

The laboratory setting was set up to maximise the focus on the face, and followed the recommendations of mugshot best practices by McCabe [100]. To reduce shadowing, lights were placed above, to the left and right of the participant. The background consisted of a uniform colour of approximately 18% grey. The camera was also rotated 90 degrees to increase the pixels available for face acquisition.

The micro-expressions in this dataset were posed by participants whom were asked to perform the 7 basic emotions with low muscle intensity and moving back to neutral as fast as possible. Posed facial expressions have been found to have significant differences to spontaneous expressions [1], therefore the micro-expressions in this dataset are not representative of natural human behaviour and highlights the requirement for expressions induced naturally. Further, this dataset is not publicly available for further study.

2.2.1.2 USF-HD

Similar to the previous dataset, USF-HD [124] includes 100 posed micro-expressions recorded at 29.7 fps. The participants were shown various micro-expressions and told to replicate them in any order of preferencethey ed. As with the Polikovly described dataset, posed not re-create a real-world scenario and replicating other people's micro-expressions does not represent how these movements would be presented by the participants themselves.

Recording at almost 30 fps can risk losing important information about the movements. In addition, this dataset defined micro-expressions as no higher than 660 ms, which is longer than the previously accepted definitions. Moreover, the categories for micro-expressions are smile, surprise, anger and sad, which is reduced from the 7 universal expressions by missing out disgust, fear and contempt. This dataset has also not been made available for public research use.

2.2.1.3 YorkDDT

As part of a psychological study named the York Deception Detection Test (YorkDDT), Warren et al. [143] recorded 20 video clips, at 320×240 resolution and 25 fps, where participants truthfully or deceptively described two film clips that were either classed as emotional, or non-emotional. The emotional clip, intended to be stressful, was of an unpleasant surgical operation. The non-emotional clip was meant to be neutral, showing a pleasant landscape scene.

The participants viewing the emotional clip were asked to describe the non-emotional video, and vice versa for the participants watching the non-emotional clip. Warren et al. [143] reported that some micro-expressions occurred during both scenarios, however these movements were not reported to be available for public use.

During their study into micro-expression recognition, Pfister et al. [115] managed to obtain the original scenario videos where 9 participants (3 male and 6 female) displayed micro-expressions. They extracted 18 micro-expressions for analysis, 7 from the emotional scenario and 11 from the non-emotional version.

Other than the very low amount of micro-expressions in this dataset, it is created through a second source that do not go into a large amount of detail about AU, or participant demographic. With the data unable to be publicly accessed, it is not possible to study these micro-expressions. It is also an issue with the frame rate being so low, the largest amount of frames for analysis would be around 12-13 frames. The lowest reported micro-expression length was 7 frames.

2.2.2 Spontaneous datasets

Developing micro-expression spontaneous datasets is one of the biggest challenges faced in this research area. It is difficult to elicit micro-expressions because they are difficult to fake, so we need to get the true emotion while the person try to hide it. Some spontaneous datasets to date include: SMIC [83], CASME [154], CASME II [150], SAMM [18] and CAS(ME)^2 [119]. SAMM was designed for micro-movements with less emphasis on the emotional side for increased objectiveness. Available datasets will be described in this section.

2.2.2.1 Chinese Academy of Sciences Micro-Expressions

Yan et al. [154] created a spontaneous micro-expression dataset called CASME. The dataset contains 195 samples of micro-expressions with a frame rate of 60 fps. These 195 samples were selected from more than 1500 facial movements, where 35 participants (13 females, 22 males) took part. The clips were divided into two classes depending on the environmental setting and cameras used.

Class A Samples in this class recorded by BenQ M31 camera at 60 fps, and the resolution is set to 1280×720 pixels. Natural light was used for recording.

Class B A GRAS-03K2C camera recording at 60 fps was used to record samples in this class with resolution set to 640×480 pixels. For class B two LED lights have been used.

Table 2.1 shows each emotion class and the frequencies at which they occur in the CASME(A and B) dataset.

Table 2.1: The frequency occurance of each emotion category in the CASME dataset

Emotion	Frequency
Amusement	5
Sadness	6
Disgust	88
Surprise	20
Contempt	3
Fear	2
Repression	40
Tense	28

2.2.2.2 Spontaneous Micro-expression Corpus

Li et al. [83] built the SMIC dataset, which was recorded in an indoor environment with four lights from the four upper corners of the room. To induce strong emotions 16 movie clips were selected and shown to participants on a computer monitor. Facial expressions

Table 2.2: Type of Emotions Frequency in SMIC [83]

Dataset	positive	negative	surprise	total
HS	51	70	43	164
VIS	28	23	20	71
NIR	28	23	20	71

have been gathered using a camera fixed on the top of monitor while participants watched movie clips.

The dataset is spontaneous, 20 participants (6 females and 14 males) participated in the experiment. A high speed (HS) camera set to 100 fps and resolution of 640×480 was used to gather the expressions from the first ten participants. A sample from this HS dataset is shown in Fig. 2.1. A normal visual camera (VIS) and near-infrared (NIR), both with 25 fps and resolution of 640×480, were used for all 20 participants. The lower frame rates of the latter two cameras can help to check whether the current method can be effective at this speed.

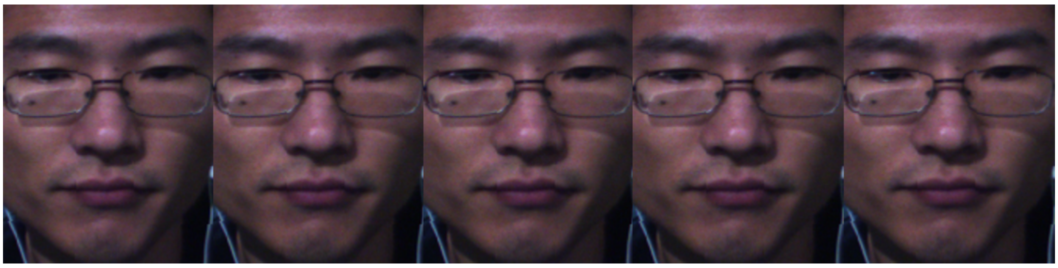


Figure 2.1: Sample of HS SMIC dataset with negative expression.

The accepted duration of micro-expression for SMIC is 1/2 second. Since not every participant showed micro-expressions when recording SMIC the final dataset includes 164 micro-expression clips from 16 participants recorded in HS dataset. While VIS and NIR datasets include 71 clips from 8 participants. Emotions in SMIC were classified into 3 classes (positive, negative and surprise). Table 2.2 show the number of emotions in any class according to the type of dataset.

2.2.2.3 Chinese Academy of Sciences Micro-Expression II

CASME II has been developed by Yan et al. [150], which succeeds the CASME dataset [154] with major improvements. All samples in CASME II are spontaneous and dynamic micro-expressions with high frame rate (200 fps). There is always a few frames kept before and after each micro-expressions, to make it suitable for detection experiments, however the amount of these frames can vary across clips. The resolution of samples is 640×480 pixels for recording, which were saved as MJPEG with a resolution of around 280×340 pixels for the cropped facial area. Fig. 2.2 shows a sample from the CASME II with a happiness-class expression. The micro-expressions were elicited in a well-controlled laboratory environment. The dataset contains 247 micro-expressions (gathered from 35 participants) that were selected from nearly 3000 facial movements and have been labeled with action units (AUs) based on the Facial Action Coding System (FACS) [34]. Lighting flickers were avoided in the recordings and highlights to the regions of the face have been reduced.

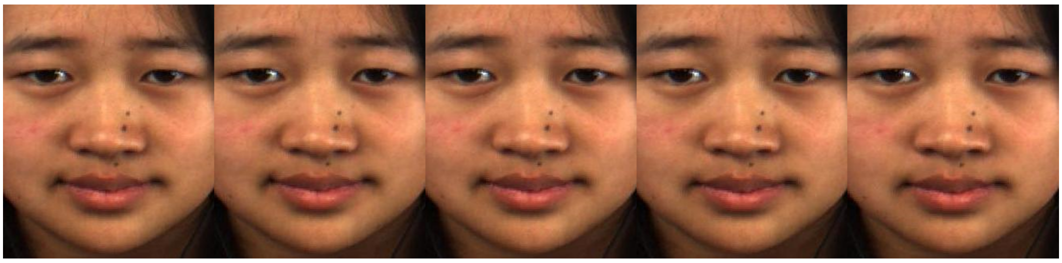


Figure 2.2: Sample of CASME II dataset with happiness expression, the participant has been FACS coded with AU1+AU12 (Inner brow raiser+lip corner puller).

Table 2.3: The frequency of each micro-expression class in the CASME II dataset.

Emotion	Frequency
Happiness	33
Disgust	60
Surprise	25
Repression	27
Others	102

2.2.2.4 Spontaneous Actions and Micro-Movements

The Spontaneous Actions and Micro-Movements (SAMM) [18] dataset is the first high-resolution dataset of 159 micro-movements induced spontaneously with the largest variability in demographics. The inducement procedure was based on the 7 basic emotions [29] and recorded at 200 fps. An example from the SAMM dataset can be seen in Fig. 2.3. As part of the experimental design, each video stimuli was tailored to each participant, rather than getting self-reports after the experiment. This allowed for particular videos to be chosen and shown to participants for optimal inducement potential. The experiment comprised of 7 stimuli used to induce emotion in the participants who were told to suppress their emotions so that micro-movements might occur. To increase the chance of this happening, a prize of £50 was offered to the participant that could hide their emotion the best, therefore introducing a high-stakes situation [29, 31]. Each participant completed a questionnaire prior to the experiment so that the stimuli could be tailored to each individual to increase the chances of emotional arousal. There is a total of 159 FACS-coded micro-movements reported in this dataset.

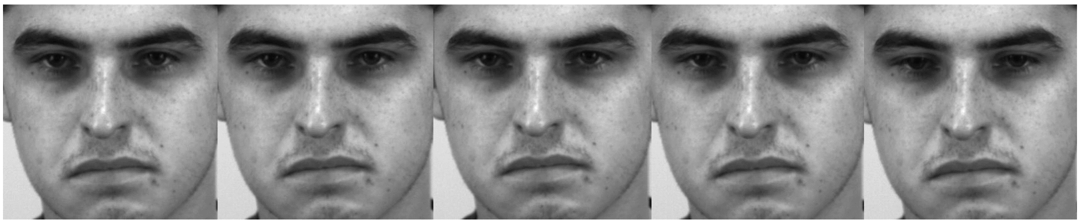


Figure 2.3: Sample of SAMM dataset with anger expression, the participant has been FACS coded with AU4+AU7 (Brow lowerer+lid tightener).

2.2.2.5 A Dataset of Spontaneous Macro-Expressions and Micro-Expressions

Qu et al. [119] presented a new facial database with macro- and micro-expressions, which included 250 and 53 samples respectively selected from more than 600 facial movements. This database has been collected from 22 participants (6 males and 16 females) with mean age of 22.59 years (standard deviation: 2.2). A Logitech Pro C920 camera was used to record samples at frame rate equal to 30 fps and resolution set to 640×480 pixels. CAS(ME)^2 has been labelled using combinations of AUs, self-reports and the emotion category decided for the emotion-evoking videos. This database contains four emotion categories: positive, negative, surprise and other which is shown in Table 2.4 with their frequency occurrence.

Table 2.4: Type of emotion and their frequencies in the CAS(ME)^2 dataset.

Emotion	Macro-expression	Micro-expression
Positive	87	6
Negative	95	19
Surprise	13	9
Other	55	19

Table 2.5: A Summary of non-spontaneous and spontaneous datasets.

Dataset	Participants	Resolution	FPS	Samples	Emotion Classes	FACS Coded	Ethnicities
Polikovsky [118]	11	640×480	200	13	7	Yes	3
USF-HD [124]	N/A	720×1280	29.7	100	4	No	N/A
YorkDDT [143]	9	320×240	25	18	N/A	No	N/A
CASME [154]	35	640×480, 1280×720	60	195	7	Yes	1
SMIC [83]	20	640×480	100 and 25	164	3	No	3
CASME II [150]	35	640×480	200	247	5	Yes	1
SAMM [18]	32	2040×1088	200	159	7	Yes	13
CAS(ME)^2 [119]	22	640×480	30	250 macro, 53 micro	4	No	1

2.2.3 Dataset comparison

Table 2.5 shows summary of a comparison of the datasets. Due to the non-spontaneous datasets were not made available, it is not been possible to provide a critical review on those datasets. Overall, CASME II has a high number of micro-expression samples collected from high number of participants (35 participants), similar to CASME but with 195 samples. There is no distribution in ethnicities in CASME and CASME II, where all participants are Chinese. SMIC have participants from 3 different ethnicities, but this limitation was overcome in SAMM which has participants from 13 different ethnicities. SAMM also has advantage over the other in age distribution with mean age of 33.24 years (SD: ±11.32). CASME II and SAMM have high frame rate (200 fps). SAMM is the first high-resolution dataset which set to 2040×1088 pixel and a facial area of 400×400. The CAS(ME)^2 has a limited number of micro-expression samples with just 53 collected. In terms of emotion stimuli for the participants, CASME and SAMM have 7 classes, CASME II has 5 classes and SMIC only with 3 classes. CASME, CASME II and SAMM have been coded using FACS. Although SAMM was stimulated by 7 emotional classes, the final label in their first release for the micro-movements only consists of FACS codes - but not emotion classes.

CASME II and SAMM become the focus of the researchers as they equipped with all the criteria needed for micro-expressions recognition: emotion classes, high frame rate, a rich number of micro-expressions and varies in term of the intensity for facial movements.

2.3 Features

The features used in micro-expression recognition will be discussed in this section. Figure 2.4 shows the total number of publications and its feature types based on our review. This is a strong evidence to support the growth of micro-expressions research. It is noted that 3DHOG was used in earlier work but not as popular as HOOF in recent years. LBP-TOP gained popularity in 2014 and maintained its number till today. On the other hand, deep learning gained popularity in the past year.

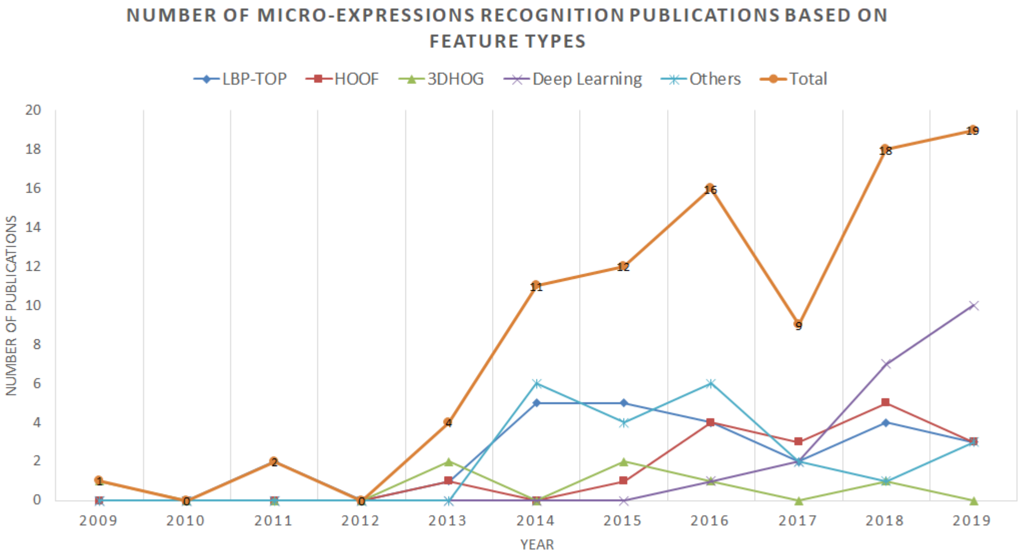


Figure 2.4: Illustration of the number of publications in micro-expressions recognition based on feature types over the past 10 years.

The full summary of the feature types, classifers and metrics used in the past decade is presented on Table 2.6 for Part I (2009-2016) and Table 2.7 for Part II (2016-2019). The detailed algorithms review are categorised into: 3D Histograms of Oriented Gradients, Local Binary Pattern-Three Orthogonal Planes, Histogram of Oriented Optical Flow, Deep Learning Approaches and Other Feature Extraction Methods.

2.3.1 3D Histograms of Oriented Gradients

Polikovsky et al. [118] presented an approach for facial micro-expression recognition. They divided face into 12 regions selected through manual annotation of points on the face and then a rectangle was centred on these points. 3D-histograms of oriented gradients (3DHOG) was used to recognise motion in each region. This approach was evaluated on a posed dataset of micro-expressions captured using a high speed camera (200 fps). 13 different micro-expressions were recognised in this experiment. Their main contribution was to measure the duration of three phases of micro-expressions; constrict of the muscles (Constrict), muscle construction (In-Action) and release of the muscles (Release).

Polikovsky and Kameda [117] used 3DHOG again this time with k-mean classifier and voting procedure. They proposed a method for detecting and measuring timing characteristic of micro-expressions. Frame-by-frame classification was done to detect AUs in 8 video cube regions. The *Onset* frame and *Offset* have higher accuracy than the *Apex* frame, which indicates that their proposed descriptor is suitable for recognition rather than classification for a static frame. To measure AU timing characteristics, the change of bin values in the 3D gradient orientation histogram have been used to reflect the changes and motion accelerations of facial movement. They claimed that this time profile could be used to identify the distinction between posed and spontaneous micro-expression.

Different facial regions having different contributions to micro-expressions as Chen et al. claimed [12] and this being largely ignored by previous studies. They proposed to used 3DHOG features with weighted method and used fuzzy classification for micro-expression recognition. They evaluated their method on 36 samples from CASME II, which contains 4 emotions at a rate of 9 samples per emotion. They compared the result with 3DHOG and weighted 3DHOG and perform better than both achieving average accuracy of 86.67%.

2.3.2 Local Binary Pattern-Three Orthogonal Planes and Variations

Pfister et al. [115] proposed a framework for recognising spontaneous facial micro-expressions. LBP-TOP [164] as a spatio-temporal local texture descriptor has been used to extract dynamic features. In classification phase, Support Vector Machine (SVM), Multiple Kernel Learning (MKL) and Random First (RF) have been used. This framework was evaluated on earlier version of SMIC where the data collected from only six participants with 77 sample of micro-expressions. Temporal Interpolation Model (TIM) has been used to increase the number of frames to achieve more statistically stable histograms. The result of SMIC were compared to York Deception Detection Test (YorkDDT) [143] which were recorded in 25 fps and resolution 320×240. Using leave-one-subject-out (LOSO), the method was evaluated on two corpora and down-sampled SMIC to 25 fps. They have two sets to classify between them emotional vs non-emotional and lie vs truthful. The best result achieved on YorkDDT to classify between first set is an accuracy of 76.2% using MKL and 10 frames. For the second set, the best result is 71.5% using MKL 10 frames and same result using SVM. For SMIC they classify between negative and positive, the best result is 71.4% using MKL 10 frames and 64.9% using MKL and 15 frames for down-sampled SMIC.

Pfister et al. [114] then proposed a method to differentiate between spontaneous and posed facial expressions (Spontaneous Vs Posed (SVP)). They extended Complete Local Binary Patterns (CLBP) which was proposed by Guo et al. [48] to work with dynamic texture descriptor and called it CLBP from Three Orthogonal Planes (CLBP-TOP). They evaluated their proposed method by leave-one-subject-out on a corpus developed by them Spontaneous vs POSed (SPOS). This SPOS provides spontaneous and posed expression for the same subject in the session. It contains 7 subjects with 84 posed and 147 spontaneous expressions. Two cameras have been used to record the corpus, one recorded data from visual (VIS) and the other from near-infrared channel (NIR). Both of cameras used 640×480

Table 2.6: Summary (Part I: 2009 - 2016) of the feature types, classifier and metrics used over the past decade for micro-expression recognition by year and authors.

Year	Authors	Datasets	Feature type	Classifier	Metrics (Best Result)
2009	Polikovsky et al. [118]	Polikovsky	3DHOG	K-means	AUs Classification
2011	Pfister et al.[115]	Earlier version of SMIC	LBP-TOP	SVM, MKL and RF	Accuracy: 71.4% using MKL
2011	Pfister et al.[114]	SPOS	CLBP-TOP	SVM, MKL and LINEAR	Accuracy: 80% using MKL
2013	Polikovsky and Kameda[117]	Polikovsky	3DHOG	K-means	Recognition of 11 AUs
2013	Li et al. [83]	SMIC(HS, VIS and NIR)	LBP-TOP	SVM	Accuracy: 52.11% on VIS
2013	Song et al. [128]	SEMAINE corpus	HOG+HOF	SVR	N/A
2014	Guo et al. [47]	SMIC	LBP-TOP	nearest neighbour	Accuracy: 65.83%
2014	Yan et al. [150]	CASME II	LBP-TOP	SVM	Accuracy: 63.41%
2014	Wang et al. [137]	CASME and CASME II	TICS	SVM	Accuracy: 61.85% on CASME 58.53% on CASME II
2014	Le et al. [78]	CASME II and SMIC	LBP-TOP+STM	AdaBoost	Accuracy: 43.78% on CASME II 44.34% on SMIC
2014	Lu et al. [95]	SMIC, CASME B and CASME II	DTCM	SVM, RF	Accuracy:82.86% on SMIC, 64.95% on CASME and 64.19% on CASME II
2014	Liong et al. [88]	CASME II and SMIC	OSW-LBP-TOP	SVM	Accuracy:57.54% on SMIC 66.40% on CASME II (LOO)
2014	Wang et al. [136]	CASME	DTSA	ELM	Accuracy: 46.90%
2015	House and Meyer [56]	SMIC	LGCP-TOP	SVM	Accuracy 48.1%
2015	Wang et al. [142]	CAMSE II and SMIC	LBP-SIP and LBP-MOP	SVM	Accuracy:66.8% on CASME using LBP-MOP
2015	Wang et al. [138]	CASME and CASME II	TICS, CIELuv and CIELab	SVM	Accuracy:61.86% on CASME 62.30% on CASME II
2015	Le et al. [77]	CASME II	DMDSP+LBP-TOP	SVM, LDA	F1-score: 0.52
2015	Huang et al.[58]	CASME II and SMIC	STLBP-IP	SVM	Accuracy:59.51% on CASME II 57.93% on SMIC
2015	Liu et al. [93]	SMIC, CASME and CASME II	MDMO	SVM	Accuracy:68.86% on CASME 67.37% on CASME II and 80% on SMIC
2015	Li et al. [82]	CASME II and SMIC	LBP, HOG and HIGO	LSVM	Accuracy:57.49% on CASME II 53.52% on SMIC
2015	Kamarol et al. [67]	CASME II	STTM	SVM one-against-one	Accuracy:91.71%
2016	Chen et al.[12]	CASME II(36 samples)	3DHOG	Fuzzy	Accuracy: 86.67%.
2016	Talukder et al. [132]	SMIC	LBP-TOP	SVM	Accuracy: 62% on SMIC-NIR
2016	Duan et al. [26]	CASME II	LBP-TOP from eye region	26 classifiers	Perform better on happy and disgust
2016	Huang et al. [59]	SMIC, CASME and CASME II	improved of STLBP-IP	SVM	Accuracy:64.33% on CASME 64.78% on CASME II and 63.41% on SMIC
2016	Zhang et al. [162]	CASME II	gabor filter+ PCA and LDA	SVM	Good performance on static image
2016	Huang et al. [60]	SMIC, CASME and CASME II	STCLQP	Codebook	Accuracy:64.02% on SMIC 57.31% CASME and 58.39% CASME II
2016	Ben et al. [7]	CASME	MMPTR	Euclidean distance	Accuracy: 80.2%
2016	Liong et al. [91]	SMIC and CASME II	Bi-WOOF	SVM	accuracy 61.0 on CASME II, 62.1 on SMIC-HS
2016	Liong et al. [89]	CASME II and SMIC	Optical Strain	SVM	Accuracy:63.41% on CSME II 52.44% on SMIC
2016	Oh et al. [103]	CASME II and SMIC	I2D	SVM	F1-score: 0.41 and 0.44 on CASME II and SMIC
2016	Wang et al. [139]	CASME and CASME II	STCCA	Nearest Neighbor, SVM	Mean recognition accuracy : 41.20% on CASME 38.39 on CASME II
2016	Zheng et al. [168]	CASME and CASME II	LBP-TOP, HOOF	RK-SVD	Accuracy:69.04% on CASME 63.25% on CASME II
2016	Kim et al. [73]	CASME II	CNN	LSTM	Accuracy: 60.98%

resolution and 25 fps. SVM, LINEAR classifier (LIN), Multiple Kernel Learning (MKL) and fusion of SVM, LIN and Random Forest through majority voting (FUS) have been used as classifiers. They showed that CLBP-TOP overcome LBP-TOP with an accuracy of 78.2%, 72% and 80% on NIR, VIS and combination, respectively.

Li et al. [83] run two experiments on SMIC database for analysing micro-expressions. The first experiment was to detect micro-expressions occurring and the other was to then recognise the type of micro-expression. The detection stage was employed to distinguish a micro-expression and a normal facial expression. On the other hand, recognition discriminated three classes of micro-expression (positive, negative and surprise). A normalisation was done to all faces, followed by a registration to a face model using 68 feature points from an Active Shape Model [15]. Then the faces were cropped according to the eye positions that has been detected using Haar eye detector [102]. LBP-TOP was used for feature extraction from cropped face sequences.

In the VIS and NIR dataset which has a limited number of frames, some problems may arise when applying LBP-TOP. To avoid these problems, TIM was used to allow up-sampling and down-sampling of the number of frames. SVM was used as a classifier and leave-one-subject-out cross validation [144] was used to compute the performance of the two experiments, which were run on three datasets (HS, VIS and NIR). The best accuracy for detection of micro-expressions was 65.55% when evaluating the method on the HS dataset and the X, Y and T parameters were equal to 5, 5 and 1 respectively for LBP-TOP. For micro-expressions recognition, the best accuracy is equal to 52.11% on VIS dataset with X, Y and T having the same value as previous. Avoiding the problem that may arise because the limitation regarding the number of frames by using TIM is considered a strength for this algorithm. However, there is a limitation in using a limited number of recognition classes, since some emotion cannot be judged under ambiguous conditions if more than one expression reported by the participant.

Guo et al. [47] used LBP-TOP features in their micro-expression recognition experiment. To classify these features, they used the nearest neighbour method to compare the distance between unknown samples with entire known samples. Euclidean distance has been used as distance measurement. This method was evaluated on SMIC database. In evaluation, firstly they used Leave-One-Subject-Out (LOSO) and Leave-One-Expression-Out (LOEO) and achieved a recognition accuracy of 53.72% and 65.83% for LOSO and LOEO respectively. In addition, they have conducted experiments for different values of LBP-TOP parameters ($R_X, R_Y, R_T, P_X, P_Y, P_T$) which refer to the radii in axis X, Y and T, and the number of neighbourhood points in the XY, XT and YT planes respectively. The best result was achieved when they set the value to (1,1,2,8,8,8) for parameters. A different distribution of training set and testing set also have been tested and the best result of 63% was achieved when portion of training and testing data with a 5:1 split.

Yan et al. [150] carried out a micro-expression recognition experiments on clips from the CASME II dataset, developed by the same authors. LBP-TOP was used in this experiment to extract the features. SVM was employed as the classifier. With radii varying from 1 to 4 for X and Y, and from 2 to 4 for T (they do not consider T=1 due to little change between two neighbouring frames on a sample rate of 200 fps), and SVM was used as the classifier which classify between five main categories of emotions provided in this experiment (happiness, disgust, surprise, repression and others). The best performance is 63.41% shown when the radii are 1, 1 and 4 for XY, YT and XT planes respectively. Developing high quality datasets with higher temporal (200 fps) and spatial resolution (about 280×340 pixels on facial area), and classify 5 categories of expression with performance 63.41% are the advantages of this method, however they use same method which used for classifying ordinary facial expressions which may not work well for micro-expressions.

Liong et al. [88] proposed Optical Strain Weighted LBP-TOP (OSW-LBP-TOP) method which used optical strain features for micro-expression recognition. They evaluated this feature on CASME II and SMIC. They used SVM as classifier and test different kernel. Their method outperformed the two baseline methods [150, 83] when evaluated on two datasets and achieved accuracy of 57.54% on SMIC when using poly kernel and 66.40% on CASME II when RBF kernel was used.

Davison et al. [21] developed a method to differentiate between micro-movements (MFMs) and neutral expression. This method has been evaluated on CASME II database. LBP-TOP and Gaussian Derivatives (GDs) features are obtained. RF and SVM used as classifiers. Normalization has been done before extract the features to make sure that all faces are in the same position. The images have been divided into 9x8 blocks with no overlapping. Local features obtained for each block after being processed separately using GDs. These local features concatenated into the overall global feature description. LBP-TOP has been calculated for each block through all three planes X, Y and T. In the classification phase data has been separated into testing and training. 100-fold cross-validation was used for testing. The best accuracy achieved is 92.60% when RF has been used and separate testing and training data into 50% and a combination of LBP-TOP and GDs features were used.

House and Meyer [56] implemented a method for micro-expression recognition and detection. They used LBP-TOP and local gray code patterns on three orthogonal planes (LGCP-TOP) as features descriptors. SVM has been used as classifier and SMIC database used to evaluate the method. LGCP-TOP is modified version of LGCP [61] that originally worked for facial expressions and re-worked for analysing the dynamic texture of micro-expressions. They did not overcome the result of LBP-TOP from [83] and they returned this to the feature vectors of LGCP-TOP, which is too large to be classified without over-fitting. They claimed that LGCP-TOP had advantage over LBP-TOP in computational time of the feature descriptor.

Wang et al. [142] inspired two feature descriptors for micro-expressions recognition from the concept of LBP-TOP, LBP-Six Intersection Points (SIP) and LBP-Three Mean Orthogonal Planes (MOP). LBP-SIP is an extension of LBP-TOP and more compact form. This compaction is based on the duplication in computing neighbour points through three planes. Therefore, they only considered the 6 unique points on intersection lines of three orthogonal planes. They claimed that these 6 points carry sufficient information to describe dynamic textures. Vector dimensions in LBP-SIP is 20, in contrast LBP-TOP produce 48 dimensions.

The basic idea of LBP-MOP is to compute features of mean planes rather than all frames in the video. Those two descriptors were evaluated on CASME II and SMIC databases use baseline settings for both datasets [150, 83]. Leave-one-video-out (LOVO) and Leave-one-subject-out (LOSO) cross-validation configurations have been tested on two datasets with different popular kernels for SVM. Also a Wiener filter has been applied for image smoothing to remove noise. LBP-MOP achieved best result (66.8%) on CASME II with linear kernel for SVM using LOVO cross validation and Wiener filter applied in preprocessing step. On SMIC the two methods did not achieve better results than the original LBP-TOP, which achieved 66.46% with Wiener filter and RBF kernel for SVM using LOVO cross validation.

Wang et al. [137] proposed a novel color space model for micro-expressions recognition using dynamic textures on Tensor Independent Color Space (TICS) in which the color components are as independent as possible. They claimed it will enhance the performance of micro-expression recognition. It differs from other literature [150] [83] in getting LBP-TOP features from color as a fourth-order tensor in addition to width, height and time. These experiments were conducted on two micro-expression databases, CASME and CASME II. SVM has been used as classifier. The results show that the performance in TICS is better than that in RGB or grayscale, where the best result achieved on CASME class B is 61.85% and 58.53% on CASME II. Although the accuracy is lower than other state of the art in same area [154, 150] but reveals that TICS may provide useful information more than RGB and grayscale.

In addition to TICS, Wang et al. [138] further show that CIELab and CIELuv are also could be helpful in recognising micro-expressions. They achieved 61.86% accuracy on

CASME class B using TICS, CIELuv and CIELab with different parameters for LBP-TOP. An accuracy of 62.30% was achieved on CASME II using TICS and CIELuv with different parameters for LBP-TOP.

Le et al. [77] proposed a preprocessing step that may enhance recognition rate for micro-expressions. Due to the redundant frames without significant motion which generated when recording with high-speed camera which have high fps, they proposed to use Sparsity-Promoting Dynamic Mode Decomposition (DMDSP) [66] to analyse and eliminate this redundancy. They used LBP-TOP to extract features, with SVM and Linear Discriminant Analysis (LDA) [13] as classifiers. This method was evaluated on CASME II. F1-score, recall and precision have been used to measure the performance. The percentages of reserved frame using DMDSP were varied between 45% and 100% of original frame length. The performance increased while the percentages of reserved frames decreased. The best performance was achieved when 45% of frames were reserved with F1-score, precision and recall equal to 0.52, 0.48 and 0.56 respectively when SVM was used, and 0.47, 0.42 and 0.53 when LDA was used. The performance was compared to the benchmark of CASME II [150] and outperformed the benchmark.

Le et al. [78] defined three difficulties that faced Micro-Expression recognition systems: difficulty of being able to differentiate between two micro-expressions for one subject, namely inter-class similarity, dissimilarity of the same micro-expression between two subjects due to the different facial morphology and behaviour, the uneven distribution of each classes and subjects. They aimed to resolved two latter problems by using facial registration, cropping and interpolation as preprocessing to remove morphological differences. They have proposed variant of AdaBoost to deal with imbalanced characteristics of micro-expressions. The experiments were evaluated on CASME II and SMIC. In addition, TIM has been used to avoid the biases that can be caused by the different frame lengths. For feature extraction LBP-TOP was used and a Selective Transfer Machine (STM) has been used to avoid imbalances which came from the mismatch between distributions of training and testing samples that caused by leave-one-subject-out (LOSO) cross validation to evaluate the datasets. The best result was achieved on CASME II (43.78% recognition rate) when STM used with AdaBoost and fixed frame length of 15 frames, for SMIC, 10 frames give the best result (44.34% recognition rate).

More recently, Talukder et al. [132] used LBP-TOP as features extraction and SVM as classifier after magnified the motion to enhance the low intensity of micro-expression. They conducted their method on the SMIC dataset. They claimed that there is improvement on the recognition result due to the motion magnification applied with average recognition rate up to 62% on SMIC-NIR.

Unlike other studies Duan et al. [26] extracted LBP-TOP from the eye region, not from the whole face. They tested this method on CASME II. They used more than 20 classifiers to train the features. Their method performed better than other methods when classifying happy and disgust expressions.

Huang et al. [58] proposed Spatio-Temporal Local Binary Pattern with Integral Projection (STLBP-IP). They used integral projection to boost the capability of LBP-TOP with experiments conducted on the CASME II and SMIC datasets using SVM as a classifier. When they tested this method on CASME II, it was been compared with several methods from different studies and was used different parameters for LBP-TOP and different kernel for SVM, and also compared with LBP-SIP [142] and LOCP-TOP [129] that achieved a promising performance over these methods with an accuracy rate of 59.51%. When they evaluated their method on SMIC they compared it with [25, 63, 164, 83] and achieved 57.93%.

Huang et al. [59] proposed facial micro-expression recognition method using discriminative spatio-temporal local binary pattern with an improved integral projection. They proposed this method to preserve the shape attribute of micro-expressions. They claimed that extracting features from the global face region lead to ignoring the discriminative information between different classes. They conducted this method on three publicly available datasets: SMIC, CASME and CASME II. They compared this new method with their previous study [58] and demonstrated better results across three datasets with accuracy rate up to 64.33% on CASME, 64.78% on CASME II and 63.41% on SMIC.

Wang et al. [140] used LBP-TOP features to recognise micro-expressions after preprocessed the CASME II dataset with Eulerian Video Magnification (EVM). SVM and k-nearest neighbour (KNN) have been used as classifiers to classify between 5 motions from CASME II dataset. They used leave-one-subject validation with comparison with baseline [150] and other methods [142, 141, 110, 88]. Their proposed method achieved accuracy of up to 75.30%.

Zhang et al. [163] combined local LBP-TOP and local Optical Flow (OF) features after extracted them from local regions of face based on AUs and conducted it on CASME II. They claimed that different local features can perform better than single global features. They compare between different classifiers with different parameters (KNN, SVM and RF), also a comparison between global features and local features has been conducted to prove their hypothesis. Accuracy up to 62.50% has been achieved when they combined two local features with RF classifier.

To solve the cross-database ME recognition problem Zong et al. [173] proposed a method to regenerate the target sample in the process of recognition to have the similar feature distributions as source sample, they called their method Target Sample Re-Generator (TSRG). They evaluated this method on CASME II and three types of SMIC, therefore six experiments have been conducted where the databases served as source and target. Uniform LBP-TOP have been used as features extractor and UAR and WAR used as performance measurement. Comparing to some state-of-the-art method TSRG overcome them in seven experiments in both weighted average recall (WAR) and unweighted average recall (UAR) of 12 in total. They improve their work in [175] and proposed a frame work called it Domain Regeneration (DR) the difference is the generating from both source and target for more similar feature distributions. And they used here three domains to regenerating samples DR-face space for target (DRFS-T), DR-face space for sample (DRFS-S) and DR-line space (DRLS).

By combining heuristic and automatic approaches Liong et al. [90] introduced a method to recognize micro-expression by selecting facial regions statically based on AUs frequency occurring(ROI-selective). They used a hybrid features (Optical Strain Flow (OSF) and block-based LBP-TOP). They tested their method on CASME II and SMIC using SVM as classifier with LOSOCV and LOVOCV to validate the effectiveness. The results have been reported using more than one measurements including accuracy and F-measure and compared with baseline of OSF and LBP-TOP. the method overcome the baseline of two features in all measurement and with both validations, in term of F-measure the best result was 0.51 and 0.31 on SMIC and CASME II respectively. Zong et al. [174] argued that extracting features of fixed-sized facial blocks for micro-expression recognition is not suitable technique. This is due to the fact that it may ignore some information about the AU if it is small or may get overlapping if it is large, leading to the extraction of confusing information. To solve the mentioned problem, they proposed hierarchical division scheme which is dividing face into regions with different densities and different size. They also proposed a learning model called it kernelized group sparse learning (KGSL). More than one feature types have extracted from those hierarchical divisions such as LBP-TOP, LBP-SIP and STLBP-IP. Evaluating of hierarchical division and KGSL have been done on CASME II and SMIC using LOSOCV. The best result achieved on CASME II, when using Hierarchical STLBP-IP + KGSL and it was 63.97% in term of accuracy.Zhao and Xu [166] proposed a novel automatic facial expression recognition framework based on necessary morphological patches (NMPs), they used LBP-TOP as features on CASME II and SMIC and achieved 67.95% on CASME II. Recently some researchers use LBP-TOP as features with non-good results such as Wang et al. [135].

Table 2.7: Summary (Part II: 2016 - 2019) of the feature type, classifier and metrics used over the past decade for micro-expression recognition by year and authors.

Year	Authors	Datasets	Feature type	Classifier	Metrics (Best Result)
2017	Zhang et al. [163]	CASME II	LBP-TOP,Optical Flow	KNN, SVM and RF	Accuracy: 62.50%
2017	Zheng [167]	SMIC, CASME and CASME II	2DGSR	SRC	Accuracy:71.19% and 64.88% on CASME and CASME II
2017	Zong et al. [173]	CASME II and SMIC	LBP-TOP	TSRG	UAR 60.15
2017	Happy and Routray [51]	CASME II, CASME and SMIC	FHOFO	SVM, KNN and LDA	F1-score was 0.5489, 0.5248 and 0.5243 CASME, CASME II and SMIC
2017	Hao et al. [50]	JAFFE	WLD and DBN	DBN	Recognition rate: 92.66
2017	Peng et al. [113]	CASMEI/II	OF	DTSCNN	Accuracy up to 66.67%
2017	Xu et al. [147]	CASMEI	FDM	SVM	accuracy=45.3 F1=0.47
2018	Liong et al. [90]	CASME II and SMIC	OSF and LBP-TOP	SVM	F-measure: 0.51 and 0.31 SMIC and CASME II
2018	Zhu et al. [172]	CASME II	LBP-TOP and OF	SVM	accuracy of 53.3%
2018	Zong et al. [174]	CASME II and SMIC	LBP-TOP, LBP-SIP and STLBP-IP	KGSL	accuracy: 63.9 on CASME II
2018	gan et al. [45]	CASME II and SMIC	BiVACNN	CNN	accuracy of 80% using 3 classes
2018	hu et al. [57]	CASME II	LGBP-TOP+CNN	CNN	accuracy of 66.2%
2018	Jia et al. [65]	CASME II and CK+	LBP and LBP-TOP	NN	65.5%
2018	Khor et al. [71]	CASME II and SAMM	OF	CNN	F1-Score=0.5 Accuracy=0.52
2018	Li et al. [85]	CASME II	CNN	CNN	63%
2018	Lin et al. [86]	CASME II and SMIC	Gabor filter	SVM	55.28
2018	Lu et al. [94]	CASME II	FMBH	SVM	69.11 %
2018	Xia et al. [145]	CASME II	CNN	CNN	65.8%
2018	Zhao and Xu [166]	CASME II	LBP-TOP	SVM	67.95%
2019	Allaert et al. [3]	CASME II	LMP	SVM	68.4%
2019	Gan et al. [46]	Cross-database	OF	CNN	74.6% 3 classes
2019	Li et al. [81]	CASME II	CNN	CNN	59.11%
2019	Liong et al. [87]	Cross-database	CNN	CNN	73.5 F1-score
2019	Peng et al. [112]	Cross-database	CNN	CNN	0.631
2019	Van Quang et al. [134]	Cross-database	CNN	CNN	0.6506
2019	Song et al. [127]	Cross-databse	CNN	CNN	73.8%
2019	Wang et al. [135]	CASME II	LBPTOP,HOOF	SVM	62.8%
2019	Xia et al. [146]	Cross-database	CNN	CNN	0.57
2019	Yu et al. [159]	SMIC	DCP	SVM	62.8%
2019	Zhi et al. [169]	CASME II	CNN	CNN	62.5%
2019	Zhou et al. [171]	CASME II and SAMM	CNN	CNN	0.75 and 0.48 3-classes
2019	Khor et al. [70]	CASME and SAMM	CNN	CNN	Accuracy=71.1 F1-score=0.71

2.3.3 Histogram of Oriented Optical Flow (HOOF)

Liu et al. [93] proposed Main Directional Mean Optical-flow (MDMO) as features for recognition micro-expression. Their MDMO consist of Regions of Interest (ROIs) based partially on AUs. One of the significant advantages of MDMO is the small features dimension, where the features vector length equal to 72 which is 2 features extracted from each region of 36 ROIs. Aligned all frames to the first frame has been applied to reduce the noise result from head movements. SVM classifier has been adopted for recognition. SMIC, CASME and CASME II datasets were used to evaluate their method. The result compared to the benchmark which used LBP-TOP and histogram of oriented optical flow (HOOF) features and achieve better result compare to benchmark which 68.86%, 67.37% and 80% on CASME, CASME II and SMIC respectively. Song et al. [128] used a Harris3D detector with combination of HOG and the Histograms of Oriented Optical Flow (HOOF) features, and used codebook to encode features in a sparse manner of micro-expressions. To predict expression they used Support Vector Regression (SVR) [126]. They evaluated this method on a subset of the SEMAINE corpus [101] dataset. Happy and Routray [51] they claimed that the changes on the face during a micro-expression is temporal changes more than spatial. Based on this claim they proposed temporal features descriptor called Fuzzy Histogram of Optical Flow Orientation (FHOFO) and it's an extension of HOOF. They evaluated their method on CASME, CASME II and SMIC. The best result was achieved in term of F1 score was 0.5489, 0.5248 and 0.5243 on the mentioned datasets respectively. In [52] They used Pair-wise feature proximity (PWFP) as features selection to improve the result in the previous study which has been slightly improved. To enhance micro-expression recognition Zhu et al. [172] transfer learning from speech to micro-expression and call their method coupled source domain targetized with updating tag vectors. LBP-TOP and OF have been used as features extractor with different vector dimension. They used SVM as classifier and evaluated their method on CASME II. The best accuracy of 53.3% achieved by OF with dictionary dimensions at 50. Lu et al. [94] proposed fusion of motion boundary histograms (FMBH) which generated by combing the horizontal and the vertical optical flow components, they evaluated their on CASME II and achieve a good accuracy which is 69.11 %. Su et al. [130] propose a ROI (Region of Interest)-based spatio-temporal feature named Dense Sampling Optical-flow's Mean Magnitude and Angle (DS-OMMA) for micro-expression recognition and they achieve accuracy up to 66.2% on CASME II.

2.3.4 Deep Learning Approaches

Over the past few years, deep learning approaches, such as convolutional neural networks (CNNs), have grown rapidly with a growing number of successful applications [79, 24]. A core feature of CNNs is the network architecture that produces the features to represent the input data. Popular architectures include LeNet [80], GoogLeNet [131] and AlexNet [76]. Many deep learning approaches focus on static images for classification, object detection or segmentation. Spatio-temporal based analysis methods using 3D CNNs are emerging with new applications, primarily on action recognition [64, 69, 160, 133].

As the datasets associated with these new methods are very large in number, for example the Sports-1M Dataset by Karpathy et al. [69], gaining discriminative data for 3D CNNs is a much easier task than collecting spontaneously induced micro-expressions. Therefore, there are very few approaches to detecting and recognising subtle motion using deep learning.

One of the first to use CNNs in micro-expressions analysis is by Kim et al. [73]. They proposed a new feature representation for micro-expressions where the spatial information at different temporal states (i.e. onset, apex and offset) are encoded using a CNN. This method used the extracted features attempted to help discriminate micro-expression classes when the model is passed to the long short-term memory (LSTM) recurrent neural network, where the temporal characteristics of the data are analysed. The overall achieved accuracy when comparing with the state-of-the-art was 60.98%, which is still relatively similar to many micro-expression recognition systems that only use accuracy for the evaluation metric. Further, the method only evaluated on single dataset, i.e. CASME II [150] dataset and does not consider more modern micro-expression datasets such as SAMM [18] and CAS(ME)^2 [119]. In 2017, Peng et al. [113] proposed a new method named Dual Temporal Scale Convolutional Neural Network (DTSCNN). Due to the data deficiency in available datasets, they designed a shallower neural network for micro-expression recognition with only 4 layers for both convolutional and pooling layers. As stated in its name, DTSCNN is a two streams network. The network has been fed with the optical-flow sequences. CASMEI/II datasets were used in the experiment and have been merged by the authors using selected data from both datasets, CASME I/II have been categorized into 4 classes: Negative, Others, Positive and Surprise. They achieved the best accuracy of 66.67%. Hao and Tian [50] used deep belief network (DBN) as the second stage features extractor to extract more global feature with less computation cost. DBN classification has been done by pre-training and fine-tuning in DBN. This was fused with the first stage local features was Weber Local Descriptor (WLD). However, their method only evaluated on a non-spontaneous dataset JAFFE [98], which was dated and difficult to compare with current literature. Gan et al. [45] proposed a deep learning method called BiVACNN for micro-expressions recognition. Their method contain three phases which are: apex detection, multi-features extraction, and learning, then classification into three categories(positive, negative and surprise) after combine CASME II and SMIC. They achieve an 80% accuracy rate. Some researchers try to combined conventional features (LGBP-TOP) with CNN features [57] and achieved 66.2% in terms of accuracy using LOSO cross-validation. One of the most challenges that faced micro-expressions recognition is the limited training samples Jia et al. [65] try to avoid this problem by transfer learning from macro to micro expressions. They used LBP and LBP-TOP for both expressions respectively. their method achieved accuracy 65.5%. Khor et al. [71] feed CNN with an optical flow frame using cross-database contain from CASME II and SAMM reach to performance equal to 0.5 and 0.52 in terms of F1-score and accuracy respectively. Li et al. [85] proposed to use DCNN with just apex frame after detect it and They achieved accuracy equal to 63% when they evaluated their method on CASME II. Other Deep learning works such as Xia et al. [145] achieve some improvements in ME recognition, but they are still significantly below state-of-the-art handcrafted features. Recently research on deep learning was increased some works achieved poor results. [169, 81], some of them used cross-database and classify three classes (positive, negative and surprise) [171, 146, 127, 134, 87, 46, 112].Khor et al. [70] proposed a lightweight dual-stream shallow network in the form of a pair of truncated CNNs with heterogeneous input features, their method have been validated on CASME II,SAMM and SMIC. They achieved result in terms of accuracy and F1-score(0.7119 0.7151,0.5735 0.4644,0.6341 0.6462) on three dataset respectively.

The review reflects the existing CNN-based methods faced similar problem in terms of data. Overall, micro-expression recognition using deep learning is still in its infancy due to a lack of available dataset. A large amount of data is crucial when training CNNs like many machine learning approaches. Micro-expressions are very complex and cannot easily be categorised into distinct classes as many approaches attempt to do [115, 153, 157]. Using 3D CNN features to understand the subtle movement would be a better approach to generalise the problem of discriminating a micro-expression on the face.

2.3.5 Other Feature Extraction Methods

Lu et al. [95] proposed Delaunay-Based Temporal Coding Model (DTCM) for Micro-Expression Recognition. Active Appearance Model (AAM) used to define facial feature points (68 points). Delaunay triangulation has been implemented based on the feature points. This process divides the facial area into number of sub-regions with triangle shape. Normalisation has been done based on standard face (neutral), this remove personal appearance difference irrelative. They used local temporal variations (LTVs) to code the features space, where the difference between mean of grayscale values of subregion and sub-region in neighbour frame were computed. Delaunay triangulation generates a large number of subregions which leads to large number of local features. To overcome this problem, they selected just subregions related to micro-expression, this selection based on standard deviation analysis. finally, the code sequences of all subregions concatenated into one feature vector. RF [9] and SVM have been used as classifiers. This method evaluated on SMIC, CASME class B and CASME II. They achieved better result than state of the art, with 82.86%, 64.95% and 64.19% on SMIC, CASME class B and CASME II respectively. Zhang et al. [162] developed micro-expression recognition system or visual platform as they claimed that there has not been much work done in designing these kind of systems. Their system includes two main parts: feature extraction and dimensional reduction, they used a gabor filter for feature extraction and principal components analysis (PCA) and LDA for dimension reduction. For classification stage, SVM has been used. To evaluate their system, CASME II and real-time videos were used. They claimed that the system have a good performance on static images counter to real-time videos. Gabor filter also been used by Wu et al. [144] but they evaluated the performance on Cohn and Kanade's dataset

(CK) [68], which was developed for facial expression analysis. Li et al. [82] evaluated the performance of three feature types (LBP, HOG and histograms of image gradient orientation (HIGO)) on two publicly available datasets (CASME II and SMIC). They extracted these three features from different planes. LSVM was employed as the classifier using LOSO validation. On CASME II, the best accuracy was 57.49%. This is achieved when they extracted HOG from both 3 orthogonal planes (HOG-TOP) and XT, YT planes (HOG-XYOT). On the other hand, three versions of SMIC were tested - VIS, NIR, HS and sub of HS, with the last one achieved the best accuracy and when features were extracted using HIGO-TOP and HOG-TOP. In addition, an effect of the interpolation length was tested with different frame lengths from 10 to 80 with fixed incremental steps of 10 frames. The best performance was achieved with an interpolation to 10 frames and it was 53.52%, 45.12% and 38.02% on SMIC-VIS, SMIC-HS and SMIC-NIR respectively. Huang et al. [60] outlined two problems of LBP-TOP. The first problem is LBP-TOP does not consider useful information, the second problem is the classical pattern used by LBP-TOP may not be good for describing local structure. To avoid those two problems, they proposed Spatio-Temporal Completed Local Quantization Patterns (STCLQP), which extracted sign, magnitude and orientation. In addition, a codebook were developed for each component in both appearance and temporal domains.Their method was evaluated on SMIC, CASME and CASME II with accuracy of 64.02%, 57.31% and 58.39%, respectively. Spatio-Temporal Texture Map (STTM) was developed by Kamarol et al. [67]. STTM used a modified version of Harris corner function [53] to extract the micro-expression features. This method evaluated on CASME II, and compared with other features (Volume Local Binary Pattern (VLBP) and LBP-TOP). They used SVM with one-against-one classification between four classes. In terms of accuracy, the average recognition rate of STTM performed slightly better than the other features which is reached to 91.71% in contrast LBP-TOP achieved 91.48%. On the other hand, in terms of computation time, there is a large difference between STTM and other features, where STTM process one frame in 1.53 seconds in contrast to 2.57 and 2.70 seconds for VLBP and LBP-TOP respectively. Wang et al. [136] introduced a micro-expression algorithm called discriminant tensor subspace analysis (DTSA). This method was evaluated on the CASME dataset. Extreme learning machine (ELM) was used as classifier. They have tested the method with various optimal dimensionality and different sets of training and testing. The best accuracy, 46.90%, was achieved when dimensionality was set to 40×40×40 and the training sample is 15. Maximum margin projection with tensor representation (MMPTR) is a micro-expression recognition algorithm contributed by Ben et al. [7]. They tested their algorithm on CASME. The best average recognition rate, which is 80.2% was achieved on tensor size of 64×64×64 and training sample was same as [136] with 15 samples. Liong et al. [91] questioned whether all frames of micro-expressions need to be processed for effective analysis. They used only the apex and the onset frame for experiments to test this theory. The frames were extracted using their proposed Bi-Weighted Oriented Optical Flow (Bi-WOOF). These features were then evaluated on CASME II and the three formats of SMIC. The best performance achieved on CASME II and SMIC-HS in terms of accuracy was 0.61 and 0.62 respectively. Liong et al. [89] proposed two sets of features: optical strain and optical strain weighted. These two features constructed by utilising facial optical strain magnitudes. They performed the features on the CASME II and SMIC and they overcame the baseline of two datasets [83, 150] with recognition rate reach to 52.44% on SMIC and 63.16% on CASME II. Oh et al. [103] claimed that there is changes on facial contour which are located in different part of face are crucial for the recognition micro-expressions. According to that they proposed a feature extraction method to represent these changes called Intrinsic Two-Dimensional local structuresm (I2D). This method was evaluated on the CASME II and SMIC dataset.The result was better than two state of the art [83, 150] with the best F1-score of 0.41 and 0.44 on CASME II and SMIC respectively. Sparse Tensor Canonical Correlation Analysis (STCCA) was proposed by Wang et al. [139] to improve the recognition rate of micro-expressions. They conducted the experiment on CASME and CASME II. They proved that their method can perform better than 3D-Canonical Correlation Analysis and three-order Discriminant Tensor Subspace Analysis. In addition to that they proved that Multi-linear Principal Component Analysis is not suitable for micro-expression recognition. Zheng et al. [168] proposed a a relaxed K-SVD classifier (RK-SVD) and tested it on LBP-TOP and HOOF features to be used for micro-expression recognition. They evaluated this proposed classifier on CASME and CASME II, and compared it with different classifiers such as SVM, MKL and RF. The results was better than other classifiers for both features and on two datasets [154, 150] with best accuracy of 69.04% and 60.82% for LBP-TOP and HOOF respectively on CASME, and on CASME II the accuracy was 63.25% 58.64% for the same features respectively. Zheng [167] proposed a method for micro-expression recognition named 2D Gabor filter and Sparse Representation (2DGSR). They evaluated their method on three publicly available datasets (SMIC, CASME and CASME II) and compared it with other popular methods (LBP-TOP, HOOF-whole and HOOF-ROIs). For classification Sparse Representations Classifier (SRC) has been used with LOSO cross validation. In terms of accuracy they achieved a result up to 71.19% and 64.88% on CASME and CASME II respectively. Gabor filter also used by Lin et al. [86] as spatiotemporal features but they didn't achieve high result. Ben et al. [6] proposed local binary feature descriptor called hot wheel patterns from three orthogonal planes (HWP-TOP) which has been inspired by dual-cross patterns from three orthogonal planes (DCP-TOP) with some rotations. They used smooth SVM (SSVM) as a classifier. They evaluated their descriptor on 61 samples from CASME II with three classes(except fear and sadness) and achieved recognition rate of 0.868. They try to solve the problem of micro-expression limited samples by leverage labeled macro-expression and shared feature between macro and micro expression, however this may be not so accurate due the difference between macro and micro characteristic. After extracting features using distance estimation between points which have been predicted using ASM Jain et al [62] using Random Walk-based (RW) to learn the features before providing it to Artificial Neural Network (ANN) classifier. RW reduces the dimensionality of the feature and this minimize the complexity of computation. They evaluated their method on CASME and SMIC and provide the result in term of AUC, which is up to 0.8812 and 0.9456 on SMIC and CASME respectively. Allaert et al. [3] proposed a method called LMP and they evaluated it CASME II achieving a promising result up to 68.4 %. While Yu et al. [159] and their method DCP didn't achieved a good result.

2.4 Performance Metrics and Validation Techniques

The spotting accuracy of humans peaks around 40% [42]. Analysis using computer algorithms incorporating machine learning and computer vision can only be evaluated fairly with a standardised metrics. This section elaborates the metrics used in the literature. Drawing from detailed review in Section 3, we summarised and explain the evaluation metrics.

2.4.1 Metrics

The metrics for micro-expressions analysis are commonly used for binary classification purposes, and so is adequate for quantifying *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)* and *False Negative (FN)* detections. More detailed information on these measures can be found in [5]. The earlier work, as illustrated in Table 2.6, the majority of the results in micro-expressions analysis are based on *Accuracy*. as defined in equation 2.1 [5].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2.1}$$

In the later stage, as illustrated in Table 2.7, the measurement of performance were reported in *F1-Score (or F-Measure)*. Other metrics such as *Recall*, *Precision*, and *Matthews Correlation Coefficient* (*MCC*) are also gradually used to report the results. By using the

Precision measure of exactness, and determines a fraction of relevant responses from results. Recall, or sensitivity, is a fraction of the results that are relevant to the experiment and that are successfully retrieved [5].

$$Precision = \frac{TP}{TP+FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.3}$$

It is unlikely to use these measures on their own as both these measure are commonly used together to form an understanding of the relevance of the results returned from experimental classification. The F-Measure is useful in determining the harmonic mean between the *Precision* and *Recall* and is used in place of accuracy as it provides a more detailed analysis of the data. The equation can be defined as [5]

$$F\text{-}Measure = \frac{2TP}{2TP+FP+FN}. \tag{2.4}$$

A downside to this measure is that it does not take into account *TN*, a value that is required to create *ROC* curves. The *MCC* uses all detection types to output a value between -1, which indicates total disagreement and $+1$, which indicates total agreement. A value of 0 would be classed as a random prediction, and therefore both variables can be deemed independent. It can be provide a much more balanced evaluation of prediction than previous measurements, however it is not always possible to obtain all four detection types (i.e. *TP*, *FP*, *FN*, *TN*). The coefficient can be calculated by [5]

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.5}$$

2.4.2 Validation Techniques

Two commonly used validation techniques in computer vision are *n-fold* cross validation and leave-one-subject-out (LOSO). From our review, the evaluation system by different researchers reported in different validation techniques, where LOSO is more widely used. While some reported their results in both validation techniques [93, 18], and some only reported in LOSO [73, 168, 114].

2.5 Challenges

Research on automated micro-expressions recognition using machine learning has witnessed good progress in recent years. A number of promising methods based on texture

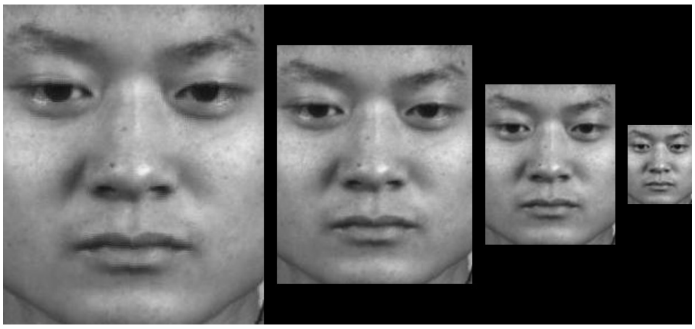


Figure 2.5: An example of different resolution by downscaling an image from CASME II dataset. From left to right: 100% (Original resolution), 75% of the original resolution, 50% of the original resolution and 25% of the original resolution.

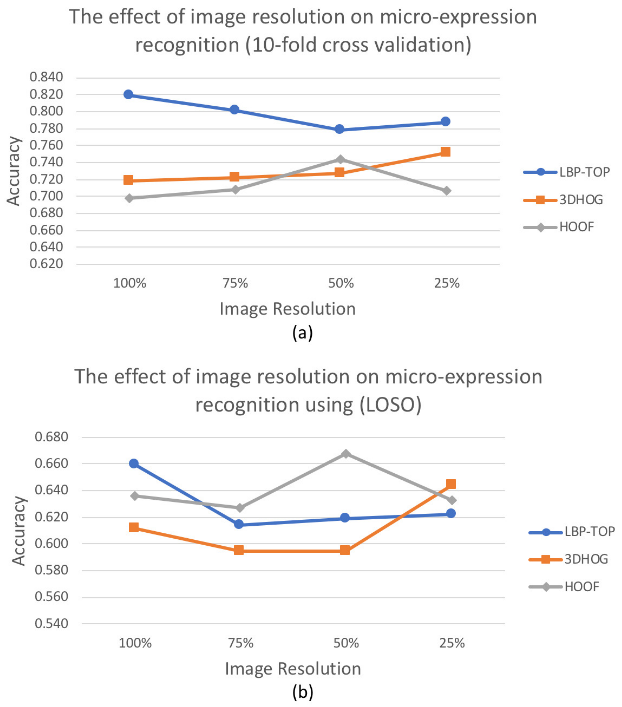
features, gradient features and optical flow features have been proposed. Many datasets was generated but lack of standardisation is indeed a great challenge. This section provides the challenges of the research in micro-expressions analysis into details.

2.5.1 The effect of Spatial Temporal Settings in Data Collection

Due to lack of communication between different research groups on experimental settings, the datasets are varied in resolution and frame rates. Some researchers [83, 82] investigated on the effect of Temporal setting to micro-expression recognition. Using TIM [114] to adjust the temporal settings is a well-known method in micro-expression analysis. However, there is a lack of thorough research in further investigating the implication of spatial-temporal changes for micro-expression recognition.

We believe resolution plays an important role for features extraction. We downscale the CASME II dataset to four scales, 100% (original resolution), 75% of the original resolution, 50% of the original resolution and 25% of the original resolution, as depicted by Figure 2.5. To address the research gap, we experiment this four resolutions with three feature types (LBP-TOP, 3DHOG and HOOF) with *10-fold* cross validation and LOSO. To reduce the effect of learning algorithm, we used a standard SVM method as the classifier. Figure 2.6 compares the performance of the experiments.

From the observation, LBP-TOP performed better in high resolution images than 3DHOG and HOOF. It is noted that HOOF performed better when we downscale the resolution to 50% and 3DHOG worked best at 25%. These results showed LBP-TOP relied on spatial information (XY), but HOOF and 2DHOG are more dependent on temporal (XT and YT). The conventional methods are relies on feature descriptors and varies from one to another.


The effect of image resolution on micro-expression recognition (10-fold cross validation)

(a)

The effect of image resolution on micro-expression recognition using (LOSO)

(b)

Figure 2.6: The effect of image resolution on micro-expression recognition using LBP-TOP, 3DHOG and HOOF on two different evaluation method: (a) 10-fold cross validation, and (b) LOSO.

Table 2.8: A breakdown of the number of clips categorised into estimated emotion classes for the SAMM dataset.

Estimated Emotion	Number of Clips
Anger	57
Contempt	12
Disgust	9
Fear	8
Happiness	26
Sadness	6
Surprise	15
Other	26

2.5.2 Emotional Classes versus Objective Classes in Data Labelling

A large focus on micro-expression research has been on the detection and recognition of emotion-based classed (i.e. discreet groups that micro-expression fit into during classification). Objective classes attempt to take away the potential bias of labelling difficult to distinguish micro-expressions into classes suited to a particular muscle movement pattern.

To date, SAMM [18] is the only dataset that moves the focus from an emotional-based classification system, to an objective one, and is designed around analysing objective physical movement of muscles. Emotion classification requires the context of the situation for an interpreter to make a meaningful interpretation. Most spontaneous micro-expression datasets have FACS ground truth labels and estimated or predicted emotion. These have been annotated by an expert and self-reports written by participants. In SAMM, Davison et al. [18] focused on objectiveness and did not report emotional classes in their dataset release. Due to this reason, it has not been widely experimented by other researchers. To address this issue, we introduced the emotional classes for SAMM in this chapter.

SAMM has estimated emotional classes based on the AUs and the emotional stimuli presented to participants to allow for comparison with previous emotion class focused papers such as CASME II [150] and SMIC [83]. The amount of clips in the SAMM dataset in each estimated emotion class can be seen in Table 2.8. Note that the categories are based around EMFACS labelling of reliable AUs to emotion [36], so any that did not fit into these categories are placed in the 'Other'class.

To this end it can be argued that keeping classification to well-defined muscles (that cannot be changed or bias) is a more optimal solution to micro-expression recognition than discreet emotion classes. Further, Yan et al. [151] state that it's inappropriate to categorise micro-expressions into emotion categories, and that using FACS AU research to inform the eventual emotional classification would be a more logical approach. In 2017, Davison et

Table 2.9: Each class represents AUs that can be linked to emotion.

Class	Action Units
I	AU6, AU12, AU6+AU12, AU6+AU7+AU12, AU7+AU12
II	AU1+AU2, AU5, AU25, AU1+AU2+AU25, AU25+AU26, AU5+AU24
III	A23, AU4, AU4+AU7, AU4+AU5, AU4+AU5+AU7, AU17+AU24, AU4+AU6+AU7, AU4+AU38
IV	AU10, AU9, AU4+AU9, AU4+AU40, AU4+AU5+AU40, AU4+AU7+AU9, AU4 +AU9+AU17, AU4+AU7+AU10, AU4+AU5+AU7+AU9, AU7+AU10
V	AU1, AU15, AU1+AU4, AU6+AU15, AU15+AU17
VI	AU1+AU2+AU4, AU20
VII	Others

al. [19] proposed new objective classes based on FACS coding. They have coded the two state-of-the-art FACS-coded datasets into seven objective classes as illustrated in Table 3.2. The objective classes were used for the first FME grand challenge conducted in 2018 [156].

2.5.3 Face Regions in Data Analysis

Recent work on the micro-expressions recognition have provided promising results on successful detection techniques, however there is room for improvement. To begin detection, current approaches follow methods of extracting local feature information of the face by splitting the face into regions, as illustrated in Figure 2.7.

The state of the art can be categorised into:

1. *Four quadrants*. Shreve et al. [123] split the face into 4 quadrants and analyse each quarter as individual temporal sequences. The advantage of this method is that it is simple to analyse larger regions, however the information to retrieve from the areas are restricted to whether there was some form of movement in a more global area.

2. *$m \times n$ blocks*. Another method is to split the face into a specific number of blocks [150, 21, 22]. The movement on the face is analysed locally, rather than a global representation of the whole face, and can focus on small changes in very specific temporal blocks. A disadvantage to this method is that it is computationally expensive to process the whole images as $m \times n$ blocks. It can also include features around the edge of the face, including hair, that do not relate to movement but could still effect the final

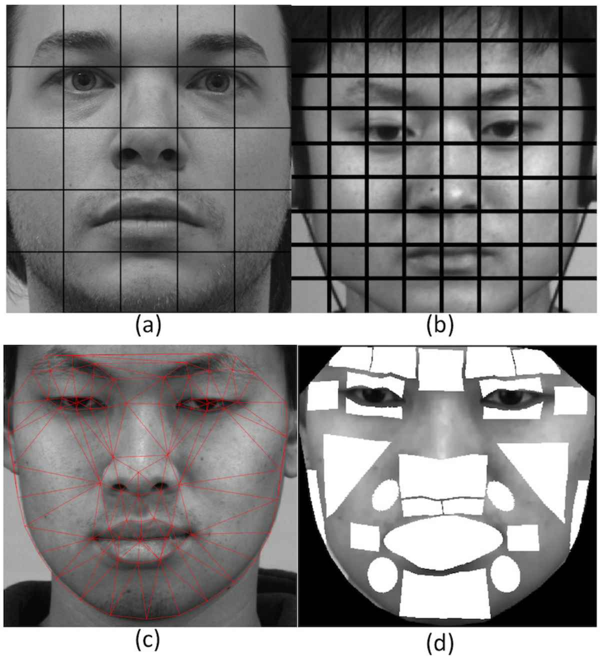


Figure 2.7: Illustration of face regions: (a) 5 × 5 blocks, (b) 8 × 8 blocks, (c) Delaunay triangulation, and (d) FACS-based regions.

feature vector. Figure 2.7(a) and Figure 2.7(b) illustrate the samples of block-based face regions.

3. *Delaunay triangulation.* Delaunay triangulation, as shown if Figure 2.7(c), has also been used to form regions on just the face and can exclude hair and neck [95], however this approach can still extract areas of the face that would not be useful as a feature and adds further computational expense.

4. *FACS-based region.* A more recent and less researched approach is to use defined regions of interest (ROIs) to correspond with one or more FACS AUs [137, 138]. These regions have more focus on local parts of the face that move due to muscle activation. Some examples of ROI selection for micro-expression recognition and detection include discriminative response map fitting [93], Delaunay triangulation [95] and facial landmark based region selection [111]. Unfortunately, currently defined regions do not cover all AUs and miss some potentially important movements such as AU5 (Upper Lid Raiser), AU23 (Lip Tightener) and AU31 (Jaw Clencher). To overcome the problem, Davison et al. [19] proposed FACS-based regions to improve local feature representation by disregarding face region that do not contribute to facial muscle movements. The defined region is presented in Figure 2.7(b).

Figure 2.7 compares different face region splitting methods. Due to FACS-based region is more relevant to facial muscle movements and suitable for AUs detection, more research should be focusing on FACS-based region than split the face into $m \times n$ blocks.

2.5.4 Deep Learning versus Conventional Approaches

The pipeline of conventional micro-expression recognition approach is very similar to macro-expressions in terms of preprocessing techniques, hand-crafted features and, if applicable, machine learning classification. However, geometric feature-based methods are rarely used as tracking feature points on a face that barely moves will not produce good results. Instead, appearance-based features are primarily used to attempt to describe the micro-movement or train machine learning to classify micro-expressions into classes.

Spatial temporal settings during data collection, preprocessing stage of dataset including face alignment and face regions split, feature extraction methods and the type of classifiers are the main factors for conventional approaches. Moving forward, end-to-end solution that is capable of handling these issues is required. Deep learning approaches have yet to have much impact on micro-expression analysis, however to ensure a rounded review of current techniques we shall provide a preliminary study on deep learning and its applications to micro-expression.

As the temporal nature of micro-expressions are a key feature to understand, modern video-analysis technique, namely 3D convolutional neural networks (3D ConvNets) [133], may be used to exploit the temporal dimension. This network expands on the typical 2D convolutional neural network (CNN) by using $3 \times 3 \times 3$ convolutional kernels where the third dimension is in the temporal domain (frames in a video). It was originally used for analysis for action recognition, however it can be expanded for any other video-analysis task easily. Using the deconvolution method described by Zeiler and Fergus [161], Tran et al. [133] was able to show that the features extracted from the 3D ConvNet focuses on the appearance of the first few frames and then tracks salient motion over the next frames. The key difference in using 2D ConvNets is the ability to extract and learn from features from both motion and appearance.

With minimal data available to train from, deep learning methods have a much more difficult time in learning meaningful patterns [24]. When independent test samples were used for validation, the model showed that further investigation is required for deep learning with micro-expression to be effective, including the use of more data. The biggest disadvantage to using video-data is not being able to load such large amounts of data into memory, even on GPUs that have 12GB of on-board memory. This leads to the minimisation of the batch size and reduction of resolution to allow for training to proceed. Further ways of being able to handle micro-expression data without having to reduce the amount of data available would be vital to retaining the discriminative information required for micro-expression analysis. Further, the time required to train the model shows the challenge of the ability to train long video-based deep learning methods.

2.5.5 Standardisation of Metrics

We recommend the researchers to standardised the performance metrics that they used in evaluation. As the majority of datasets are inbalanced [78], reporting the result in *F-Measure* (or *F1-Score*) seems to be the best option. Using the conventional *Accuracy* measure may result in a bias towards classes with large number of samples, hence over-estimating the capability of the evaluated method. F-Measure micro-average across the whole dataset and is computed based on the total true positives, false negatives and false positives, across 10-fold cross validation and Leave-one-subject-out (LOSO).

Due to each dataset with small micro-expression samples, the researchers are encourage to use more datasets for their experiment. For cross datasets evaluation, unweighted average recall (UAR) and weighted average recall (WAR) are recommended as these measurements were shown promising in speech emotion recognition [121]. WAR is defined as number of correctly classified samples divided by the total number of samples, while UAR is defined as sum of accuracy of each class divided by the number of classes without considerations of samples per class. To obtain the overall scores, the results from all the folds are averaged. These metrics had been recommended in the First Micro-expressions Grand Challenge Workshop in conjunction with Face and Gesture 2018 Conference [156].

2.5.6 Real-world Implementation

For implementation of the micro-expressions recognition in real-world, the challenges to be addressed include:

1. **Cross-Cultural Analysis** Micro-facial expressions occur when people attempt to hide their true emotion, and so the possibility of how well some cultures manage this suppression would be interesting to learn. By using software to detect micro-expressions across cultures, the results of different suppression of emotion can be studied. Therefore people in East Asian cultures could be different from Western cultures, which can be analysed to find any correlation between the psychological studies and automated micro-expressions recognition. Something to note in this type of investigation would be to ensure the different participants originate and live in

their respective countries, as people living with different cultures for a long time may not exhibit the same behaviour.

2. **Dataset Improvements**. Further work can be done to improve micro-movement datasets. Firstly, more datasets or expanding previous sets would be a simple improvement that can help move the research forward faster. Secondly, a standard procedure on how to maximise the amount of micro-movements induced spontaneously in laboratory controlled experiments would be beneficial. If collaboration between established datasets and researchers from psychology occurred, dataset creation would be more consistent. As using human participants is required, and emotions are induced, ethical concerns are always going to play a part in future studies of this kind. Any work moving forward must take into account these concerns and draw from previous experiments to ensure no harm will come to the psychological welfare of participants.

3. **Real-Time Micro-Facial Expressions Recognition**. To be able to implement any form of micro-movement detection system into a real-world scenario, it must perform the processes required in real-time (or near to real-time). As the accuracy of facial expression analysis is already quite high, transitioning to real-time has already produced decent results. However there is currently no known systems that is able to detect micro-expressions.

The accuracy of many state-of-the-art methods is still too low to be deployed effectively in a real-world environment. The progress in research of micro-expressions recognition can aid in the paradigm shift in affect computing for real-world applications in psychology, health study and security control.

2.6 Summary

We have presented a comprehensive review on datasets, features and metrics for micro-expressions analysis. To summarise, the future direction to advance automated micro-expression recognition should take into consideration on how the dataset is capture (spatial temporal settings), labeling of the dataset based on Action Unit based objective classes, FACS-based face regions for better localisation, end-to-end solution using deep learning, fair evaluation using standardised metrics (ideally F1-Score and MCC) and LOSO as the validation technique. More importantly, the openness and better communication within the research communities are crucial to crowd-source the data labelling and using the standard evaluation system.

Chapter 3

Methodology

3.1 Introduction

This chapter focuses on the theories and techniques used throughout the thesis and the methodology that has been followed in the thesis contributions. The chapter also will introduce some of the most popular features descriptions for FME recognition. It also will describe the classifier that is used in this research. In addition to the Facial Action Coding System (FACS)

3.2 FME Recognition Pipeline

The process of recognition FMEs in conventional computer vision usually involves:

1. Preprocessing.

2. Feature description.

3. Classification.

3.2.1 Preprocessing

In FME recognition usually need convert colored image into gray-scale due most of features extraction algorithms work on one channel rather than three channels of RGB image to reduce computation and time, also removing noise and correcting brightness can be involved as preprocessing step especially in FME's researches because the subtle nature of FME be affected by the noise and this could affect in turn the final result.One of the most important preprocessing is face alignment, which is applied to all frames of micro-expression so that all the faces are in the same position based on a constant reference point and this step comes after cropping the facial area because it is the area of interest for FME.

3.2.2 Features description

In this section, the features extraction algorithms which have been used in the research experiments will be described. Due to the dynamic nature of FME (video datasets), just some of an algorithms could extract features from FME videos. These algorithms could be classified into three main categories which are; texture-based algorithms presented by Local Binary Pattern Three Orthogonal Planes(LBP-TOP), gradient-based presented by Histogram of Oriented Gradients on 3D (HOG3D) and optical flow-based which presented by Histogram of Optical Flow (HOOF).

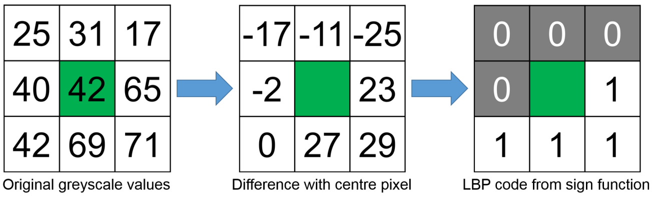
3.2.2.1 Local Binary Pattern-Three Orthogonal Planes(LBP-TOP)

The LBP operator forms labels for each pixel in an image by thresholding a 3×3 neighbourhood of each pixel with the centre value. The result is a binary number where if the outside pixels are equal to or greater than the centre pixel, it is assigned a 1, otherwise it is assigned a 0. The amount of labels will therefore be $2^8 = 256$ labels.

This operator was extended to use neighbourhoods of different sizes. Using a circular neighbourhood and bilinearly interpolating values at non-integer pixel coordinates allow any radius and number of pixels in the neighbourhood. The grey-scale variance of the local neighbourhood can be used as the complementary contrast method. The following notation of (P,R) will be used for pixel neighbourhoods, where P are sampling points on a circle of radius R. Fig. 3.1 shows an example of LBP computation.

Uniform patterns can be used to reduce the length of the overall feature vector and implement a single rotation-invariant descriptor. An LBP that is uniform when the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. So 00000000 (0 transitions), 01110000 (2 transitions) and 11001111 (2 transitions) are uniform whereas the patterns 11001001 (4 transitions) and 01010010 (6 transitions) are not. In the computation of the LBP labels, uniform patterns are used so that there is a separate label for each uniform pattern and all the non-uniform patterns are labelled with a single label. For example, when using $(8,R)$ neighbourhood, there are a total of 256 patterns, 58 of which are uniform, which yields in 59 different labels.

Based on the LBP operator, LBP-TOP was first described as a texture descriptor [165] that used XT and YT temporal planes rather than just the 2D XY spatial plane. Yan et al. [150] used this method to report initial findings in the CASME II dataset, and Pfister et al. [115] and Davison et al. [21] used it as feature descriptors in their work.



25	31	17
40	42	65
42	69	71

Original greyscale values

-17	-11	-25
-2		23
0	27	29

Difference with centre pixel

0	0	0
0		1
1	1	1

LBP code from sign function

Figure 3.1: LBP code calculation by using the difference of the neighbourhood pixels around the centre.

Each region has the standard LBP operator applied [105] with c being the centre pixel and P being neighbouring pixels with a radius of R [165]

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{3.1}$$

where g_c is the grey value of the centre pixel and g_p is the grey value of the p-th neighbouring pixel around R. 2^p defines weights to neighbouring pixel locations and is used to obtain the decimal value. The sign function to determine what binary value is assigned to the pattern is calculated as [165]

$$s(\mathbf{A}) = \begin{cases} 1, & \text{if } \mathbf{A} \geq 0 \\ 0, & \text{if } \mathbf{A} < 0 \end{cases} \tag{3.2}$$

If the grey value of P is larger than or equal to c, then the binary value is 1, otherwise it will be 0. Fig. 3.1 illustrates the sign function on a neighbourhood of pixels. After the image has been assigned LBP, the histogram can be calculated by [165]

$$H_i = \sum_{x,y} I\{LBP_l(x,y) = i\}, i = 0, \ldots, n-1 \tag{3.3}$$

where $LBP_l(x,y)$ is the image labelled with LBP. As this method is incorporating temporal data, the histogram can be extended to be calculated for all three planes [165]

$$H_{i,j} = \sum_{x,y,t} I\{LBP_j(x,y,t) = i\}, i = 0, \ldots, n_j - 1 \tag{3.4}$$

where n_j is the number of labels produced by the LBP operator in the jth plane. $j = 0,1,2$ which represents the XY, XT and YT planes respectively. $LBP_i(x,y,t)$ expresses the LBP code of the central pixel (x,y,t) in the jth plane. The $I\{\mathbf{A}\}$ function is the equivalent to Eq. 3.3 that refers to the sign function in Eq. 3.2. An illustration of the LBP-TOP histogram concatenation process can be seen in Fig. 3.2.

The neighbouring points and radius parameters (P,R) can be defined as $P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T$ for each plane and axis, with the overall feature descriptor defined as $LBPTOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$. We chose to use the best case results from [150] and set the neighbouring points and radii parameters to $LBPTOP_{4,4,4,1,1,4}$.

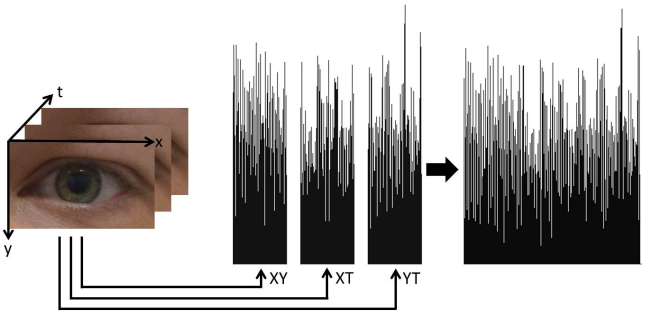


Figure 3.2: LBP is calculated on every block in all three planes. Each plane is then concatenated to obtain the final LBP-TOP feature histogram.

3.2.2.2 Histograms of Oriented Gradient on 3D (HOG3D)

Histograms of Oriented Gradient (HOG3D) [75] is adapted version of histograms of oriented gradient (HOG) for static images to be suitable for dynamic texture in micro-expressions. 2D can be represent as I(x,y), orientation and gradient could be calculated [118][20] as follows [75]

$$\begin{aligned} m_{2D}(x,y) &= \sqrt{\delta I_x(x,y)^2 + \delta I_y(x,y)^2} \\ \theta_{2D}(x,y) &= tan^{-1}(\delta I_y(x,y)^2/\delta I_x(x,y)^2) \end{aligned} \tag{3.5}$$

where $\delta I_x(x,y)$ and $\delta I_y(x,y)$ stand for image partial derivative. In 3D case the video v(x,y,t) where t refer to time, firstly partial derivative should be calculated along x,y and t. then compute magnitude $m_xy(x,y,t)$,$m_xt(x,y,t)$ and $m_yt(x,y,t)$ and orientation $\theta_xy(x,y,t)$,$\theta_xt(x,y,t)$ and $\theta_yt(x,y,t)$ for each couple $(\delta v_x, \delta v_y)$, $(\delta v_x, \delta v_t)$ and $(\delta v_y, \delta v_t)$ using equation 3.6 [118]

$$\begin{aligned} m_{xy}(x,y,t) &= \sqrt{\delta v_x(x,y,t)^2 + \delta v_y(x,y,t)^2} \\ \theta_{xy}(x,y,t) &= tan^{-1}\left(\frac{\delta v_x(x,y,t)^2}{\delta v_y(x,y,t)^2}\right) \\ m_{yt}(x,y,t) &= \sqrt{\delta v_y(x,y,t)^2 + \delta v_t(x,y,t)^2} \\ \theta_{yt}(x,y,t) &= tan^{-1}\left(\frac{\delta v_y(x,y,t)^2}{\delta v_t(x,y,t)^2}\right) \\ m_{xt}(x,y,t) &= \sqrt{\delta v_x(x,y,t)^2 + \delta v_t(x,y,t)^2} \\ \theta_{xt}(x,y,t) &= tan^{-1}\left(\frac{\delta v_x(x,y,t)^2}{\delta v_t(x,y,t)^2}\right) \end{aligned} \tag{3.6}$$

gradient orientation histograms are computed for every frame for the couples $(\delta v_x, \delta v_y)$ gradient orientation histogram contains 8 bins $(\delta v_x, \delta v_t)$ and $(\delta v_y, \delta v_t)$ contains 12 bins.

After computing histograms in $(\delta v_x, \delta v_y)$ $(\delta v_x, \delta v_t)$ and $(\delta v_y, \delta v_t)$ for every frame, all histograms corresponding to the same frame are concatenated to one feature vector and normalized.

3.2.2.3 Histogram of Oriented Optical Flow (HOOF)

HOOF [10] features compute optical flow for each frame, then the vector binned according to the orientation and weighted according to the magnitude, where each optical flow consist of pair of angle and magnitude and can be represented as $v = [x, y]^T$ with direction $\theta = tan^{-1}(\frac{y}{x})$. Fig. 3.3 explain how to build HOOF feature with 4 bins.

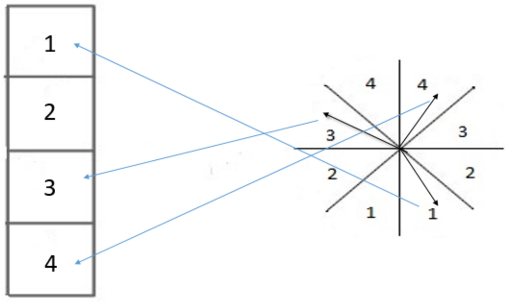


Figure 3.3: Build HOOF features with 4 bins

One of the HOOF-based methods is Main Directional Mean Optical Flow (MDMO) [93] which is a ROI-based normalised statistic feature. Discriminative Response Map Fitting (DRMF) [4] used to locate 68 facial feature points, 66 of them have been used (two inner corner points of lip ignored) to normalize faces base on the first frame. Normalized face has been partitioned to 36 regions of interest (ROIs) determined by 66 feature points and partially based on FACS. Using optical flow the change in intensity between two pixels has been detected between two frames over time motion of objects. Changing in intensity can be represented by [93]

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t) \tag{3.7}$$

the optical flow value of a pixel between two frames at time t in Euclidean coordinates can be represented as a two-dimensional vector [93]

$$[V_x^t, V_y^t]^T \tag{3.8}$$

To compute MDMO feature for micro-expression recognition the Euclidean coordinates $[V_x^t, V_y^t]^T$ has been converted into polar coordinates $(\rho i, \theta i)$, where ρi is the magnitude and θi is orientation of the optical flow vectors. Histogram of oriented optical flow

(HOOF) [10] has been computed for each ROI in each frame R_i^k, where i the index of frames and k is the index of ROIs and the optical has been classified into 8 bins. A mean vector $\overline{u}_i^k$ has been computed for optical flow vectors in bin with maximum count. $\overline{u}_i^k = (\overline{\rho}_i^k, \overline{\theta}_i^k)$, $\overline{\theta}_i^k$ called *main direction*. A feature vector Ψ_i has been built by $\Psi_i = (\overline{u}_i^1, \overline{u}_i^2, \ldots, \overline{u}_i^{36})$, this make The dimension of feature vector is $36 \times 2 = 72$, where 36 is the number of ROIs. Then micro-expression represented by concatenated features vector of each frame $\Gamma = (\Psi_1, \Psi_2, ..., \Psi_n)$, where n the number of frames in micro-expression. Finally, a normalisation in Cartesian coordinate has been done before converted back to polar coordinates and represented by [93]

$$\overline{\Psi} = [(\overline{\rho_1}, \overline{\theta_1})^T, (\overline{\rho_2}, \overline{\theta_2})^T, ..., (\overline{\rho_{36}}, \overline{\theta_{36}})^T] \tag{3.9}$$

In this experiment, block-based HOOF has been used and the parameters was set to $pRow = 6$, $pCol = 6$ and $pFrames = 6$ where $pRow$, $pCol$ and $pFrames$ is sub-block size in pixels for the row, column and frame respectively. The number of blocks was set to 3×3 spatial blocks and 2 temporal blocks, Horn-Schunck method has been used for optical flow computing and the last parameter is quantization, which is set to 8 orientations.

3.2.3 Classification

3.2.3.1 Support vector machine(SVM) using sequential minimal optimization (SMO)

First proposed by Cortes and Vapnik [16], a Support Vector Machine (SVM) attempts to find a linear decision surface (hyperplane) that can separate classes and has the largest distance between support vectors (elements in data closest to each other across classes). If a linear surface does not exist, then an SVM is able to use kernel functions to map the data into a higher dimensional space where a decision surface can be found.

We use the Sequential Minimal Optimization (SMO) [116] algorithm to train the SVMs. SMO is able to break down large quadratic programming problems into a series of the smallest possible problems, which are solved analytically and avoids using a time-consuming numerical quadratic programming optimisation as an inner loop. SMO is also able to handle large training sets and is one of the computationally fastest methods of evaluating linear SVMs.

3.3 The Implication of Spatial Temporal Changes on Facial Micro-Expression Analysis

3.3.1 Overview

Facial micro-expression datasets lack consistency and standardisation, with different research groups using various experimental settings, in particular, where the datasets are varied in resolution and frame rates. To provide new insights into the roles of frame rate and resolution, we conduct an investigation into the use of different frame rates and resolution on current benchmark datasets (SMIC and CASME II). By using Temporal Interpolation Model, we subsample SMIC (original frame rate is 100 fps) to 50 fps and CASME II (original frame rate is 200 fps) into 100 fps and 50 fps. In addition, the resolution settings are adjusted to three scaling factors: 100% (original resolution), 75% and 50%. Three feature types are used to test the performance of these settings, which are Local Binary Patterns in Three Orthogonal Planes, 3D Histograms of Oriented Gradient and Histogram of Oriented Optical Flow. The results showed that the frame rate and resolution could affect the performance of micro-expression recognition, which behave distinctively dependent on feature types. This work provides new guidelines for future research in selecting frame rate, resolution and feature descriptors in micro-expressions recognition. There are a limited amount of datasets available for facial micro-expressions (henceforth micro-expressions) analysis, and the ones that do exist vary in standards, especially with the frame rates and resolution chosen for capturing the videos. Early datasets were created with low specification such as low resolution and frame rate. Recently with new and advanced technologies for capturing and gathering datasets, researchers start to create high quality dataset. The non-publicly available datasets include the USF-HD [125], and the Polikovsky dataset [118] which have a frame rate of 29.7 and 200 fps respectively. The publicly available datasets include the CASME [154] dataset using 60 fps, the SMIC [83] dataset using 100 fps, CASME II [150] using 200 fps, SAMM [18] using 200 fps and CAS(ME)^2 [119] using 30 fps. The researchers in this field have been collecting data using different settings for frame rate, resolution, experimental design, with or without stimuli, lighting condition and camera model. While some suggested high frame rate [150, 18], the most recent work [119] in this field still use a low frame rate. The question that arises here is that are these high quality datasets needed to improve micro-expressions analysis, and among those different standards which is the best?

To address the above question, we provide new insights of the implication of spatial temporal changes on micro-expression recognition by conducting a comparative study using the most popular feature types on two high frame rate and popular benchmark datasets, i.e. SMIC and CASME II. First we review the relevant spatial temporal work in micro-expression recognition outline our method in generating various frame rates and resolution. Then we summarise the three basic feature descriptors and a classifier used for this work. Finally, we present the results and discuss the future work.

3.3.2 Frame Rate Subsampling

To subsample each micro-expression video clip to different frame rates, we use a Temporal Interpolation Model (TIM) [114]. This uses graph embedding to interpolate at random points with the micro-expression clips. This method allows for a more statistically stable feature extraction when reducing the original frame rate of SMIC and CASME II.

A micro-expression video is seen as a set of images sampled along a curve, and a continuous function is created in a low-dimensional manifold by representing the video as a path graph P_n with n vertices. The vertices correspond to video frames and edges to the adjacency matrix $\mathbf{W} \in \{0,1\}^{n\times n}$ with $W_{i,j} = 1$ if $|i-j| = 1$ and 0 otherwise. To complete manifold embedding in the graph, P_n is mapped to a line that minimise the distance between connected vertices. If $y = (y_1, y_2, \ldots, y_n)^T$ is the map, y is obtained by minimising the following [114].

$$\sum_{i,j}(y_i - y_j)^2 W_{i,j}, \quad i,j = 1,2,\ldots,n \tag{3.10}$$

where this equation is equivalent to calculating the eigenvectors of the Laplacian graph P_n. The Laplacian graph is created with the eigenvectors $\{y_1, \ldots, y_{n-1}\}$ and allows y_k to be viewed as a set of points described by [114].

$$f_k^n(t) = \sin(\pi k t + \pi(n-k)/(2n)), t \in [1/n, 1] \tag{3.11}$$

sampled at $t = 1/n, 2/n, \ldots, 1$. The resulting curve described by [114].

$$\mathscr{F}^n(t) = \begin{bmatrix} f_1^n(t) \\ f_2^n(t) \\ \vdots \\ f_{n-1}^n(t) \end{bmatrix} \tag{3.12}$$

This curve is then used to temporally interpolate images at random positions within a micro-expression. To find the correspondences for the curve $\mathscr{F}^n$ within the image space, the image frames are mapped to points defined by $\mathscr{F}^n(1/n), \mathscr{F}^n(2/n), \ldots, \mathscr{F}^n(1)$. A linear extension of graph embedding [148] is then used to learn a transformation vector w that minimises [114].

$$\sum_{i,j}(w^T x_i - w^T x_j)^2 W_{i,j}, \quad i,j = 1,2,\ldots,n \tag{3.13}$$

where $x_i = \xi_i - \bar{\xi}$ is a mean-removed vector and ξ_i is the vectorised image. The resulting eigenvalue problem was solved by He et al. [54]

$$XLX^T w = \lambda' XX^T w \tag{3.14}$$

by using the singular value decomposition with $X = U\Sigma V^T$. A new image ξ can then be created using interpolation by al. [54]

$$\xi = UM\mathscr{F}^n(t) + \bar{\xi} \tag{3.15}$$

where M is a square matrix. There is an assumption that ξ_i are linearly independent, and the validity of the TIM method depends on this.

The interpolated frames of a micro-expression clip preserves the characteristics of the original movement well, whilst smoothing out the temporal profile. For the proposed method, we chose to interpolate the original frame rate of 200 fps down to 150 and 100 fps. The amount of frames chosen was determined by

$$\alpha = \gamma(\theta \in \Omega) \tag{3.16}$$

where α is the amount of frames chosen for subsampling, γ is the scaling factor and $\theta \in \Omega$ is the original amount of frames θ within the movement Ω. For instance, the scaling factor for CASME II is represented by 0.5 for 100 frames and 0.25 for 50 frames.

3.3.3 Resolution Down-Scale

CASME II has been captured in 640×480 pixels in the raw section of the dataset. The pre-processed part of the dataset have about 280×340 pixels for the cropped facial area. SMIC high speed (HS) camera set to 100 fps and resolution of 640×480 was used to gather the expressions. The facial resolution for SMIC is 190×230 pixels, which has lower resolution than CASME II. In order to test the effects of resolution variations in micro-expressions recognition we scaled down both datasets (SMIC and CASME II) by 75% and 50% from the original resolution which also included as shown in Fig. 3.4.

3.3.4 Feature Representation and Classification

Three feature types are used to test the performance, which are Local Binary Patterns in Three Orthogonal Planes, 3D Histograms of Oriented Gradient and Histogram of Oriented Optical Flow. These features have extracted from the original dataset and eight variations of the CASME II and five variations of SMIC. These features have training and testing using SMO classifier and have been evaluated and validated using 10-fold cross-validation and leave-one-subject-out (LOSO).

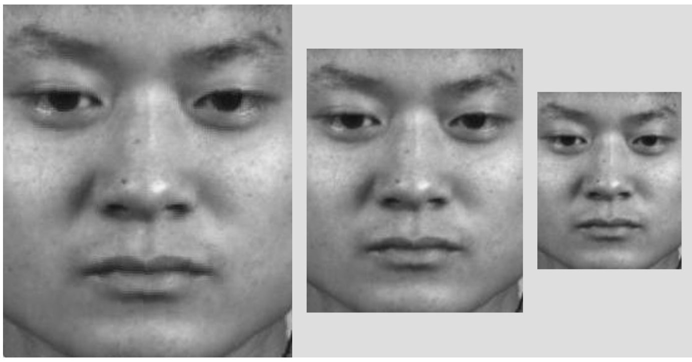


Figure 3.4: Down scale resolution: 100% (original resolution), 75% and 50% from the original resolution.

3.4 Objective Classes for Micro-Facial Expression Recognition

3.4.1 Overview

Micro-facial expression analysis is less established and harder to implement due to being less distinct than normal facial expressions. Feature representations, such as Local Binary Patterns (LBP) [106, 104, 164], Histogram of Oriented Gradients (HOG) [17] and Histograms of Oriented Optical Flow (HOOF) [10], are commonly used to describe micro-expressions. Although FME analysis is very difficult, the popularity in recent years has grown due to the potential applications in security and interrogations [108, 43, 41], healthcare [55, 14] and automatic detection in real-world applications where the detection accuracy of humans peaks around 40% [43].

Generally, the process of recognising normal facial expressions involves preprocessing, feature extraction and classification. Micro-expression recognition is not an exception, but the features extracted should be more descriptive due the small movement in micro-expressions compared with normal expressions. One of the biggest problems faced by research in this area is the lack of publicly available datasets, which the success in facial expression recognition [158] research largely relies on. Gradually, datasets of spontaneously induced micro-expression have been developed [84, 155, 149, 18], but earlier research was centred around posed datasets [118, 125].

Eliciting spontaneous micro-expression is a real challenge because it can be very difficult to induce the emotions in participants and also get them to conceal them effectively in a lab-controlled environment. Micro-expression datasets need decent ground truth labelling with Action Units (AUs) using the Facial Action Coding System (FACS) [35]. FACS objectively assigns AUs to the muscle movements of the face. If any classification of movements take place for FMEs, it should be done with AUs and not only emotions. Emotion classification requires the context of the situation for an interpreter to make a meaningful interpretation. Most spontaneous micro-expression datasets have FACS ground truth labels and estimated or predicted emotion. These have been annotated by an expert and self-reports written by participants.

We contend that using AUs to classify micro-expressions gives more accurate results than using predicted emotion categories. By organising the AUs of the two most recent FACS coded state-of-the-art datasets, CASME II [149] and SAMM [18], into objective classes, we ensure that the learning methods train on specific muscle movement patterns and therefore increase accuracy. Yan et al. [152] also state that it is inappropriate to categorise micro-expressions into emotion categories, and that using FACS AU research to inform the eventual emotional classification.

To date, experiments on micro-expression recognition using categories based purely on AU movements, has not been completed. Additionally, the SAMM dataset was designed for micro-movement analysis rather than recognition. We contribute by completing recognition experiments on the SAMM dataset for the first time with three features previously used for micro-expression analysis: LBP-TOP [164], HOOF [11] and HOG 3D [17, 117]. Further, the proposed objective classes could inform future research on the importance of objectifying movements of the face.

The proposed classes will show that classifying expressions using Action Units, instead of predicted emotion, removes the potential bias of human reporting. The proposed classes are tested using LBP-TOP, HOOF and HOG 3D feature descriptors. The experiments are evaluated on two benchmark FACS coded datasets: CASME II and SAMM.

3.4.2 Datasets Analysis

This section will describe two datasets which are used in the experiments. A comparative summary of the datasets can be seen in Table 3.1. Previously developed micro-expression recognition systems are also discussed using established features to represent each micro-expression.

3.4.2.1 CASME II

When analysing the FACS codes of the CASME II dataset, it was found that there are many conflicts to the coded AUs and the estimated emotions. These inconsistencies do not help

Table 3.1: A summary of the different features of the CASME II and SAMM datasets.

Feature	CASME II [149]	SAMM [18]
Micro-Movements	247	159
Participants	35	32
Resolution	640×480	2040×1088
Facial Resolution	280×340	400×400
FPS	200	200
Spontaneous/Posed	Spontaneous	Spontaneous
FACS Coded	Yes	Yes
No. Coders	2	3
Emotion Classes	5	7
Mean Age (SD)	22.03 (SD = 1.60)	33.24 (SD = 11.32)
Ethnicities	1	13

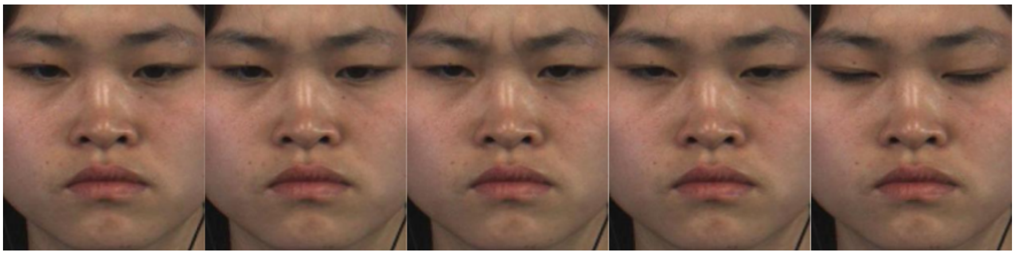


Figure 3.5: Sample frames showing Subject 11's micro-expression clip 'EP19_03f' that was coded as an AU4 in the 'others' category.

when attempting to train distinct machine learning classes, and adds further justification for the proposed introduction of new classes based on AUs only.

For example, Subject 11 with the micro-expression clip filename of 'EP19_03f', was coded as an AU4 in the 'others' estimated emotion category (shown in Fig. 3.5). However, Subject 26 with the micro-expression clip filename of 'EP18_50', was also coded with AU4 but in the 'disgust' estimated emotion category (shown in Fig. 3.6). As can be seen in the apex frame (centre image) of both Fig. 3.5 and 3.6, AU4, the lowering of the brow, is present. Having the same movement in different categories is likely to have an effect on any training stage of machine learning.

3.4.2.2 SAMM

The SAMM dataset was originally designed to investigate micro-facial movements by analysing muscle movements of the face rather than recognising distinct classes [23]. We

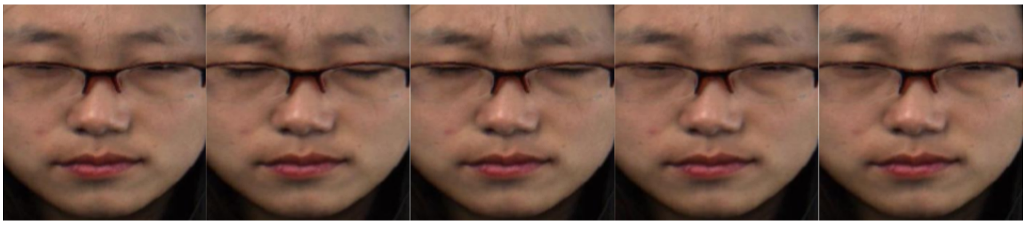


Figure 3.6: Sample frames showing Subject 26's micro-expression clip 'EP18_50' that was coded as an AU4 in the 'disgust' category.

Table 3.2: Each class represents AUs that can be linked to emotion.

Class	Action Units
I	AU6, AU12, AU6+AU12, AU6+AU7+AU12, AU7+AU12
II	AU1+AU2, AU5, AU25, AU1+AU2+AU25, AU25+AU26, AU5+AU24
III	A23, AU4, AU4+AU7, AU4+AU5, AU4+AU5+AU7, AU17+AU24, AU4+AU6+AU7, AU4+AU38
IV	AU10, AU9, AU4+AU9, AU4+AU40, AU4+AU5+AU40, AU4+AU7+AU9, AU4 +AU9+AU17, AU4+AU7+AU10, AU4+AU5+AU7+AU9, AU7+AU10
V	AU1, AU15, AU1+AU4, AU6+AU15, AU15+AU17
VI	AU1+AU2+AU4, AU20
VII	Others

are the first to categorise SAMM based on the FACS AUs and then use these categories for FME recognition.

3.4.3 Classes Restructuring

To overcome the conflicting classes in CASME II, we restructure the classes around the AUs that have been FACS coded. Using EMFACS [37], a list of AUs and combinations are proposed for a fair categorisation of the SAMM [18] and CASME II [149] datasets. Categorising in this way removes the bias of human reporting and relies on the ground truth movement data, feature representation and recognition technique for each micro-expression clip. Table 3.2 shows 7 classes and the corresponding AUs that have been assigned to that class. Classes I-VI are linked with happiness, surprise, anger, disgust, sadness and fear. Class VII relates to contempt and other AUs that have no emotional link in EMFACS [37]. It should be noted that the classes do not directly correlate to being these emotions, however the links used are informed from previous research [33, 35, 37]. Each movement in both datasets were classified based on the AU categories of Table 3.2, with the resulting frequency of movements being shown in Table 3.4.

Table 3.3: The total number of movements assigned to the new classes for both SAMM and CASME II.

Class	CASME II	SAMM	Total
I	25	24	49
II	15	13	28
III	99	20	119
IV	26	8	34
V	20	3	23
VI	1	7	8
VII	69	84	153
Total	255	159	415

3.4.4 Feature Representation and Classification

Micro-expression recognition experiments are run on two datasets: CASME II and SAMM. For this experiment, three types of feature representations are extracted from a sequence of grey images which represent the FME. These image sequences are divided into 5×5 blocks that are non-overlapping. The LBP-TOP features [164] radii parameters for X, Y and T are set to 1, 1 and 4 respectively and all neighbours in three planes set to 4. The HOG3D [117] and HOOF [11] features are set to the parameters described in the original implementations.

Sequential Minimal Optimization (SMO) [116] is used in the classification phase with 10-fold cross validation and leave-one-subject-out (LOSO) to classify between I-V, I-VI and I-VII classes.

3.4.5 Evaluating objective classes on a composite database

To show that objective classes can provide more standardization for the classes between datasets a method to evaluate them on a composite database has been proposed by introducing selective block-based with fused features representation. Base on objective classes, this task combines CASME II and SAMM (Two benchmark Facial Action Coding System (FACS)) into a single composite database (Composite Database Evaluation task (CDE)) [156] and uses Leave-One-Subject-Out cross-validation to evaluate the performance. CASME II consists of five emotion classes (happiness, disgust, surprise, repression and others).

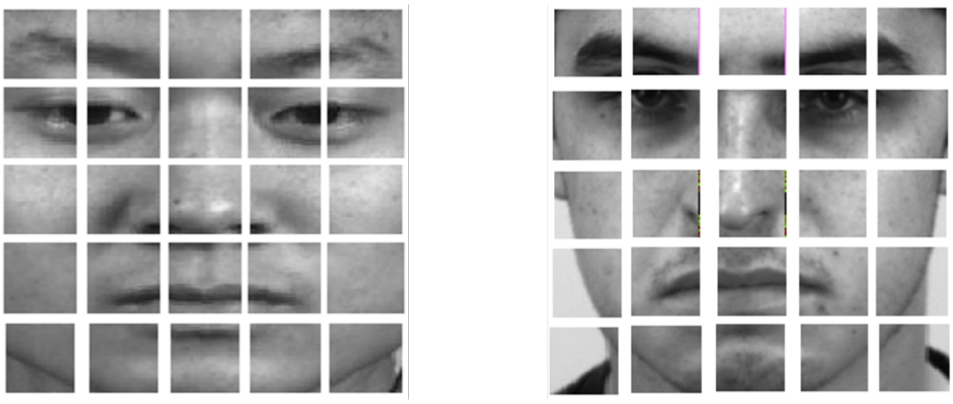


Figure 3.7: Images split into 5x5 blocks. On the left: an example from CASME II; and on the right: an example from SAMM. The total is 25 blocks

However, SAMM consists of seven emotion classes (happy, sad, surprise, angry, disgust, fear and contempt). To form standardised the databases, we focus on objective classes that based on Action Units(AUs) than self-reports. A summary of the objective classes are as illustrated in Table 3.4. A single composite database for this experiment has a total of 253 micro-expressions.

Table 3.4: The total number of movements assigned to the new classes for both SAMM and CASME II.

Class	CASME II	SAMM	Composite
I	25	24	49
II	15	13	28
III	99	20	119
IV	26	8	34
V	20	3	23
Total	185	68	253

We propose a selective block-based feature fusion representation method for CDE. The limitation of the existing 5x5 blocks is not all 25 blocks are correspond to facial movement. Figure 3.8 shows our proposed selective blocks. Then, we extract LBP-TOP, HOOF and HOG3D from each block., These features are fused and all blocks are concatenated into a single histogram. Leave-one-subject-out (LOSO) cross validation is use to evaluate our proposed method on the CDE for I-V objective classes. The next section discuss and compare the result of our proposed method with baseline methods.

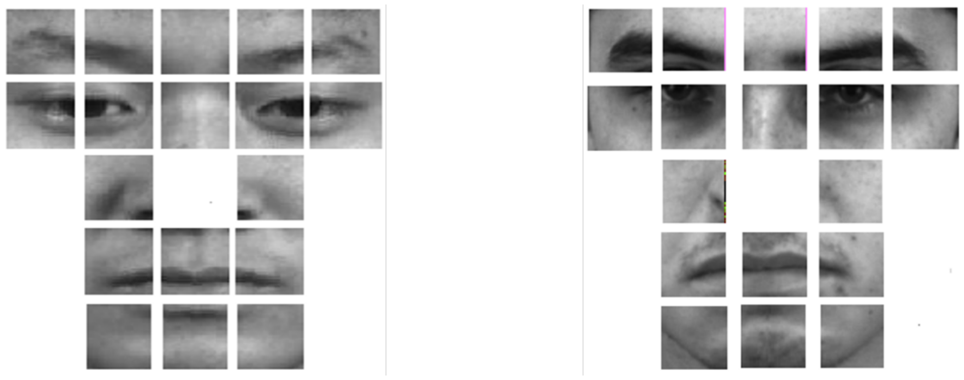


Figure 3.8: Our proposed selective block-based. On the left: an example from CASME II; and on the right: an example from SAMM. The block is reduced from 25 blocks to 18 blocks.

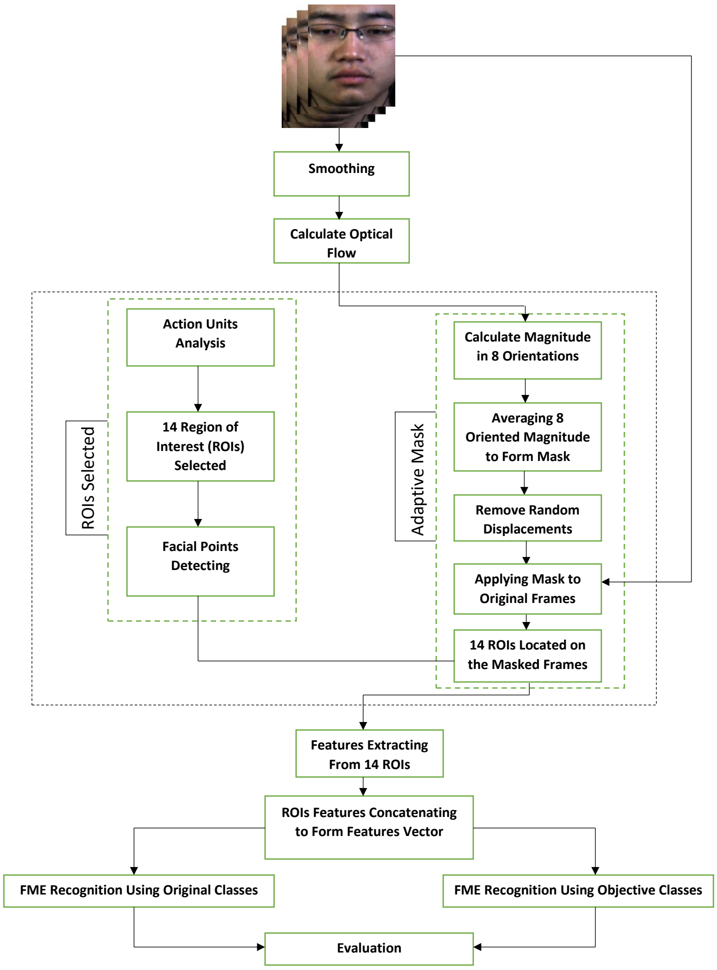
3.5 Adaptive Mask for Region-based Facial Micro-Expression Recognition

3.5.1 Overview

Facial micro-expression can be characterized by its short duration and subtle movements. In facial micro-expression recognition, these subtle movements require more specific feature descriptors due to only a few parts of the face produce information that helps us to recognize micro-expressions. In addition to that, the brightness and head-movements may confuse machine learning which considered it as facial-movements. Over the past decade, researchers designed different methods to study micro-expressions recognition. To further study this aspect, we proposed a region-based method with an adaptive mask for facial micro-expression recognition. Based on the most frequent Action Units on the two publicly available datasets, i.e. CASME II and SAMM, 14 ROIs are defined where the adaptive mask is created by calculating the optical flow after Gaussian smoothing Figure 3.9 describes the methodology of the proposed method, each part in the figure will be described into more details in the following sections.

3.5.2 Facial Landmarks Detection

Faces have quite distinct features, such as eyebrows, mouths, noses and eyes. Most humans have these features in about the same position, and so identifying the points where these features occur on the face is an interest research problem. Zhou et al. [170] proposed a way of detecting facial points using a deep learning approach named convolutional neural networks (CNN). The research toolkit developed requires an Internet connection to allow



Smoothing
Calculate Optical Flow
Action Units Analysis
14 Region of Interest (ROIs) Selected
Facial Points Detecting
ROIs Selected
Calculate Magnitude in 8 Orientations
Averaging 8 Oriented Magnitude to Form Mask
Remove Random Displacements
Applying Mask to Original Frames
14 ROIs Located on the Masked Frames
Adaptive Mask
Features Extracting From 14 ROIs
ROIs Features Concatenating to Form Features Vector
FME Recognition Using Original Classes
FME Recognition Using Objective Classes
Evaluation

Figure 3.9: Methodology of proposed method.

for the images to be processed on the Face++ servers. We use Face++ to detect the facial landmarks. Figure 3.11(a) shows the facial landmark points.

3.5.3 The Facial Action Coding System

The Facial Action Coding System (FACS) was first published by Ekman and Friesen [34] as a research tool to measure any facial expression that a human can perform. It was designed to objectively understand the facial muscle movements with no inference to emotion i.e., how muscular action is related to facial appearances. Each observable component of facial movement is called an Action Unit (AU) and all facial expressions can be broken down into their constituent AUs. The AUs in the FACS manual are presented in two main groups: upper face and lower face actions. Each main group is then split into sub-groups.

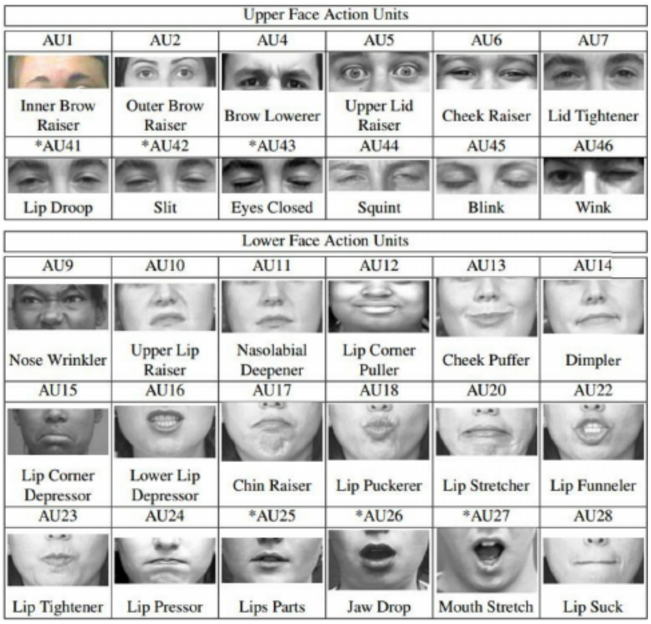


Figure 3.10: FACS AUs defined as two main groups and split into sub-groups.

A breakdown of the groupings and AUs belonging to that group is shown in figure 3.10.

3.5.4 FACS-Based Regions

Analysis have been done on two of the most popular publicly available micro-expressions datasets, CASMEII [150] and SAMM [18], to find the most frequently occur AUs in micro-expressions. Table 3.5 summarizes the frequency of the AUs with the highest occurrence on AU4 and the lowest occurrence on AU31. This step is to ensure only relevant movements are detected. Additionally, the advantage of this step is that the features can be locally analyzed without processing insignificant parts of the face. Table 3.6 summarises the name of each region and its correspondence AUs. ROIs have located on face after detect facial landmarks points. Figure 3.11(b) illustrates our proposed ROIs.

3.5.5 Smoothing

Due to the subtle nature of micro-expression and the noise could affect the extracted features, noise reduction should be applied. One of the simplest method in noise reduction is using the smoothing algorithm. For our work, Gaussian smoothing operator have been applied to reduce the noise, as shown in Figure 3.12.

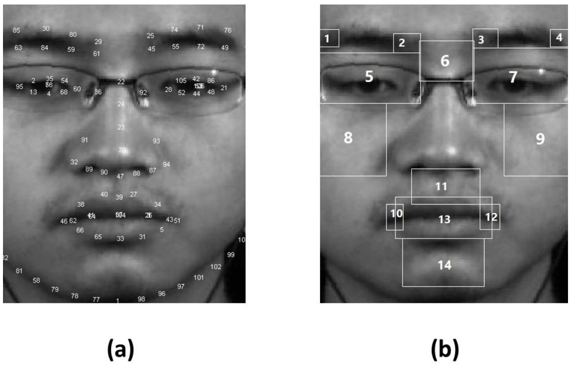


Figure 3.11: (a)Facial landmark points on a sample subject of CASMEII (b) 14 ROIs based on the frequency of AUs occurrences.

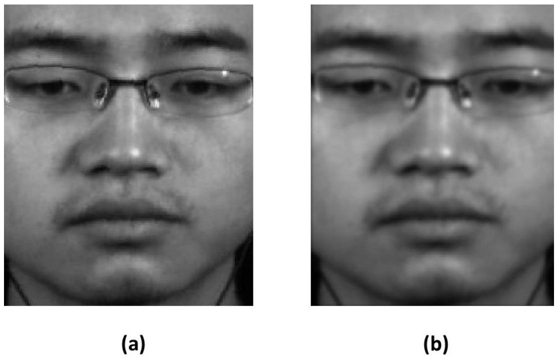


Figure 3.12: Preprocessing step: (a) Before smoothing (b) after smoothing.

3.5.6 Optical Flow

Optical flow is the pattern of apparent motion. It tries to calculate the motion between two image frames at every pixel position. The estimation of optical flow is based on two assumptions: brightness constancy and temporal persistence. A pixel at location (x, y, t) with intensity I (x, y, t) will move by $\triangle x$, $\triangle y$ and $\triangle t$ between the two image frames, and it satisfies the following brightness constancy constraint: [96]

$$I(x,y,t) = I(x+\triangle x, y+\triangle y, t+\triangle t) \tag{3.17}$$

According to temporal persistence, the image constraint at I (x, y, t) with Taylor series can be developed to get the following equation: [96]

$$I(x+\triangle x, y+\triangle y, t+\triangle t) = I(x,y,t) + \frac{\partial I}{\partial x}\triangle x + \frac{\partial I}{\partial x}\triangle y + \frac{\partial I}{\partial x}\triangle t + H.O.T \tag{3.18}$$

Where H.O.T is higher-order term.
Rearrange equation 3.18 to get the following equation: [96]

Table 3.5: A Summary of AUs frequency on CASME II and SAMM.

AU	CASME II	SAMM	Total
1	26	6	32
2	22	18	40
4	129	23	152
5	2	10	12
6	13	5	18
7	38	46	84
9	13	5	18
10	16	6	22
12	34	30	64
13	0	3	3
14	27	13	40
15	16	4	20
17	25	7	32
18	0	4	4
20	0	7	7
24	2	10	12
25	2	7	9
26	0	6	6
31	0	2	2

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial x}V_y + \frac{\partial I}{\partial x}V_t = 0 \tag{3.19}$$

where V_x,V_y are the horizontal and vertical components of the optical flow field. The equation 3.18 is called optical flow constraint equation. There are many methods to solve constraint equation. In th proposed method Lucas-Kanade method [97] have been used for optical flow estimation.

3.5.7 Oriented Magnitude Mask

The optical flows are calculated from the frames sequences after smoothing operation to represent motion information of each pixel in FME in the form of horizontal and vertical displacements V_x and V_y. The optical flows are computed from each pair of frames between the first frame and the rest.

Table 3.6: The region number, name and its associated AUs.

Region Number	Region Name	Associated AU(s)
1	Right Brow - Right	2,4
2	Right Brow - Left	1,4
3	Left Brow - Right	1,4
4	Left Brow - Left	2,4
5	Right Eye	5,7
6	Glabella	1, 4, 9
7	Left Eye	5,7
8	Right Cheek	6, 12
9	Left Cheek	6, 12
10	Dimple - Right	12, 13, 14, 18, 20
11	Upper Lip	10
12	Dimple - Left	12, 13, 14, 18, 20
13	Mouth	12, 13, 14, 15, 16 17, 18, 19, 20, 22 23, 24, 25, 26, 28
14	Chin	15, 17, 25, 26

Using V_x and V_y orientation and magnitude could be calculated for each pixel displacements, where the magnitude calculated using the following equation: [135]

$$M = \sqrt{V_x^2 + V_y^2} \tag{3.20}$$

and the orientation calculated by: [135]

$$\Theta = tan^{-1}(\frac{V_x}{V_y}) \tag{3.21}$$

Using equation 3.20 and 3.21 the optical flow have been visualize using magnitude value for each pixel among eight orientations, this visualization have been done using the following equation:

$$om(x,y,i)(\Theta) = \begin{cases} m & (\Pi/4)(i-1) <= \Theta < (\Pi/4)(i-1) + (\Pi/4) \\ 0 & else \end{cases} \tag{3.22}$$

where x,y the position of the current pixel, i the number of orientations 1<=i<=8, according to that $2\Pi/8 = \Pi/4$. These eight oriented magnitude is then go through an averaging process as shown in Figure 3.13 to form the mask (in the centre of Figure 3.13). The mask was adaptive because every single frame in the sequence has its own mask due to the different movement for each frame.

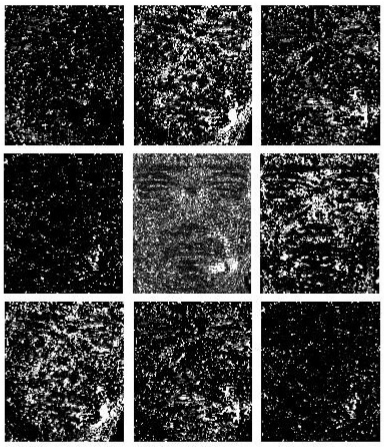


Figure 3.13: Illustrations of the magnitude calculation of optical flow in 8 orientation (surrounding images), the centre image shows the average of 8 surrounding images.

3.5.8 Remove Random displacements

The vision of the optical flow field may be affected by brightness changes and produce pixel displacements which could be considered as facial-movement and this leads to confusion for machine learning especially in the micro-expression circumstances because micro-expressions are subtle.

Noticed that, the displacements in optical flows caused by facial-movements are direction consistent among neighboring frames, whereas the displacements caused by the light conditions are random and direction inconsistent. Therefore, detecting these random displacements and removing it can enhance the directional consistent displacements caused by micro-expression movements and decrease the random displacements caused by brightness changes. This process has been done by calculating the number of flips between 0 and 1 for correspondence pixels through the frame, where the facial-movement displacements have less flipping in binary pattern as the opposite of random displacement which tends to be more flipping as shown in Figure 3.14(c). Algorithm 1 explains removing random displacements

Algorithm 1: Removing random displacements

```
Input: Mask Sequences
Output: Mask Sequences without random displacements
for $x = 1, x{+}{+}, while\ x < Mask\_height$ do
    for $y = 1, y{+}{+}, while\ x < Mask\_width$ do
        for $i = 2, i{+}{+}, while\ i < Masks\_length$ do
            if mask[x,y,i]!=mask[x,y,i-1] then
                flips[x,y]=flips[x,y]+1
for $x = 1, x{+}{+}, while\ x < Mask\_height$ do
    for $y = 1, y{+}{+}, while\ x < Mask\_width$ do
        sum=sum+flips[x,y]
mean = sum/$Mask\_height * Mask\_width$ for $x = 1, x{+}{+}, while\ x < Mask\_height$
 do
    for $y = 1, y{+}{+}, while\ x < Mask\_width$ do
        for $i = 2, i{+}{+}, while\ i < Masks\_length$ do
            if flips[x,y]>mean then
                mask[x,y,i]=0
```

The final mask applied to the original image to produce a masked image as shown in Figure 3.14(d) before locating the ROIs as in Figure 3.14(f).

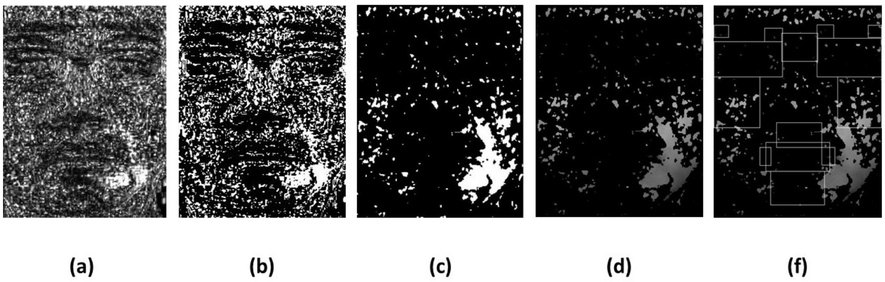


Figure 3.14: Optical flow mask: (a) Oriented magnitude (b) Black and White (c) Removing random displacements (d) applying mask to original image (f) regions after mask.

3.5.9 Feature Decriptors, Classification and Validation

LBP-TOP has been used as a features descriptor and set the neighboring points and radii parameters to $LBPTOP_{4,4,4,1,1,4}$ as in [150]. LBP-TOP has been extracted from each ROI through all the frames to form features vector for this region this has been done for all regions before all ROIs features concatenated in one large vector which describes the specific

micro-expression.
SMO used as classifier to classify between two type of classes due the proposed method has been evaluated using the original classes and proposed objective classes. Leave-one-subject-out (LOSO) has been used to validate the proposed method, where LOSO is well-established and widely used in FME evaluation.

3.6 Summary

Research on automated facial micro-expression recognition using machine learning has witnessed good progress in recent years. A number of promising methods based on texture features, gradient features and optical flow features have been proposed. Many datasets were generated but lack of standardisation is indeed a great challenge. Therefore, comparing and discussing the effect of different frame rate and resolution have been conducted. Three of the most famous feature descriptors have been used to represent micro-expression. These features have variation in their nature, which is very suitable to test the effect of frame rate and resolution based on these features. LBP-TOP used as an examples of texture-based features, 3DHOG to represent gradient-based features and HOOF to represent optical flow-based features.
Currently, emotion classes within the CASME II dataset are based on Action Units and self-reports, creating conflicts during machine learning training. We will show that classifying expressions using Action Units, instead of predicted emotion, removes the potential bias of human reporting. The proposed classes are tested using LBP-TOP, HOOF and HOG 3D feature descriptors. The experiments are evaluated on two benchmark FACS coded datasets: CASME II and SAMM.
A new method for FME recognition have been proposed. The method is region-based, where 14 ROIs have selected based on AUs analysis on CASMEII and SAMM. ROIs have been proposed for locally analyzed the features to avoid unimportant information of the face. Further, to be more specific to the movement related to FME, adaptive mask based on the micro-motion using optical flow have applied to each frame of ME. Then the random displacements which caused by the light condition and could be consider as micro-movements. LBP-TOP features extracted from each region and SMO is implemented as the classifier. The proposed method evaluated on two of benchmark datasets: CASME II and SAMM.

Chapter 4

Result and Discussion

4.1 Introduction

Results of the thesis contributions as shown in chapter 3 will be introduced in this chapter and a discussion about these results will be done. The start will be to show the effect of changes in the Spatio-temporal settings for FME datasets. Then the result of FME recognition using objective classes will be shown to prove that restructuring classes around action units have an advantage on the labeling using self-report, also to show that objective classes provide more standardized and unified classes, cross datasets results using objective classes on a composite datasets (CASME II and SAMM) will be shown. Finally, the results of proposed FME recognition method will be shown to demonstrate the proposed method was the appropriate solution to overcome the thesis problem.

4.2 The Implication of Spatial Temporal Changes on Facial Micro-Expression Analysis

We have conducted comprehensive evaluation on the performance of three popular feature representations on micro-expressions recognition using difference frame rates and resolution. We observed that the top performers across different categories were varied.

Table 4.2 summarised the results of two validation methods, 10-fold cross validation and Leave-one-subject-out (LOSO), on CASME II. For 10-fold cross validation, the best result was LBP-TOP with 200 fps and 100% of the original resolution, achieved an F-Measure of 0.637. However, when validated with LOSO, the best results was HOOF with 200 fps and 50% resolution, achieved an F-Measure of 0.439.

Table 4.3 summarised the results of 10-fold cross validation and LOSO on SMIC. For 10-fold cross validation, the best result was 3DHOG with 50 fps and 75% resolution, achieved an F-Measure of 0.624. When validated with LOSO, the best result was HOOF with 100 fps and 75% resolution, achieved an F-Measure of 0.614.

4.2.1 Comparison of the State-of-the-art Methods

Table 4.1 compared the performance of the state-of-the-art methods on SMIC and CASME II. The majority of the state-of-the-art results were based on the original resolution and frame-rate. Therefore, this comparison was based on the original properties of the dataset. In addition, the majority of previous works reported their results using accuracy as the performance metric. Therefore, we compared these methods based on accuracy. From our observation, although Li et al. [82] and Liu et al. [93] achieved good accuracy on SMIC dataset, the performance on CASME II were comparable to the basic features of HOOF, LBP-TOP and 3DHOG.

Table 4.1: The performance of the state-of-the-art methods on SMIC and CASME II. Note that we have only included some popular previous works that reported their results on both datasets.

Method	SMIC	CASME II
Li et al. [82]	0.5352	0.5749
Lu et al. [95]	0.8286	0.6419
Le et al. [78]	0.4434	0.4378
Huang et al. [58]	0.5793	0.5951
Liu et al [93]	0.8000	0.6737
LBP-TOP	0.561	0.66
3DHOG	0.538	0.611
HOOF	0.593	0.636

4.2.2 Temporal Analysis

Since LOSO is a better approach in performance measure, we further analyse its performance for temporal analysis. As shown in Fig. 4.1 and Fig. 4.2, we observed that LBP-TOP and HOOF performed better in high frame rate on CASME II and SMIC. For the majority of the resolution, F-Measure decreased or maintained as the frame rates dropped. In contrast, when compared the performance of 3DHOG on different frame rates, the F-Measure increased on lower frame rate (50 fps for CASME II and SMIC), as illustrated in Fig. 4.3.

Table 4.2: The results of CASME II for both the 10-fold cross-validation and leave-one-subject-out for the 3DHOG, HOOF and LBP-TOP features with a varying resolutions and frame rates.

		10-Fold Cross-Validation				Leave-One-Subject-Out (LOSO)			
Resolution	Frame Rate	Accuracy	TPR	FPR	F-Measure	Accuracy	TPR	FPR	F-Measure
3DHOG									
100%	200	0.718	0.478	0.290	0.425	0.611	0.368	0.271	0.319
	100	0.713	0.474	0.295	0.421	0.595	0.345	0.276	0.301
	50	0.731	0.509	0.281	0.463	0.627	0.396	0.266	0.348
75%	200	0.722	0.486	0.286	0.436	0.595	0.341	0.265	0.291
	100	0.714	0.474	0.293	0.423	0.590	0.345	0.270	0.295
	50	0.727	0.501	0.284	0.456	0.604	0.356	0.258	0.301
50%	200	0.727	0.494	0.279	0.446	0.594	0.352	0.267	0.308
	100	0.723	0.490	0.283	0.444	0.581	0.345	0.291	0.292
	50	0.735	0.509	0.269	0.472	0.603	0.368	0.272	0.311
HOOF									
100%	200	0.698	0.475	0.319	0.423	0.636	0.427	0.304	0.383
	100	0.679	0.443	0.374	0.318	0.615	0.404	0.345	0.314
	50	0.676	0.443	0.365	0.362	0.616	0.404	0.346	0.317
75%	200	0.708	0.494	0.305	0.450	0.627	0.408	0.275	0.368
	100	0.666	0.420	0.389	0.285	0.614	0.392	0.329	0.309
	50	0.673	0.435	0.357	0.367	0.609	0.376	0.315	0.307
50%	200	0.744	0.553	0.258	0.536	**0.668**	**0.475**	0.283	**0.439**
	100	0.720	0.518	0.318	0.455	0.610	0.388	0.326	0.321
	50	0.673	0.431	0.357	0.357	0.580	0.365	0.332	0.297
LBP-TOP									
100%	200	**0.819**	**0.635**	**0.131**	**0.637**	0.66	0.435	**0.193**	0.415
	100	0.802	0.603	0.139	0.609	0.610	0.384	0.238	0.348
	50	0.779	0.552	0.166	0.553	0.618	0.360	0.222	0.346
75%	200	0.801	0.600	0.144	0.603	0.613	0.376	0.240	0.354
	100	0.802	0.603	0.146	0.606	0.595	0.352	0.256	0.318
	50	0.772	0.545	0.172	0.545	0.620	0.368	0.220	0.342
50%	200	0.778	0.568	0.168	0.571	0.618	0.380	0.202	0.346
	100	0.781	0.572	0.163	0.575	0.600	0.348	0.210	0.322
	50	0.763	0.533	0.186	0.533	0.598	0.325	0.219	0.297

Table 4.3: The results of SMIC for both the 10-fold cross-validation and leave-one-subject-out for the 3DHOG, HOOF and LBP-TOP features with varying resolutions and frame rates.

		10-Fold Cross-Validation				Leave-One-Subject-Out (LOSO)			
Resolution	Frame Rate	Accuracy	TPR	FPR	F-Measure	Accuracy	TPR	FPR	F-Measure
3DHOG									
100%	HS(100 FPS)	0.720	0.597	0.250	0.588	0.538	0.378	0.274	0.333
	HS(50 FPS)	0.716	0.591	0.245	0.586	0.567	0.426	0.320	0.407
75%	HS(100 FPS)	0.707	0.579	0.261	0.569	0.594	0.469	0.265	0.429
	HS(50 FPS)	**0.739**	**0.628**	**0.230**	**0.624**	0.533	0.402	0.313	0.374
50%	HS(100 FPS)	0.721	0.591	0.242	0.580	0.556	0.408	0.252	0.405
	HS(50 FPS)	0.723	0.597	0.233	0.594	0.593	0.457	0.264	0.458
HOOF									
100%	HS(100 FPS)	0.734	0.621	0.231	0.619	0.593	0.493	0.278	0.498
	HS(50 FPS)	0.713	0.597	0.264	0.586	0.574	0.475	0.357	0.467
75%	HS(100 FPS)	0.722	0.603	0.243	0.599	**0.687**	**0.609**	**0.254**	**0.614**
	HS(50 FPS)	0.698	0.573	0.275	0.564	0.619	0.536	0.372	0.523
50%	HS(100 FPS)	0.723	0.609	0.259	0.598	0.627	0.530	0.310	0.534
	HS(50 FPS)	0.683	0.554	0.305	0.521	0.609	0.506	0.376	0.486
LBP-TOP									
100%	HS(100 FPS)	0.703	0.573	0.241	0.572	0.561	0.402	0.279	0.349
	HS(50 FPS)	0.695	0.560	0.241	0.561	0.567	0.402	0.294	0.349
75%	HS(100 FPS)	0.693	0.554	0.244	0.554	0.549	0.384	0.285	0.344
	HS(50 FPS)	0.673	0.524	0.267	0.522	0.538	0.384	0.345	0.335
50%	HS(100 FPS)	0.700	0.567	0.247	0.566	0.545	0.402	0.268	0.395
	HS(50 FPS)	0.701	0.567	0.245	0.564	0.488	0.317	0.306	0.313

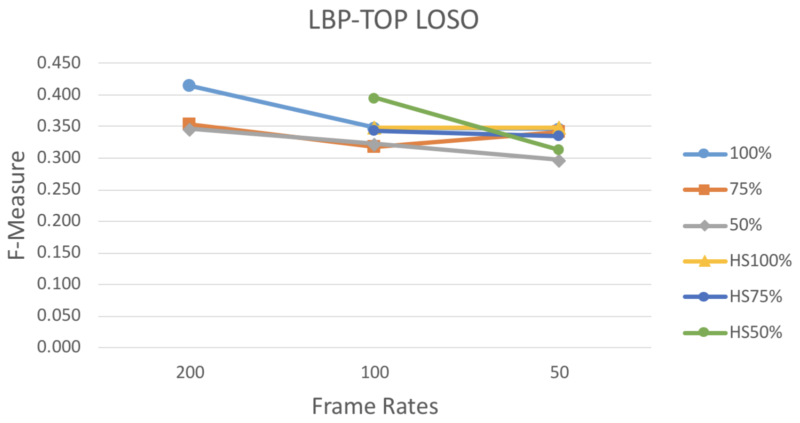


Figure 4.1: Comparison of F-Measure using LBP-TOP with varying resolution and frame-rates. The graph shows the best result is using high frame rate and high resolution when evaluated using LOSO validation on CASME II (100%, 75% and 50%) and SMIC (HS100%, HS75% and HS50%).

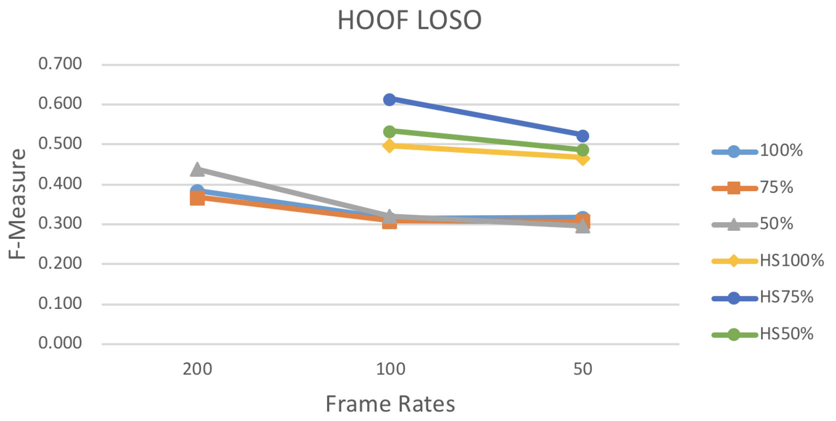


Figure 4.2: Comparison of F-Measure using HOOF with varying resolution and frame-rates. The graph shows the best result is using high frame rate and lower resolution when evaluated using LOSO validation on CASME II (100%, 75% and 50%) and SMIC (HS100%, HS75% and HS50%).

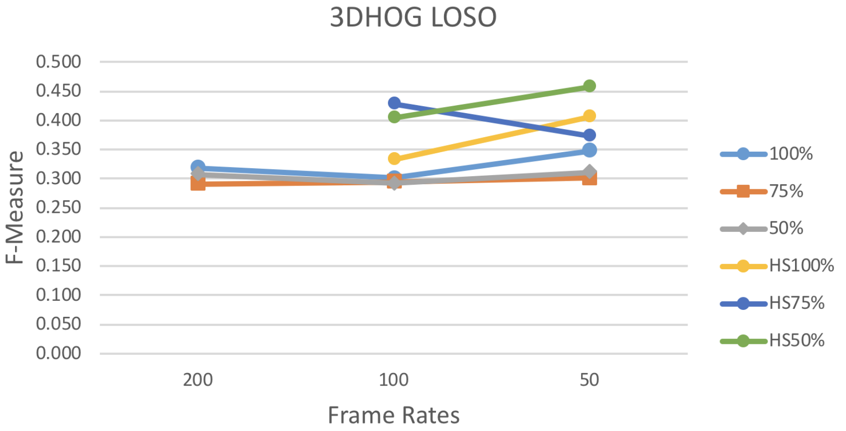


Figure 4.3: Comparison of F-Measure using 3DHOG with varying resolution and frame-rates. The graph shows the overall best result is achieved by using low frame rate when evaluated using LOSO validation on CASME II (100%, 75% and 50%) and SMIC (HS100%, HS75% and HS50%).

4.2.3 Spatial Analysis

Regarding the effect on varying resolutions on micro-expressions recognition, there is also an inconsistency for optimal resolution through different features. We found that LBP-TOP achieved better result when the full resolution of CASME II and SMIC were used. For HOOF, a low resolution (50% from the original resolution) achieved the best results on CASME II and the mid resolution (75% from the original resolution) achieved the best results on SMIC. This might be due to the lower facial resolution of SMIC when compared to CASME II. For 3DHOG, the result on spatial analysis is inconclusive. As illustrated in Fig. 4.3, the results are varied but the majority of the results performed better with low frame rate.

4.2.4 Features Analysis

By taking into account the effect of resolution and frame rate on micro-expressions recognition when using different feature descriptors, we observed that LBP-TOP as an example of texture-based features performed better on high resolution and high frame rate as illustrated in Fig. 4.1, texture-based features depended on pixels to extract the informations of micro-expressions, so more pixels(high resolution) means more informations. Also, high frame rate increases the number of pixels on the 3rd plane, this is why LBP-TOP performs better on high settings. Gradient-based features such as 3DHOG in these experiments does not need a high specification to achieve good results as shown in Fig. 4.3, with the best result achieved on 50 fps across different resolutions. On the other hand, HOOF features, which is optical flow-based, as shown in Fig. 4.2, performed better in a high frame rate scenario. Whilst there was no considerable need for high resolution, we found that the best result was recorded in 50% of the original resolution. As the opposite of texture-based features, we found that optical flow extracts the information by calculating the motion between frames, so it depends on temporal more than spatial this is why it needs high frame rate for better performance rather than high resolution.

4.2.5 Result Summary

The LOSO method of evaluation shows a decrease in accuracy compared with the 10-fold cross validation. While both have been used in previous research, the differences show the challenging nature of finding the correct way to determine a method's success. Further, the lower performance seen overall with LOSO can be attributed to the fairer nature of testing on a subject completed omitted from the training stage.

The frame rate is certainly important, with drops seen as this is decreased. However, obtaining equipment and data storage for 200 FPS recording can be difficult. A good trade-off could reduce the frame rate, but keep the best performing resolution and feature.

LBP-TOP and 3DHOG are relatively simple feature types, with HOOF being somewhat more informative based on temporal data. As micro-expression movements can look unique, even though the same muscle are used, simple features tend to pick out the obvious changes and struggle to model how a real micro-expression differs from noise.

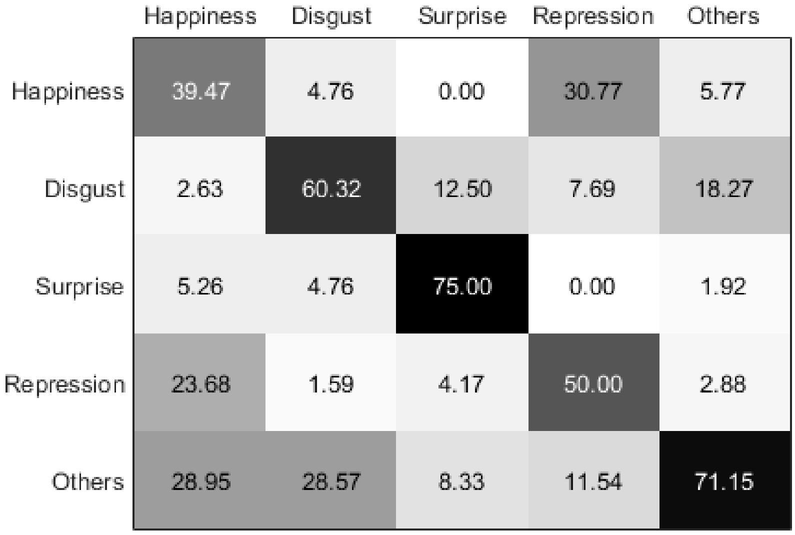
Analysis on 10 folds of 10-fold cross validation have been done, the distribution of classes through 10 folds are well distributed with low standard deviation, this implies that the sampling variations are minimal as shown table 4.4. Based on this analysis, the results variation were caused by the spatial and temporal changes.

Table 4.4: The sampling variation of classes distribution through the 10 folds.

Classes	1	2	3	4	5
fold 1	2	3	4	5	11
fold 2	1	7	5	0	12
fold 3	3	6	2	2	12
fold 4	3	1	0	4	17
fold 5	3	8	4	0	10
fold 6	12	5	2	2	4
fold 7	2	10	0	8	5
fold 8	3	5	5	0	12
fold 9	1	6	0	6	12
fold 10	2	8	3	0	12
STD	3.19	2.60	2.01	2.90	3.74

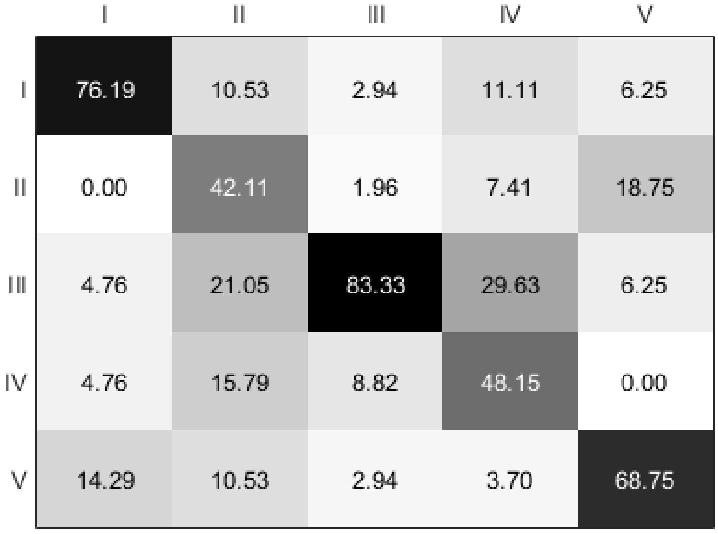
4.3 Objective Classes for Micro-Facial Expression Recognition

Evidence to support the proposed AU-based categories can be seen in the confusion matrix in Fig. 4.4. A high proportion of micro-expressions have been classified as ‘others’, for example 28.95% of the ‘happiness’ and 28.57% of the ‘disgust’ categories are classified as ‘others’ respectively. The original chosen emotions, including many placed in the ‘others’ category, leads to a lot of conflict at the recognition stage. It should be noted that the CASME II dataset [149] included self-reporting, which adds another layer of complexity during classification.



	Happiness	Disgust	Surprise	Repression	Others
Happiness	39.47	4.76	0.00	30.77	5.77
Disgust	2.63	60.32	12.50	7.69	18.27
Surprise	5.26	4.76	75.00	0.00	1.92
Repression	23.68	1.59	4.17	50.00	2.88
Others	28.95	28.57	8.33	11.54	71.15

Figure 4.4: Confusion matrix of the original CASME II classes using the LBP-TOP feature, using SMO as a classifier.



	I	II	III	IV	V
I	76.19	10.53	2.94	11.11	6.25
II	0.00	42.11	1.96	7.41	18.75
III	4.76	21.05	83.33	29.63	6.25
IV	4.76	15.79	8.82	48.15	0.00
V	14.29	10.53	2.94	3.70	68.75

Figure 4.5: Confusion matrix of the proposed classes I-V on the CASME II dataset using the LBP-TOP feature and SMO as a classifier.

Table 4.5: Results on the CASME II dataset showing each feature, proposed classes, and the original classes defined in [149] for comparison.

		10-Fold Cross-Validation					Leave-One-Subject-Out (LOSO)				
Feature	Class	Accuracy (%)	TPR	FPR	F-Measure	AUC	Accuracy (%)	TPR	FPR	F-Measure	AUC
LBP-TOP	Original	77.17	0.56	0.22	0.53	0.74	66.0	0.49	0.17	0.48	0.63
	I-V	77.94	0.63	0.33	0.58	0.70	67.80	0.54	0.14	0.51	0.44
	I-VI	76.84	0.59	0.32	0.55	0.69	67.94	0.53	0.14	0.51	0.44
	I-VII	76.13	0.50	0.23	0.45	0.70	61.92	0.39	0.17	0.35	0.63
HOOF	Original	78.83	0.61	0.19	0.60	0.78	63.6	0.51	0.24	0.49	0.61
	I-V	82.70	0.69	0.22	0.67	0.80	69.64	0.59	0.18	0.56	0.47
	I-VI	82.41	0.68	0.23	0.66	0.79	73.52	**0.62**	0.18	**0.60**	0.47
	I-VII	83.94	0.64	0.14	0.63	0.79	**76.60**	0.57	**0.14**	0.55	**0.72**
HOG3D	Original	80.93	0.62	0.14	0.62	0.79	59.59	0.38	0.24	0.35	0.50
	I-V	**86.35**	**0.72**	0.13	**0.72**	**0.84**	69.53	0.56	0.18	0.51	0.40
	I-VI	83.49	0.68	0.16	0.67	0.80	69.87	0.56	0.18	0.51	0.40
	I-VII	82.59	0.58	**0.12**	0.58	0.79	61.33	0.39	0.30	0.31	0.51

Table 4.6: Results on the SAMM dataset showing each feature and proposed classes.

		10-Fold Cross-Validation					Leave-One-Subject-Out (LOSO)				
Feature	Class	Accuracy (%)	TPR	FPR	F-Measure	AUC	Accuracy (%)	TPR	FPR	F-Measure	AUC
LBP-TOP	I-V	79.21	0.54	0.16	0.51	0.74	44.70	0.38	0.19	0.35	0.31
	I-VI	**81.93**	0.55	**0.13**	0.52	**0.74**	45.89	0.34	0.17	0.31	0.36
	I-VII	79.52	0.57	0.18	**0.56**	0.74	54.93	0.42	0.22	0.39	**0.40**
HOOF	I-V	78.95	0.56	0.16	0.55	0.74	42.17	0.32	**0.06**	0.33	0.32
	I-VI	79.53	0.52	0.15	0.51	0.73	40.89	0.28	0.07	0.27	0.35
	I-VII	72.80	0.52	0.32	0.50	0.65	60.06	0.49	0.25	**0.48**	0.30
HOG3D	I-V	77.18	0.51	0.17	0.49	0.74	34.16	0.22	0.15	0.22	0.24
	I-VI	79.41	0.48	0.15	0.45	0.71	36.39	0.19	0.14	0.19	0.26
	I-VII	79.09	**0.59**	0.25	0.55	0.71	**63.93**	**0.50**	0.22	0.44	0.30

The proposed classes I-V classification results using LBP-TOP can be seen in the confusion matrix in Fig. 4.5. In contrast, the classification rates are more stable and outperforming the original classes overall. The results are by no means perfect, however it shows that the most logical direction is to use objective classes based on AUs rather than estimated emotion categories. Further investigation using an objective selection of FACS-based regions [20] supports this with AUC results for detecting relevant movements to be 0.7512 and 0.7261 on SAMM and CASME II, respectively.

Table 4.5 shows the experimental results on CASME II with each result metric being a weighted average calculation to account for imbalanced numbers within classes. Each experiment was completed for each feature and within the original classes defined in [149] and the proposed classes. To compare with the state-of-the-art 5-class emotional-based classification in CASME II testing have done to classify 5 proposed classes(I-V). In addition to that and for more information and details and because we have 7 proposed classes testing to classify 6 classes(I-VI) and 7 classes(I-VII) also has been done and reporting as shown. Both the 10-fold cross-validation results and leave-one-subject-out (LOSO) are shown.

The top performing feature achieves a weighted accuracy score of 86.35% for the HOG 3D feature in the proposed class I-V. This shows a large improvement over the original classes which achieved 80.93% for the same feature. Using LOSO, the results were comparable with the original classes. The highest accuracy was 76.60% from the HOOF feature, in the proposed I-VII classes. For the CASME II dataset results, using LBP-TOP and 10-fold cross-validation, the original method outperformed the classes I-VI and I-VII. In addition, for HOG3D LOSO, the original method outperforms in class I-VII when using F-measure as a measurement.

The experiment based on the same conditions were then repeated for SAMM and can be seen in Table 4.6. Overall the recognition rates were good for SAMM, with the best result achieving an accuracy of 81.93% using LBP-TOP in I-VI classes for 10-fold cross validation. The best result using LOSO was from the HOG 3D feature, in the proposed I-VII classes and achieved 63.93%, however due to the lower amount of micro-expressions within the SAMM dataset compared with CASME II, the LOSO results were lower.

Some results show that using LOSO, HOOF outperforms in CASME II while HOG3D outperforms in SAMM and in CASME II using LOSO, the HOOF feature achieves a higher accuracy for classes I-VII over I-VI, but not for the F-measure metric. Explanations of this comes down to the data, and how large some variations of the settings, such as resolution and capture methods, are set. The imbalance of data, specifically the low amounts of micro-expression data, can skew LOSO results with low amounts of testing and training. This shows how using LOSO for micro-expression recognition is difficult to quantify with a fair amount of significance. Further data collection of spontaneous micro-expressions is required to rectify this.

4.3.1 Evaluating objective classes on a composite database

Table 4.7 shows the baseline results and our proposed method for CDE task. Amongst the baseline methods, HOOF is outperformed in the CDE. We proved that our proposed method is outperformed the baseline methods in all the performance metrics, where it achieved F1-Score of 0.575 and accuracy of 0.718.

Table 4.7: The results for Task B based on LOSO cross validation. LBP-TOP, 3DHOG, HOOF are the baseline methods using 5x5 blocks. Proposed is our proposed method using selective block-based feature fusion representation.

Methods	Accuracy	TPR	FPR	F1-Score	AUC
3DHOG	0.663	0.498	0.287	0.441	0.446
HOOF	0.690	0.573	0.239	0.526	0.557
LBP-TOP	0.686	0.533	0.197	0.515	0.554
Proposed	**0.718**	**0.592**	**0.162**	**0.575**	**0.606**

4.4 Adaptive Mask for Region-based Facial Micro-Expression Recognition

Table 4.8 shows the result achieved by the proposed method against the state of the art. The proposed method performs better than the majority of the handcrafted methods and it is comparable to some of the deep learning methods. Amongst handcrafted methods, the proposed perform better than FMBH [94] in terms of accuracy when they evaluated their method on CASME II, although they used manually created mask to separate the background from the face, which makes it difficult to use the method in automatic systems, unlike our proposed method, which is all automatic. In addition to that, they did not evaluate their method with F1-Score. When compared with other methods in hand-crafted, we achieved the best result.

Deep learning methods in FME recognition used augmented datasets to increase the number of samples due to the need for big data when creating a deep networks. according to that there no fair compare between the proposed method which has been evaluated on original datasets and deep learning methods.

Table 4.8: Comparison between proposed method with the state-of-the-art methods on CASME II and SAMM.

Method	CASME II		SAMM	
	Accuracy	F1-Score	Accuracy	F1-Score
LBP-TOP (baseline) [154]	63.4	0.33	41.38	-
LBP-MOP [142]	66.8	-	42.72	-
HOOF [10]	44	-	46.13	-
FDM [147]	45.3	0.47	-	-
STCLQP [60]	58.39	0.57	-	-
STLBP-IP [58]	64.75	-	-	-
Bi-WOOF [92]	61.0	50	-	-
HIGO[82]	57.40	-	-	-
MDMO [93]	67.37	-	-	-
FMBH [94]	69.11	-	-	-
FHOFO [51]	56.64	0.52	-	
Proposed Method	**69.6**	**0.59**	**59.7**	**0.51**
Proposed Method + Objective Classes	**77.9**	**0.72**	-	-

Table 4.9: Results of different experiments.

Method	Accuracy	F1-Score
Global HOOF	44	46.31
14 ROIs HOOF	56.8	0.23
14 ROIs HOOF+Smoothing	58.9	0.4
Global LBP-TOP	63.4	0.33
Global LBP-TOP + (Smoothing,Mask)	61.3	0.38
8 ROIs LBP-TOP + (Smoothing,Mask)	62.2	0.56
14 ROIs LBP-TOP	63.4	0.45
14 ROIs LBP-TOP + Smoothing	63.5	0.51
14 ROIs LBP-TOP + Mask	65.09	0.55
14 ROIs LBP-TOP+ (Smoothing,Mask)	**68.2**	**0.57**
14 ROIs LBP-TOP + (Smoothing, Mask,Removing displacements) Proposed	**69.6**	**0.59**

To justify the importance of the steps used of the proposed method, we conduct ablation studies. Table 4.9 compares the effect of smoothing, adaptive mask, removing random displacements and number of ROIs on FME recognition. We observed that extracting features locally is better than globally. In addition, we found that the use of adaptive mask has improved the results, as well as the use of smoothing before creating the mask and remove random displacements after creating the mask, which has a positive effect on the result as it removes some of the noise (that can be confused with some of the micro-movements). It also shows a sample of experiment using different ROIs (8 ROIs) by combining some ROIs like AU1, AU2, AU3, and AU4, and removing some like AU11 and AU13. The experiment proved that the selected 14 ROIs is a better choice for FME recognition.

4.5 Summary

In this chapter the result of thesis contributions have shown and discussed.Firstly, Important insights for researchers has been provided in this field to consider the settings when conducting new experiment in the future. The progress in research of micro-expressions recognition can aid in the paradigm shift in affect computing for real-world applications in psychology, health study and security control.

We show that restructuring micro-expression classes objectively around the AUs, recognition results outperform the state-of-the-art, emotion-based classification approaches. As micro-expressions are so subtle, the best way to categorise is objectively as possible, so using AU codes is the most logical. Categorising using a combination of AUs and self-reports [149] can cause many conflicts when training a machine learning method. Further, dataset imbalances can be very detrimental to machine learning algorithms, and this is further emphasised with the relatively low amount of movements in both datasets.

Finally, the result of the proposed method for FME recognition has been shown. The method is a region-based, adaptive mask based on the micro-motion using optical flow have applied. The proposed method evaluated on two benchmark datasets: CASME II and SAMM and achieved a promising result which overcomes the state-of-the-art results when compared to hand-crafted approach. Comparing to the deep learning approach at this time will not be fair due to the different datasets used resulting from the augmented of these datasets.

Chapter 5

Conclusion and Future work

5.1 Conclusion

Facial Micro-expression recognition is challenging area, which can be utilized in different application such as interrogation because it can be a good cue for lie detection due to the unique features of FME(might be uncontrollable and its short duration makes it difficult to fake) rather than normal expression.

A novel method for micro-facial expressions recognition introduced, Region-Based method with 14 ROIs selected based on AUs analysis on CASMEII and SAMM. ROIs have been proposed for locally analyzed the features to avoid unimportant information of the face. Further, to be more specific to the movement related to FME adaptive mask based on the micro motion using optical flow have applied to each frame of ME, in addition to removing random displacement which caused by light condition. LBP-TOP features extracted from each region before use SMO as classifier. The proposed method evaluated on two of benchmark datasets: CASME II and SAMM, and achieved a promising result up to 69.6 on CASMEII.

The effects of resolution and frame rate changes on FME has been investigated and identified. classifying emotion labels of dataset around AUs instead of predicted emotion has been introduced, which its accuracy reaches 76.60% on CASME II.

As micro-expression recognition is still in its infancy when compared to the macro-expression, it requires combined efforts from multidisciplinary (including psychology, computer science, physiology, engineer and policy maker) to achieve reliable results for practical real-world application. A controversial point is whether or not it should be allowed to detect these micro-expressions, as the theory behind it states that the person attempting to conceal their emotion experience these movements involuntarily and likely unknowingly. If we are able to detect them with high accuracy, then we are effectively robbing a person of being able to hide something that is private to them. From an ethical point of view, knowing when someone is being deceptive would be advantageous but takes away the freedom you had in your emotions.

5.2 Future work

Further work can be done to improve FME datasets. Firstly, more datasets or expanding previous sets would be a simple improvement that can help move the research forward faster. Secondly, a standard procedure on how to maximise the amount of micro-movements induced spontaneously in laboratory controlled experiments would be beneficial. If collaboration between established datasets and researchers from psychology occurred, dataset creation would be more consistent.

Regarding to FME method more work on the alignment should be done to avoid the random displacement which caused by head-movements and didn't covered in this research.

Deep learning has emerged as a new area of machine learning research [8, 24, 2], and micro-expression analysis has yet to exploit this trend. Unfortunately, the amount of high-quality spontaneous micro-expression data is low and deep learning requires a large amount of data to work well [24]. Many video-based datasets previously used have over 10,000 video samples [72] and even over 1 million actions extracted from YouTube videos [69]. A real effort to gather spontaneous micro-expression data is required for deep learning approaches to be effective in the future.

Bibliography

[1] Shazia Afzal and Peter Robinson. Natural affect data: Collection and annotation. In *New perspectives on affect and learning technologies*, pages 55–70. Springer, 2011.

[2] Jhan Alarifi, Manu Goyal, Adrian Davison, Darren Dancey, Rabia Khan, and Moi Hoon Yap. Facial skin classification using convolutional neural networks. In *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings*, volume 10317, page 479. Springer, 2017.

[3] Benjamin Allaert, Ioan Marius Bilasco, and Chaabane Djeraba. Micro and macro facial expression recognition using advanced local motion patterns. *IEEE Transactions on Affective Computing*, 2019.

[4] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[5] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[6] Xianye Ben, Xitong Jia, Rui Yan, Xin Zhang, and Weixiao Meng. Learning effective binary descriptors for micro-expression recognition transferred by macro-information. *Pattern Recognition Letters*, 2017.

[7] Xianye Ben, Peng Zhang, Rui Yan, Mingqiang Yang, and Guodong Ge. Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation. *Neural Computing and Applications*, 27(8):2629–2646, 2016.

[8] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[9] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE, 2009.

[11] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and Rene Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939, June 2009.

[12] Mengting Chen, Heather T Ma, Jie Li, and Huanhuan Wang. Emotion recognition using fixed length micro-expressions sequence and weighting method. In *Real-time Computing and Robotics (RCAR), IEEE International Conference on*, pages 427–430. IEEE, 2016.

[13] Jen-Tzung Chien and Chia-Chen Wu. Linear discriminant analysis (lda). 2005.

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

[15] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[17] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[18] 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(1):116–129, Jan 2018.

[19] Adrian Davison, Walied Merghani, Cliff Lansley, Choon-Ching Ng, and Moi Hoon 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)*, pages 642–649. IEEE, 2018.

[20] Adrian Davison, Walied Merghani, Cliff Lansley, Choon-Ching Ng, and Moi Hoon Yap. Objective micro-facial movement detection using facs-based regions and baseline evaluation. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 642–649. IEEE, 2018.

[21] Adrian K Davison, Moi Hoon Yap, Nicholas Costen, Kevin Tan, Cliff Lansley, and Daniel Leightley. Micro-facial movements: An investigation on spatio-temporal descriptors. In *Computer Vision-ECCV 2014 Workshops*, pages 111–123. Springer, 2014.

[22] Adrian K Davison, Moi Hoon Yap, and Cliff Lansley. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 1864–1869. IEEE, 2015.

[23] Adrian K. Davison, Moi Hoon Yap, and Cliff Lansley. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 1864–1869, Oct 2015.

[24] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[25] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[26] Xiaodong Duan, Qiguo Dai, Xinhan Wang, Yuangang Wang, and Zhichao Hua. Recognizing spontaneous micro-expression from eye region. *Neurocomputing*, 217:27–36, 2016.

[27] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[28] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton, 2001.

[29] Paul Ekman. *Emotions Revealed: Understanding Faces and Feelings*. Phoenix, 2004.

[30] Paul Ekman. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009.

[31] Paul Ekman. Lie catching and microexpressions. In Clancy W. Martin, editor, *The Philosophy of Deception*, pages 118–133. Oxford University Press, 2009.

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

[33] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[34] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[35] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[36] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

[37] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

[38] Paul Ekman and Erika L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Series in Affective Science. Oxford University Press, 2005.

[39] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[40] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[41] Mark Frank, Malgorzata Herbasz, Kang Sinuk, Amy Marie Keller, Anastacia Kurylo, and Courtney Nolan. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In *International Communication Association*, 2009.

[42] Mark G Frank, Carl J Maccario, and Venugopal l Govindaraju. Behavior and security. In *Protecting airline passengers in the age of terrorism*. Greenwood Pub. Group, 2009.

[43] Mark G Frank, Carl J Maccario, and Venugopal l Govindaraju. Behavior and security. In *Protecting airline passengers in the age of terrorism*. Greenwood Pub. Group, 2009.

[44] MG Frank, M Herbasz, K Sinuk, A Keller, and C Nolan. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In *The Annual Meeting of the International Communication Association. Sheraton New York, New York City*, 2009.

[45] YS Gan and Sze-Teng Liong. Bi-directional vectors from apex in cnn for micro-expression recognition. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 168–172. IEEE, 2018.

[46] YS Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019.

[47] Yanjun Guo, Yantao Tian, Xu Gao, and Xuange Zhang. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 3473–3479. IEEE, 2014.

[48] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *Image Processing, IEEE Transactions on*, 19(6):1657–1663, 2010.

[49] Ernest A Haggard and Kenneth S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*, pages 154–165. Springer, 1966.

[50] Xiao-li Hao and Miao Tian. Deep belief network based on double weber local descriptor in micro-expression recognition. In *Advanced Multimedia and Ubiquitous Engineering*, pages 419–425. Springer, 2017.

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

[52] SL Happy and Aurobinda Routray. Recognizing subtle micro-facial expressions using fuzzy histogram of optical flow orientations and feature selection methods. In *Computational Intelligence for Pattern Recognition*, pages 341–368. Springer, 2018.

[53] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[54] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1208–1213. IEEE, 2005.

[55] Hanns C Hopf, Wibke Muller-Forell, and Nikolai J Hopf. Localization of emotional and volitional facial paresis. *Neurology*, 42(10):1918–1918, 1992.

[56] Chris House and Rachel Meyer. Preprocessing and descriptor features for facial micro-expression recognition. 2015.

[57] Chunlong Hu, Dengbiao Jiang, Haitao Zou, Xin Zuo, and Yucheng Shu. Multi-task micro-expression recognition combining deep and handcrafted features. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 946–951. IEEE, 2018.

[58] Xiaohua Huang, Su-Jing Wang, Guoying Zhao, and Matti Piteikainen. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[59] Xiaohua Huang, Sujing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng, and Matti Pietikainen. Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection. *arXiv preprint arXiv:1608.02255*, 2016.

[60] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 175:564–578, 2016.

[61] Mohammad Shahidul Islam et al. Local gray code pattern (lgcp): A robust feature descriptor for facial expression recognition. *International Journal of Science and Research (IJSR), India Online ISSN*, pages 2319–7064, 2013.

[62] Deepak Kumar Jain, Zhang Zhang, and Kaiqi Huang. Random walk-based feature learning for micro-expression recognition. *Pattern Recognition Letters*, 2018.

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

[64] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[65] Xitong Jia, Xianye Ben, Hui Yuan, Kidiyo Kpalma, and Weixiao Meng. Macro-to-micro transformation model for micro-expression recognition. *Journal of Computational Science*, 25:289–297, 2018.

[66] Mihailo R Jovanović, Peter J Schmid, and Joseph W Nichols. Sparsity-promoting dynamic mode decomposition. *Physics of Fluids (1994-present)*, 26(2):024103, 2014.

[67] Siti Khairuni Amalina Kamarol, Nor Syazana Meli, Mohamed Hisham Jaward, and Nader Kamrani. Spatio-temporal texture-based feature extraction for spontaneous facial expression recognition. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 467–470. IEEE, 2015.

[68] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[69] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[70] Huai-Qian Khor, John See, Sze-Teng Liong, Raphael CW Phan, and Weiyao Lin. Dual-stream shallow networks for facial micro-expression recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 36–40. IEEE, 2019.

[71] Huai-Qian Khor, John See, Raphael Chung Wei Phan, and Weiyao 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)*, pages 667–674. IEEE, 2018.

[72] Amir Roshan Zamir Khurram Soomro and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

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

[74] Satoshi Kimura and Masahiko Yachida. Facial expression recognition and its degree estimation. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 295–300. IEEE, 1997.

[75] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[77] Anh Cat Le Ngo, Sze-Teng Liong, John See, and Raphael Chung-Wei Phan. Are subtle expressions too sparse to recognize? In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 1246–1250. IEEE, 2015.

[78] Anh Cat Le Ngo, Raphael Chung-Wei Phan, and John See. Spontaneous subtle expression recognition: Imbalanced databases and solutions. In *Computer Vision–ACCV 2014*, pages 33–48. Springer, 2014.

[79] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[80] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[82] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Reading hidden emotions: Spontaneous micro-expression spotting and recognition. *arXiv preprint arXiv:1511.00423*, 2015.

[83] Xiaobai Li, Thorsten Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[84] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *10th IEEE International Conference on automatic Face and Gesture Recognition*, 2013.

[85] Yante Li, Xiaohua Huang, and Guoying Zhao. Can micro-expression be recognized based on single apex frame? In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3094–3098. IEEE, 2018.

[86] Chenhan Lin, Fei Long, JianMing Huang, and Jun Li. Micro-expression recognition based on spatiotemporal gabor filters. In *2018 Eighth International Conference on Information Science and Technology (ICIST)*, pages 487–491. IEEE, 2018.

[87] Sze-Teng Liong, YS Gan, John See, Huai-Qian Khor, and Yen-Chang 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)*, pages 1–5. IEEE, 2019.

[88] Sze-Teng Liong, John See, Raphael C-W Phan, Anh Cat Le Ngo, Yee-Hui Oh, and KokSheik Wong. Subtle expression recognition using optical strain weighted features. In *Asian Conference on Computer Vision*, pages 644–657. Springer, 2014.

[89] Sze-Teng Liong, John See, Raphael C-W Phan, Yee-Hui Oh, Anh Cat Le Ngo, Kok-Sheik Wong, and Su-Wei Tan. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Processing: Image Communication*, 47:170–182, 2016.

[90] Sze-Teng Liong, John See, Raphael C-W Phan, KokSheik Wong, and Su-Wei Tan. Hybrid facial regions extraction for micro-expression recognition system. *Journal of Signal Processing Systems*, 90(4):601–617, 2018.

[91] Sze-Teng Liong, John See, Raphael Chung-Wei Phan, and KokSheik Wong. Less is more: Micro-expression recognition from video using apex frame. *arXiv preprint arXiv:1606.01721*, 2016.

[92] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018.

[93] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transaction of Affective Computing*, 2015.

[94] Hua Lu, Kidiyo Kpalma, and Joseph Ronsin. Motion descriptors for micro-expression recognition. *Signal Processing: Image Communication*, 67:108–117, 2018.

[95] Zhaoyu Lu, Ziqi Luo, Huicheng Zheng, Jikai Chen, and Weihong Li. A delaunay-based temporal coding model for micro-expression recognition. In *Asian Conference on Computer Vision*, pages 698–711. Springer, 2014.

[96] Bruce D Lucas. Generalized image matching by the method of differences. 1986.

[97] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[98] Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

[99] David Matsumoto, Seung Hee Yoo, and Sanae Nakagawa. Culture, emotion regulation, and adjustment. *Journal of personality and social psychology*, 94(6):925, 2008.

[100] M McCabe. Best practice recommendation for the capture of mugshots. *http://www. itl. nist. gov/iaui/894.03/face/bprmug3. htm*, 2009.

[101] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[102] Zhiheng Niu, Shiguang Shan, Shengye Yan, Xilin Chen, and Wen Gao. 2d cascaded adaboost for eye localization. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 1216–1219. IEEE, 2006.

[103] Yee-Hui Oh, Anh Cat Le Ngo, Raphael C-W Phari, John See, and Huo-Chong Ling. Intrinsic two-dimensional local structures for micro-expression recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 1851–1855. IEEE, 2016.

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

[105] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, Jul 2002.

[106] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[107] Maureen O'Sullivan, Mark G Frank, Carolyn M Hurley, and Jaspreet Tiwana. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530, 2009.

[108] Maureen O'Sullivan, Mark G Frank, Carolyn M Hurley, and Jaspreet Tiwana. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530, 2009.

[109] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445, 2000.

[110] Sung Yeong Park, Seung Ho Lee, and Yong Man Ro. Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 911–914. ACM, 2015.

[111] Devangini Patel, Guoying Zhao, and Matti Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *Advanced Concepts for Intelligent Vision Systems*, pages 369–380. Springer, 2015.

[112] Min Peng, Chongyang Wang, Tao Bi, Yu Shi, XiangDong Zhou, and Tong Chen. A novel apex-time network for cross-dataset micro-expression recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–6. IEEE, 2019.

[113] Min Peng, Chongyang Wang, Tong Chen, Guangyuan Liu, and Xiaolan Fu. Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in psychology*, 8:1745, 2017.

[114] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 868–875. IEEE, 2011.

[115] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1449–1456. IEEE, 2011.

[116] John Platt et al. Fast training of support vector machines using sequential minimal optimization. 1999.

[117] Senya Polikovsky and Yoshinari Kameda. Facial micro-expression detection in hi-speed video based on facial action coding system (facs). *IEICE transactions on information and systems*, 96(1):81–92, 2013.

[118] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*, pages 1–6. IET, 2009.

[119] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. Cas (me)ˆ 2: A database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 2017.

[120] James A Russell and Jose Miguel Fernández-Dols. *The psychology of facial expression*. Cambridge university press, 1997.

[121] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010.

[122] Xun-Bing Shen, Qi Wu, and Xiao-Lan Fu. Effects of the duration of expressions on the recognition of microexpressions. *Journal of Zhejiang University SCIENCE B*, 13(3):221–230, 2012.

[123] Matthew Shreve, Jesse Brizzi, Sergiy Fefilatyev, Timur Luguev, Dmitry Goldgof, and Sudeep Sarkar. Automatic expression spotting in videos. *Image and Vision Computing*, 32(8):476 – 486, 2014.

[124] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 51–56, 2011.

[125] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 51–56, 2011.

[126] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[127] Baolin Song, Ke Li, Yuan Zong, Jie Zhu, Wenming Zheng, Jingang Shi, and Li Zhao. Recognizing spontaneous micro-expression using a three-stream convolutional neural network. *IEEE Access*, 7:184537–184551, 2019.

[128] Yale Song, Louis-Philippe Morency, and Randall 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*, pages 237–244. ACM, 2013.

[129] Lip-based Speaker Authentication Spatiotemporal. Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication. 2011.

[130] Wenchao Su, Yanyan Wang, Fei Su, and Zhicheng Zhao. Micro-expression recognition based on the spatio-temporal feature. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2018.

[131] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[132] BMS Bahar Talukder, Brinta Chowdhury, Tamanna Howlader, and SM Mahbubur Rahman. Intelligent recognition of spontaneous expression using motion magnification of spatio-temporal data. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 114–128. Springer, 2016.

[133] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[134] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[135] Lei Wang, Hai Xiao, Sheng Luo, Jie Zhang, and Xiyao Liu. A weighted feature extraction method based on temporal accumulation of optical flow for micro-expression recognition. *Signal Processing: Image Communication*, 78:246–253, 2019.

[136] Su-Jing Wang, Hui-Ling Chen, Wen-Jing Yan, Yu-Hsin Chen, and Xiaolan Fu. Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural processing letters*, 39(1):25–43, 2014.

[137] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4678–4683. IEEE, 2014.

[138] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. *IEEE Transactions on Image Processing*, 24(12):6034–6047, 2015.

[139] Su-Jing Wang, Wen-Jing Yan, Tingkai Sun, Guoying Zhao, and Xiaolan Fu. Sparse tensor canonical correlation analysis for micro-expression recognition. *Neurocomputing*, 214:218–232, 2016.

[140] Yandan Wang, John See, Yee-Hui Oh, Raphael C-W Phan, Yogachandran Rahulamathavan, Huo-Chong Ling, Su-Wei Tan, and Xujie Li. Effective recognition of facial micro-expressions with video motion magnification. *Multimedia Tools and Applications*, pages 1–26, 2016.

[141] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Asian Conference on Computer Vision*, pages 525–537. Springer, 2014.

[142] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PloS one*, 10(5):e0124674, 2015.

[143] Gemma Warren, Elizabeth Schertler, and Peter Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69, 2009.

[144] Qi Wu, Xunbing Shen, and Xiaolan Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. In *Affective Computing and Intelligent Interaction*, pages 152–162. Springer, 2011.

[145] Zhaoqiang Xia, Xiaoyi Feng, Xiaopeng Hong, and Guoying Zhao. Spontaneous facial micro-expression recognition via deep convolutional network. In *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2018.

[146] Zhaoqiang Xia, Huan Liang, Xiaopeng Hong, and Xiaoyi Feng. Cross-database micro-expression recognition with deep convolutional networks. In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*, pages 56–60, 2019.

[147] Feng Xu, Junping Zhang, and James Z Wang. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing*, 8(2):254–267, 2017.

[148] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):40–51, 2007.

[149] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE*, 9(1):e86041, 01 2014.

[150] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014.

[151] Wen-Jing Yan, Su-Jing Wang, Yong-Jin Liu, Qi Wu, and Xiaolan Fu. For micro-expression recognition: Database and suggestions. *Neurocomputing*, 136:82–87, 2014.

[152] Wen-Jing Yan, Su-Jing Wang, Yong-Jin Liu, Qi Wu, and Xiaolan Fu. For micro-expression recognition: Database and suggestions. *Neurocomputing*, 136:82–87, 2014.

[153] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37(4):217–230, 2013.

[154] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[155] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *IEEE conference on automatic face and gesture recognition*, 2013.

[156] Moi Hoon Yap, John See, Xiaopeng Hong, and Su-Jing Wang. Facial micro-expressions grand challenge 2018 summary. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 675–678. IEEE, 2018.

[157] Moi Hoon Yap, Hassan Ugail, and Reyer Zwiggelaar. Facial behavioral analysis: A case study in deception detection. *British Journal of Applied Science & Technology*, 4(10):1485, 2014.

[158] Moi Hoon Yap, Hassan Ugail, and Reyer Zwiggelaar. Facial behavioral analysis: A case study in deception detection. *British Journal of Applied Science & Technology*, 4(10):1485, 2014.

[159] Xinhe Yu, Zhihua Xie, and Wenjun Zong. Dual-cross patterns with rpca of key frame for facial micro-expression recognition. In *International Conference on Image and Graphics*, pages 750–759. Springer, 2019.

[160] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[161] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[162] Peng Zhang, Xianye Ben, Rui Yan, Chen Wu, and Chang Guo. Micro-expression recognition system. *Optik-International Journal for Light and Electron Optics*, 127(3):1395–1400, 2016.

[163] Shiyu Zhang, Bailan Feng, Zhineng Chen, and Xiangsheng Huang. Micro-expression recognition by aggregating local spatio-temporal patterns. In *International Conference on Multimedia Modeling*, pages 638–648. Springer, 2017.

[164] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.

[165] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.

[166] Yue Zhao and Jiancheng Xu. Necessary morphological patches extraction for automatic micro-expression recognition. *Applied Sciences*, 8(10):1811, 2018.

[167] Hao Zheng. Micro-expression recognition based on 2d gabor filter and sparse representation. In *Journal of Physics: Conference Series*, volume 787, page 012013. IOP Publishing, 2017.

[168] Hao Zheng, Xin Geng, and Zhongxue Yang. A relaxed k-svd algorithm for spontaneous micro-expression recognition. In *Pacific Rim International Conference on Artificial Intelligence*, pages 692–699. Springer, 2016.

[169] Ruicong Zhi, Mengyi Liu, Hairui Xu, and Ming Wan. Facial micro-expression recognition using enhanced temporal feature-wise model. In *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*, pages 301–311. Springer, 2019.

[170] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[171] Ling Zhou, Qirong Mao, and Luoyang Xue. Cross-database micro-expression recognition: A style aggregated and attention transfer approach. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 102–107. IEEE, 2019.

[172] Xuena Zhu, Xianye Ben, Shigang Liu, Rui Yan, and Weixiao Meng. Coupled source domain targetized with updating tag vectors for micro-expression recognition. *Multimedia Tools and Applications*, 77(3):3105–3124, 2018.

[173] Yuan Zong, Xiaohua Huang, Wenming Zheng, Zhen Cui, and Guoying Zhao. Learning a target sample re-generator for cross-database micro-expression recognition. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 872–880. ACM, 2017.

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

[175] Yuan Zong, Wenming Zheng, Xiaohua Huang, Jingang Shi, Zhen Cui, and Guoying Zhao. Domain regeneration for cross-database micro-expression recognition. *IEEE Transactions on Image Processing*, 27(5):2484–2498, 2018.