

**DEEP LEARNING BASED REAL-TIME
SEQUENTIAL FACIAL EXPRESSION ANALYSIS
USING GEOMETRIC FEATURES**

**A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of**

MASTER OF SCIENCE

in Electronics and Communication Engineering

**by
Talha Enes KÖKSAL**

July 2023

İZMİR

We approve the thesis of **Talha Enes KÖKSAL**

Examining Committee Members:

Asst. Prof. Dr. Abdurrahman GÜMÜŞ

Department of Electrical and Electronics Engineering, İzmir Institute of Technology

Prof. Dr. Bilge KARAÇALI

Department of Electrical and Electronics Engineering, İzmir Institute of Technology

Prof. Dr. Devrim ÜNAY

Department of Electrical and Electronics Engineering, İzmir Democracy University

18 July 2023

Asst. Prof. Dr. Abdurrahman GÜMÜŞ

Supervisor, Department of Electrical and
Electronics Engineering

İzmir Institute of Technology

Prof. Dr. Mustafa Aziz ALTINKAYA

Head of the Department of
Electrical and Electronics Engineering

Prof. Dr. Mehtap EANES

Dean of the Graduate School of
Engineering and Science

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Asst. Prof. Dr. Abdurrahman GÜMÜŞ for his invaluable guidance, patience, and support throughout my research. His insightful comments and suggestions helped me to navigate the challenges of my research and achieve my goals.

I would also like to extend a special thanks to my wife Büşra, who has been my unwavering source of motivation. Her constant love, encouragement, and understanding have been essential to my success. I am particularly grateful for her support during this time as she carries our daughter to be. I am excited to welcome our little one into the world and look forward to sharing in the joys of parenthood.

ABSTRACT

DEEP LEARNING BASED REAL-TIME SEQUENTIAL FACIAL EXPRESSION ANALYSIS USING GEOMETRIC FEATURES

In this thesis, macro and micro facial expression sequences from various datasets are trained using neural networks to classify them in one of the basic emotions. In macro expression experiments, for each frame of the sequences facial landmarks are extracted using MediaPipe FaceMesh solution and geometric features using both spatial and temporal information based on these landmarks are created. To classify the features, ConvLSTM2D followed by multilayer perceptron blocks are used. In order to achieve real time classification performance, all algorithms are implemented compatible to run on GPU. The proposed method for macro expressions is tested with CK+, Oulu-CASIA VIS, Oulu-CASIA NIR and MMI datasets. In micro expression experiments, apart from geometric features also blendshape features provided by MediaPipe are used. In order to improve classification performance, Phase-Based Video Motion Processing technique is used to magnify subtle facial movements of micro expressions. Experiments are conducted separately on same classification layers that consist of ConvLSTM1D followed by multilayer perceptron blocks. The proposed method for micro expressions is tested with SAMM and CASME II datasets. The datasets utilized in this study were accessed upon signing corresponding license agreements. Each dataset is specifically designated for academic purposes and is made available under these agreements. Only data from subjects who provided consent for their information to be used in publications was included in the thesis. The license agreements for each dataset can be found in the appendices section.

ÖZET

DERİN ÖĞRENME TABANLI GEOMETRİK ÖZELLİKLERİ KULLANARAK GERÇEK ZAMANLI SIRALI YÜZ İFADESİ ANALİZİ

Bu tezde, çeşitli veri setlerinden makro ve mikro yüz ifadesi dizileri, temel duygulardan birinde sınıflandırmak için sinir ağları kullanılarak eğitilmiştir. Makro ifade deneylerinde, dizilerin her bir karesi için MediaPipe FaceMesh çözümü kullanılarak yüz işaretleri çıkarılır ve bu noktalara dayalı olarak hem uzamsal hem de zamansal bilgiler kullanılarak geometrik özellikler oluşturulur. Öznitelikleri sınıflandırmak için ConvLSTM2D ve ardından çok katmanlı algılayıcı blokları kullanılır. Gerçek zamanlı sınıflandırma performansı elde etmek için, tüm algoritmalar GPU üzerinde çalışacak şekilde uyarlanmıştır. Makro ifadeler için önerilen yöntem CK+, Oulu-CASIA VIS, Oulu-CASIA NIR ve MMI veri setleri ile test edilmiştir. Mikro ifade deneylerinde geometrik özelliklerin yanı sıra MediaPipe tarafından sağlanan blendshape özellikleri de kullanılmaktadır. Sınıflandırma performansını iyileştirmek için, mikro ifadelerin ince yüz hareketlerini büyütme için Faz Tabanlı Video Hareket İşleme tekniği kullanılır. Deneyler, ConvLSTM1D'yi takip eden çok katmanlı algılayıcı bloklardan oluşan aynı sınıflandırma katmanları üzerinde ayrı ayrı yürütülür. Mikro ifadeler için önerilen yöntem, SAMM ve CASME II veri setleri ile test edilmiştir. Bu çalışmada kullanılan veri setlerine, ilgili lisans sözleşmelerinin imzalanmasından sonra erişilmiştir. Her veri setinin akademik amaçlar için kullanmaya uygunluğu sözleşmelerde belirtilmiş ve bu anlaşmalar kapsamında kullanmıştır. Tezde sadece bilgilerinin yayınlarda kullanılmasına izin veren kişilerden elde edilen verilere yer verilmiştir. Her veri seti için lisans sözleşmeleri ekler bölümünde bulunmaktadır.

TABLE OF CONTENTS

LIST OF FIGURES	viii
LIST OF TABLES	xi
CHAPTER 1. INTRODUCTION	1
1.1. Motivation	3
1.1.1. Contributions.....	5
CHAPTER 2. REAL-TIME SEQUENTIAL MACRO EXPRESSION RECOGNITION USING GEOMETRIC FEATURES	7
2.1. Introduction.....	7
2.2. Literature Review.....	7
2.3. Methodology	9
2.3.1. Computational Setup	9
2.3.2. Datasets	9
2.3.3. Feature Creation Algorithm	12
2.3.4. Classification Algorithm.....	16
2.4. Results and Discussion	19
2.4.1. Experiments with Datasets Individually	20
2.4.2. Composite Dataset Experiments	24
2.4.3. Real-time Implementation	26
2.5. Conclusion.....	28
CHAPTER 3. SEQUENTIAL MICRO EXPRESSION RECOGNITION USING GEOMETRIC FEATURES	30
3.1. Introduction.....	30
3.2. Literature Review.....	30
3.3. Methodology	32
3.3.1. Computational Setup	32
3.3.2. Datasets	32
3.3.3. Feature Creation Algorithm	33
3.3.3.1. Phase-Based Video Motion Processing	36
3.3.4. Classification Algorithm.....	41
3.4. Results and Discussion	42

3.5. Conclusion.....	43
CHAPTER 4. CONCLUSION	45
REFERENCES	47
APPENDICES.....	52
APPENDIX A. LICENSE AGREEMENT OF CK+ DATASET.....	52
APPENDIX B. LICENSE AGREEMENT OF OULU-CASIA DATASET	53
APPENDIX C. LICENSE AGREEMENT OF MMI DATASET	55
APPENDIX D. LICENSE AGREEMENT OF CASME II DATASET	57
APPENDIX E. LICENSE AGREEMENT OF SAMM DATASET	58

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
Figure 1.1. List of action units described in Facial Action Coding System (FACS) (Source: Barrett et al., 2019).....	4
Figure 1.2. Comparison of macro and micro expressions displaying happiness emotion at their peak intensity. (a) Macro expression example taken from CK+ dataset (Source: Lucey et al., 2010). (b) Micro expression example taken from CASME II dataset (Source: Yan et al., 2014).....	5
Figure 2.1. Collage of images in datasets which are used in macro-expression experiments. (a) Cohn-Kanade (CK+) dataset (Source: Lucey et al., 2010). (b) Oulu-CASIA dataset Near Infrared (NI) variation (Source: Zhao et al., 2011). (c) Oulu-CASIA dataset Visible Light (VIS) variation (Source: Zhao et al., 2011). (d) MMI dataset, only subjects that have sequential data is used (Source: Pantic et al., 2005).....	11
Figure 2.2. Feature creation algorithm flow for macro-expression experiments. Sequential camera frames are fed into MediaPipe FaceMesh Solution get extract facial landmark positions. For each selected landmark pair, Euclidean distance and angle θ features are calculated. Neutral frame features are taken as basis and at each frame after neutral frame, calculated features are subtracted with basis features. These two features are concatenated to create feature vector. Images of the subject are taken from CK+ dataset (Source: Lucey et al., 2010)	15
Figure 2.3. Classification algorithm for macro-expression experiments. First $N - 1, A$ shaped array that holds feature vectors for whole sequence is scaled using standard scaler. Scaled 1D data is converted to image format that will be feed into ConvLSTM2D block. Output of ConvLSTM2D block is flattened and data is classified using multi-layer perceptron layers. Where, n : frame count, e : emotion count, a : feature count, b : $\text{ceil}(\text{sqrt}(a))$	16
Figure 2.4. Inner structure of ConvLSTM cell	18

Figure 2.5.	Tracking mean value of Euclidean distance features of frames. Y axis shows pixel difference from neutral state, x axis shows frame number starting from onset (frame number 0) till offset. Images below plot shows actual images that these features are extracted. Frames number in between 0 to 10 corresponds to onset phase, 10 to 40 corresponds to apex phase and after 40 offset phase starts. Images of the subject are taken from MMI dataset (Source: Pantic et al., 2005)	20
Figure 2.6.	Selected facial landmark sets to be used in macro-expression experiments. 61, 122 and 250 landmark points are selected manually from 478 facial landmark.	21
Figure 2.7.	Accuracy box-plots of datasets used in macro-expression experiments. Results for all six experiments that are combinations of 61, 122, 250 point landmarks with AU grouping and without any grouping. Red dashed line shows mean value of accuracy for 5-fold cross validation.	22
Figure 2.8.	Confusion matrices of datasets used in macro-expression experiments. Results are created with average values of 5-fold cross validation.	23
Figure 2.9.	Real time prediction results. First frame is taken in between neutral and onset phases. Second frame shows onset phase where facial expression is started. Third frame shows apex phase where emotion is predicted with 99.97% accuracy. Image sequence is taken from MMI dataset (Source: Pantic et al., 2005)	27
Figure 3.1.	Collage of images in datasets which are used in micro-expression experiments. (a) SAMM dataset (Source: Davison et al., 2016). (b) CASME II dataset (Source: Yan et al., 2014).	33
Figure 3.2.	PBVM is applied for subject 006_1_2 in SAMM dataset. Subject demonstrates anger emotion and Brow Lowerer (AU 4) movement is enhanced using PBVM. Image of the subject is taken from SAMM dataset (Source: Davison et al., 2016)	34
Figure 3.3.	The algorithm flow for micro-expression experiments involves preprocessing each frame and feeding it into MediaPipe Face Landmarker to gather facial landmark positions and blendshape scores. Out of all the facial landmark positions, 61 are selected and Euclidean distance features are created from them. Informative blendshape scores are selected and blendshape features are created from them. These Euclidean distance features and blendshape features are then used as input for classification model. Image of the subject is taken from SAMM dataset (Source: Davison et al., 2016)	35

Figure 3.4.	PBVM method involves analyzing local phase signals over time in different spatial scales and orientations using complex steerable pyramids. The amplitude of local wavelets is separated from their phase, and the phases are temporally filtered independently at each location, orientation, and scale. Spatial smoothing can be applied to increase the phase signal-to-noise ratio, which improves the results. The temporally-bandpassed phases are then amplified or attenuated, and the video is reconstructed (Source: Wadhwa et al., 2013)	37
Figure 3.5.	ROI extraction using facial landmarks and minimum bounding rectangle function of OpenCV. Image of the subject is taken from SAMM dataset (Source: Davison et al., 2016)	38
Figure 3.6.	Mediapipe blendshape scores for original apex frame of the subject seen in Figure 3.2 (a)	39
Figure 3.7.	Mediapipe blendshape scores for PBVM applied apex frame of the subject seen in Figure 3.2 (b)	40
Figure 3.8.	Classification algorithm for micro-expression experiments. First (n, a) shaped array that holds feature vectors for whole sequence is scaled using quantile transformer. Scaled 1D data is fed into ConvLSTM1D block. Output of ConvLSTM1D block is flatten and data is classified using multi-layer perceptron layers. Where, n : frame count, e : emotion count, a : feature count	41
Figure 3.9.	Confusion matrices for micro-expression experiments using Euclidean distance features	43
Figure 3.10.	Emotions and number of subjects mapping	43

LIST OF TABLES

<u>Table</u>		<u>Page</u>
Table 2.1.	Summary of datasets used in macro expression experiments	10
Table 2.2.	Relation between emotion, action units and category mapping. Correlation between action units and resulted emotion is proposed by (Source: Friesen, Ekman, et al., 1983)	13
Table 2.3.	Landmark grouping categories (Source: Buhari et al., 2020). Dividing facial landmarks into meaningful categories reduces total landmark pair counts that will be processed by feature creation algorithm.	14
Table 2.4.	Comparison of processing times to create features for GPU and CPU execution. Full means using all landmark pairs and AU means using selected landmark pairs based on FACS based method.	24
Table 2.5.	Comparison of accuracy of different methods in the literature which use only geometric features	24
Table 3.1.	Comparison of accuracy of different methods in the literature for micro-expression recognition which use only geometric features	31
Table 3.2.	List of predicted blendshapes	36
Table 3.3.	Accuracy table for micro-expression experiments	42

CHAPTER 1

INTRODUCTION

Humans possess the unique ability to communicate emotions through their facial expressions, which are considered one of the most powerful, natural, and universal forms of expression (Li and Deng, 2020). Connection between emotions and facial expression was found in a collaborative study with Ekman, Levenson and Friesen. In that study it is discovered that performing certain facial muscular actions generates emotion physiology (Ekman, 1992b). In 1972, Ekman et al. conducted a review of prior research on the interpretation of facial expressions in western cultures and discovered that all studies found evidence of six basic emotions: happiness, surprise, fear, sadness, anger, and disgust with a hint of contempt. They observed that in some cultures, fear and surprise can be identical and hard to classify (Ekman, 1992a).

The analysis of facial expressions relies on the extraction of specific features, typically categorized into two types of according to their feature representations: spatial and spatio-temporal (Li and Deng, 2020). Spatial features represent information derived from static images, like a single photo capturing a distinct facial expression that reveals a specific emotion such as joy, anger, or surprise. On the other hand, spatio-temporal features encapsulate data extracted from a series of images, akin to a video that sequences a person's emotional display over time. This method captures the dynamic progression of facial movements, offering a comprehensive and nuanced insight into how a person's emotional state might evolve. For instance, it could trace the transition from surprise to delight, or from calm to fury, which helps to track the temporal development and complexity of emotions (Pantic and Patras, 2006).

The process of facial expression recognition task includes four main steps (Sharma, Singh, and Gautam, 2019). The first of these is pre-processing, which includes detecting the face within an image or a series of images. Once the face is detected, the process advances to the second step: generating additional facial content, such as identifying and mapping the facial landmarks. These landmarks, which include key features like the eyes, nose, mouth, and contour of the face, provide a detailed facial structure that becomes instrumental in the subsequent stage of feature extraction. The third step is feature extraction, where the system isolates important attributes from the face using the generated landmarks. This captures the unique aspects of each facial expression and prepares the data for the final stage. In the concluding step, emotion classification, the system interprets the extracted facial features, assigning them to specific emotional states,

which could range from basic emotions like happiness, sadness, or anger, to more complex emotional nuances (Xie et al., 2022).

There are three commonly used techniques in facial feature extraction that are prevalent in the literature: geometric, appearance-based, and motion-based methods (Kumari, Rajesh, and Pooja, 2015; Mollahosseini, Chan, and Mahoor, 2016). Appearance-based methods, one of the foremost approaches, utilize a pixel-based approach to extract facial features. State-of-the-art techniques often incorporate attributes such as pixel intensities, Gabor filters, Local Binary Patterns (LBP), Local Phase Quantization (LPQ), and Histogram of Oriented Gradients (HoG) to obtain information about the face (Mollahosseini, Chan, and Mahoor, 2016). Meanwhile, motion-based methods focus on characteristics related to movement, such as shifts in position and shape. These alterations are predominantly driven by the contractions and relaxations of facial muscles during emotional expressions (Zhang and Tjondronegoro, 2011). Techniques in this domain might involve optical flow, Motion History Images (MHI), and volume LBP to capture these dynamic changes (Mollahosseini, Chan, and Mahoor, 2016).

In feature extraction perspective, Convolutional Neural Network (CNN) which is widely used by deep learning frameworks also utilized as the most common methodology to extract features for pixel centric approaches (Aloysius and Geetha, 2017). Geometric features that are extracted with the help of landmarks are often mathematical attributes like Euclidean distance, slope, angle and coordinates of landmarks (Álvarez et al., 2018; Buhari et al., 2020; Khan, 2018; Qiu and Wan, 2019; Rohith Raj et al., 2020; Sharma, Singh, and Gautam, 2019).

One of the key advantages of geometric features is their ease of computation, as they require relatively less processing power compared to more complex feature extraction methodologies such as CNNs. This results in faster processing of frames in a sequence, making geometric features an attractive choice for real-time applications. Furthermore, geometric features are highly robust to unwanted disruptions in the facial image, such as variations in illumination, rotation, and misalignment. These disruptions can often confound conventional pixel-centric and motion-based features, leading to reduced accuracy and reliability in facial expression recognition tasks. By contrast, geometric features are able to circumvent such disruptions by focusing on the underlying structure of the face, resulting in more accurate and reliable recognition of facial expressions.

Given the critical role of facial expressions in understanding emotions (Ekman, 1992b), the study of emotion is highly dependent on the measurement of facial expressions, leading to the development of several observer-based systems. Among these systems, the Facial Action Coding System (FACS) stands out as the most extensively used and recognized for its comprehensive methodology, psychometric rigor, and broad applicability

across diverse scenarios (Cohn, Ambadar, and Ekman, 2007). FACS is the most commonly utilized scheme for breaking down facial expressions into their individual muscle movements, referred to as Action Units (AUs). FACS enables the description of any facial expression as a combination of specific Action Units (AUs) seen in Figure 1.1, providing a systematic approach for analyzing and understanding the complexities of facial expressions. Ekman and Friesen initially proposed the FACS and later updated it in 2002 to account for micro-expressions (Xie et al., 2022).

Macro-expressions are commonly observed during daily interactions. Typically, lasting between 0.5 to 4 seconds, they manifest with noticeable visibility and intensity. In contrast, micro-expressions are fleeting, existing for no longer than half a second and can easily be overlooked without focused attention. This duration and intensity differentiate these two types of expressions. Generally, macro-expressions present themselves with higher visibility and intensity than micro-expressions, making them easier to recognize (Xie et al., 2022). Micro-expressions, however, are brief and often involuntary facial expressions. They commonly surface when individuals attempt to conceal their true emotions, especially under high-stress situations. Their ephemeral and unconscious nature makes them an important topic in understanding human emotion and its triggers. Figure 1.2 shows the difference between macro and micro expression when the emotion intensity at its peak level. The study of micro-expressions and their role in nonverbal communication has been explored by a range of disciplines, including psychology, sociology, neuroscience, and computer vision. This interdisciplinary approach has led to a heightened awareness and sensitivity to these subtle facial behaviors, enabling us to better understand comprehensive human communication beyond spoken language (Xie et al., 2022).

The process of emotional expression in the face can be categorized into three consecutive temporal phases: onset, apex and offset. Onset is the starting phase of emotional expression, where the first hints of an emotion begin to surface. Next, the emotional expression escalates to the apex phase. Here, the emotion is fully visible, reaching its peak intensity. This is the stage where the emotion is most pronounced and easily identifiable. The final phase is offset, characterized by a gradual relaxation of the facial muscles post-apex. In this stage, the intensity of the emotional expression slowly diminishes, signaling the end of the emotional display (Wu, Lin, and Wei, 2014).

1.1. Motivation

Facial expressions are a fundamental aspect of human communication and play a crucial role in conveying emotions and facilitating social interaction. Recognizing and





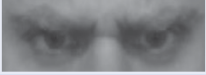
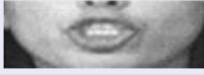



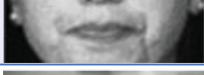






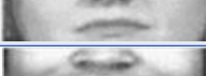





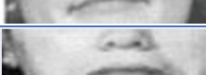



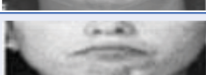

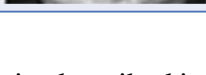

AU	Description	Facial Muscles (Type of Activation)		AU	Description	Facial Muscles (Type of Activation)	
1	Inner Brow Raiser	<i>Frontalis (pars medialis)</i>		18	Lip Pucker	<i>Incisivii labii superioris and incisivii labii inferioris</i>	
2	Outer Brow Raiser	<i>Frontalis (pars lateralis)</i>		20	Lip Stretcher	<i>Risorius with platysma</i>	
4	Brow Lowerer	<i>Corrugator supercilii, depressor supercilii</i>		22	Lip Funneler	<i>Orbicularis oris</i>	
5	Upper-Lid Raiser	<i>Levator palpebrae superioris</i>		23	Lip Tightener	<i>Orbicularis oris</i>	
6	Cheek Raiser	<i>Orbicularis oculi (pars orbitalis)</i>		24	Lip Pressor	<i>Orbicularis oris</i>	
7	Lid Tightener	<i>Orbicularis oculi (pars palpebralis)</i>		25	Lips Part	<i>Depressor labii inferioris or relaxation of mentalis, or orbicularis oris</i>	
9	Nose Wrinkle	<i>Levator labii superioris alaeque nasi</i>		26	Jaw Drop	<i>Masseter, relaxed temporalis and internal pterygoid</i>	
10	Upper-Lip Raiser	<i>Levator labii superioris</i>		27	Mouth Stretch	<i>Pterygoids, digastric</i>	
11	Nasolabial Deepener	<i>Zygomaticus minor</i>		28	Lip Suck	<i>Orbicularis oris</i>	
12	Lip-Corner Puller	<i>Zygomaticus major</i>		41	Lid Droop	<i>Levator palpebrae superioris</i>	
13	Cheeks Puffer	<i>Levator anguli oris</i>		42	Slit	<i>Orbicularis oculi</i>	
14	Dimpler	<i>Buccinator</i>		43	Eyes Closed	<i>Orbicularis oculi</i>	
15	Lip-Corner depressor	<i>Depressor anguli oris</i>		44	Squint	<i>Orbicularis oculi</i>	
16	Lower-Lip depressor	<i>Depressor labii inferioris</i>		45	Blink	<i>Orbicularis oculi</i>	
17	Chin Raiser	<i>Mentalis</i>		46	Wink	<i>Orbicularis oculi</i>	

Figure 1.1: List of action units described in Facial Action Coding System (FACS) (Source: Barrett et al., 2019)

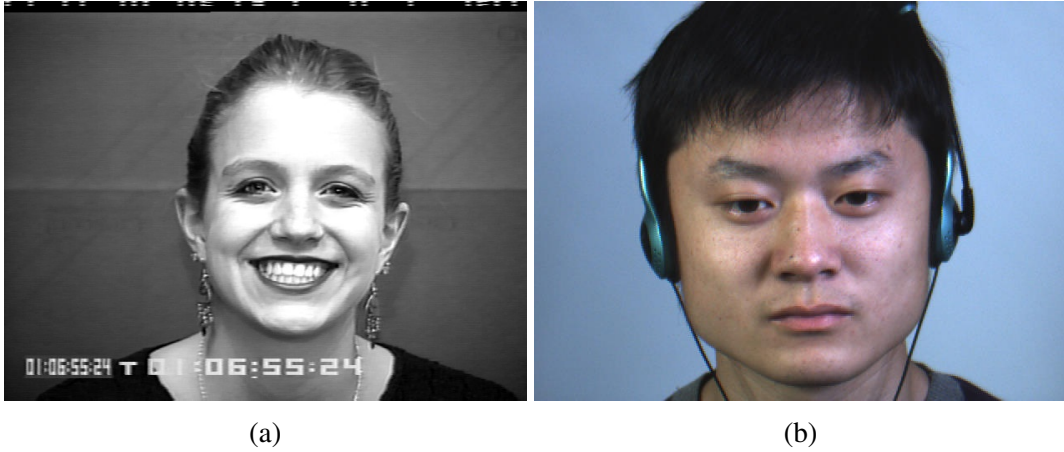


Figure 1.2: Comparison of macro and micro expressions displaying happiness emotion at their peak intensity. (a) Macro expression example taken from CK+ dataset (Source: Lucey et al., 2010). (b) Micro expression example taken from CASME II dataset (Source: Yan et al., 2014).

interpreting these expressions accurately is essential for a wide range of applications, such as affective computing, human-robot interaction, and emotion recognition (Kumari, Rajesh, and Pooja, 2015). However, accurate emotion recognition from facial expressions is challenging due to various factors, such as complex pre-processing of images, variations in illumination (Zhao et al., 2011), facial accessories, background noise, and differences in skin color (Adyapady and Annappa, 2023). Additionally, most existing methods lack of real-time processing speed. These challenges can significantly impact the accuracy and reliability of facial expression recognition systems and limit their practical applicability. Therefore, there is a need for a more robust and efficient approach to facial expression recognition that can handle these challenges and predict emotions accurately in real-world scenarios.

1.1.1. Contributions

This thesis proposes a novel approach to facial expression recognition and emotion classification using geometric features extracted from facial landmarks. Our method utilizes a faster and more efficient facial landmark extraction algorithm, resulting in a significant reduction in pre-processing time. We introduce two separate studies focusing on predicting emotions from macro and micro expressions. In the macro expression study, we extract geometric features that capture changes of movement of facial muscles from neutral to most intense emotion state(apex), enabling more accurate emotion recognition. In the micro expression study, we extract subtle changes in facial muscle movements

and use them to predict emotions with high precision. We evaluate our approach on several benchmark datasets and show that it performs competitive accuracies compared to state-of-the-art methods in terms of recognition accuracy and out-performs in terms of processing speed.

CHAPTER 2

REAL-TIME SEQUENTIAL MACRO EXPRESSION RECOGNITION USING GEOMETRIC FEATURES

2.1. Introduction

In this chapter, we present real-time sequential macro expression recognition method using geometric features extracted from facial landmarks. We begin with a literature review of existing methods for macro facial expression recognition, with a particular focus on geometric features. We then describe our methodology for real-time sequential macro expression recognition, which includes facial landmark detection, feature extraction, and classification using machine learning techniques. We present experimental results and provide a comprehensive discussion regarding the performance of our approach in terms of recognition accuracy, processing speed, and robustness in handling variations in facial expression intensity. Finally, we conclude with a summary of our contributions and future directions for research in this area.

2.2. Literature Review

In this study, we employed a landmark-based facial feature extraction approach, deviating from the more commonly utilized appearance-based (pixel centric) methodology prevalent in the literature. Here, several prominent landmark-based studies that have contributed significantly to facial expression recognition are highlighted. Choi et al. adopted sequential approach for representing facial features, employed facial landmarks to calculate the distances between all points and deriving their differences for consecutive frames, produced construct termed as Landmark Feature Maps (LFM). These LFMs were normalized to a range 0-255, generating LFM images. The facial features for each LFM were then extracted using a VGG13-based Convolutional Neural Network (CNN). The final layer incorporated Long Short-Term Memory (LSTM) and Multilayer Perceptron (MLP) for the classification (Choi and Song, 2020). Building on the LFM methodology, Kim et al. proposed “squeezed LFM” designed to eliminate redundant duplicate data within the LFM. They noted the inherent symmetry of LFMs about a diagonal axis,

given that the distance from point x to point y mirrors that from point y to point x (Kim et al., 2021). Apart from comparing distances between two points, alternate distance feature approaches have been proposed. For instance, Raj et al. proposed identifying a central point by calculating the mean of both axes, followed by determining the distance of all points relative to this central point (Rohith Raj et al., 2020). Meanwhile, Alvarez et al. fed a multilayer perceptron with two inputs: the first being the facial landmark coordinates of a person showing an emotion, and the second being the distance between this coordinates and the neutral state landmark coordinates of the same individual (Álvarez et al., 2018). Similarly, Sharma et al. proposed raw landmark coordinates and certain Euclidian distances that are manually picked as features (Sharma, Singh, and Gautam, 2019). Qui and Wan proposed to create an input vector by subtracting all landmark points relative to their regional center points. They divided the face into four regions, each with its own center point (Qiu and Wan, 2019). Beh et al. proposed six Euclidian distances selected manually over eyebrow, eye and mouth regions of face. The ratios of these distances to a reference distance were selected as features (Beh and Goh, 2019). Khan et al. proposed all Euclidian distances of each pair of extracted landmarks and additionally Euclidian distance of all landmarks relative to average point on the face to be used as features (Khan, 2018). Buhari et al. proposed both Euclidian distances and slopes of landmark pairs to be used as features. Facial regions were created based on Facial Action Coding System proposed by Paul Ekman. Features extracted from full face landmarks and landmarks belongs to created regions were experimented separately (Buhari et al., 2020).

In the literature mostly Dlib library's pre-trained facial landmark detector is used to extract facial landmarks (Álvarez et al., 2018; Beh and Goh, 2019; Buhari et al., 2020; Rohith Raj et al., 2020). There are also several other algorithms like incremental Parallel Cascade of Linear Regression (iPar-CLR) (Sharma, Singh, and Gautam, 2019) and landmark detector of IntraFace software package (Khan, 2018). In the study by Choi et al, facial landmark detection method which is called Supervision-by-Registration (SBR) is used (Choi and Song, 2020). Also, some datasets like the Extended Cohn-Kanade Dataset (CK+) comes with ready to use landmarks along with it which is tracked by an Active Appearance Model algorithm (Lucey et al., 2010). In preprocessing perspective, pixel centric approaches often require preprocessing before feeding image into neural network. These can be face alignment, scaling, rotation, illumination and color fixes, background and noise removal (Li and Deng, 2020). Landmark based approaches do not require additional preprocessing as long as landmark detection algorithm can detect required landmarks since all preprocessing is handled by the algorithm itself.

In this study MediaPipe Face Mesh is used to detect facial landmarks since it has an impressive performance on GPUs and can deliver 478 landmarks in total that is higher than

other landmark detection methods (Bazarevsky et al., 2019). To compare performance of MediaPipe Face Mesh and another popular package dlib, average processing time of both algorithms on a subject is measured. For each frame in the image sequence of subject, Euclidean distance and angle features are calculated using extracted facial landmark positions for selected landmark pairs. At neutral state calculated features are taken as basis and the features from rest of the sequence are subtracted from basis features. This approach is beneficial to reduce emotion intensity differences from person to person and calibrating automatically to the neutral state of the person. We did not introduce any preprocessing before using images from datasets, only raw images are used. Also, no data augmentation is applied to increase subject count.

2.3. Methodology

2.3.1. Computational Setup

Our experiments are conducted in the environments with following specs: Ubuntu OS, Intel i5-12600K CPU, 64 GB RAM, NVIDIA GeForce RTX 3060 12 GB GPU. Training of the proposed framework and the preprocessing operations have functioned using the Python programming language. We implemented our model in the Keras backend of the TensorFlow 2.1 framework. The categorical cross-entropy loss function and the Adam optimizer with default settings are used during the training. The batch size and number of epochs are selected as 32 and 200.

2.3.2. Datasets

Given the focus of this study on the spatio-temporal features, only sequential datasets are employed for the analysis. Figure 2.1 shows example images of subjects from datasets and table 2.1 gives summary of used datasets.

The CK+ is a fully FACS-compatible dataset with 593 sequences captured from 123 subjects at 30 frames per second (FPS) with either 640x490 or 640x480 pixels resolution. With the aid of two precisely synchronized Panasonic AG-7500 cameras, the facial expressions of 210 individuals were meticulously captured and analyzed. The participants, who ranged in age from 18 to 50, were predominantly female (69%) and of Euro-American descent (81%), with Afro-American (13%) and other groups (6%) making

Table 2.1: Summary of datasets used in macro expression experiments

Dataset	Subject Count	Sequence Count	FPS	Resolution	Emotion Count	Ethnicity
CK+	123	593	30	640x480	7	Euro-American (81%) Afro-American (13%) Other groups (6%)
Oulu-CASIA	80	2472	25	320x240	6	Finnish (~60%) Chinese (~40%)
MMI	19	848	24	720x576	6	European Asian South American

up the remainder. Under the guidance of an experimenter, the subjects were directed to execute a series of 23 facial expressions, which encompassed both individual action units and various combinations thereof. Each sequence starts from a neutral state, ends at the apex phase, and captures duration; hence frame count is different for each sequence. The apex frame of each sequence is validated and labeled by emotion researchers with reference to FACS Investigators Guide. The dataset consists of seven emotions, namely anger, contempt, disgust, fear, happiness, sadness, and surprise (Kanade, Cohn, and Tian, 2000; Lucey et al., 2010). The Oulu-CASIA is a sequential dataset with two variations: one is captured with visible light conditions (VIS), and the other one is captured with near-infrared conditions (NIR). A total of 80 people between 23 and 58 years old participated, and all expressions were captured at 25 FPS with an image resolution of 320×240 pixels. The database is comprised of two distinct parts. The first segment was captured in February 2008 by the Machine Vision Group of the University of Oulu in Finland, featuring a total of 50 subjects, a majority of whom were Finnish individuals. The second portion was recorded in April 2009 in Beijing by the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, and consisted of 30 subjects who were all of Chinese descent. Participants were instructed to take a seat in front of the camera in an observation room, with a distance of approximately 60 cm between their face and the camera. They were then prompted to imitate a facial expression as demonstrated in a series of pictures. Images with three different illumination conditions: weak, normal, and dark, are present in the dataset. Each sequence starts from a neutral state and ends at the apex phase (Zhao et al., 2011). The MMI Face Database is a complex source that contains both static and sequential images captured at frontal and profile views of faces. Every video sequence in the database was captured at a standard rate of 24 frames per second using a PAL camera. The collection comprises roughly 30 profile-view and 750 dual-view facial expression video sequences. These sequences differ in length, ranging from 40 to 520 frames, and

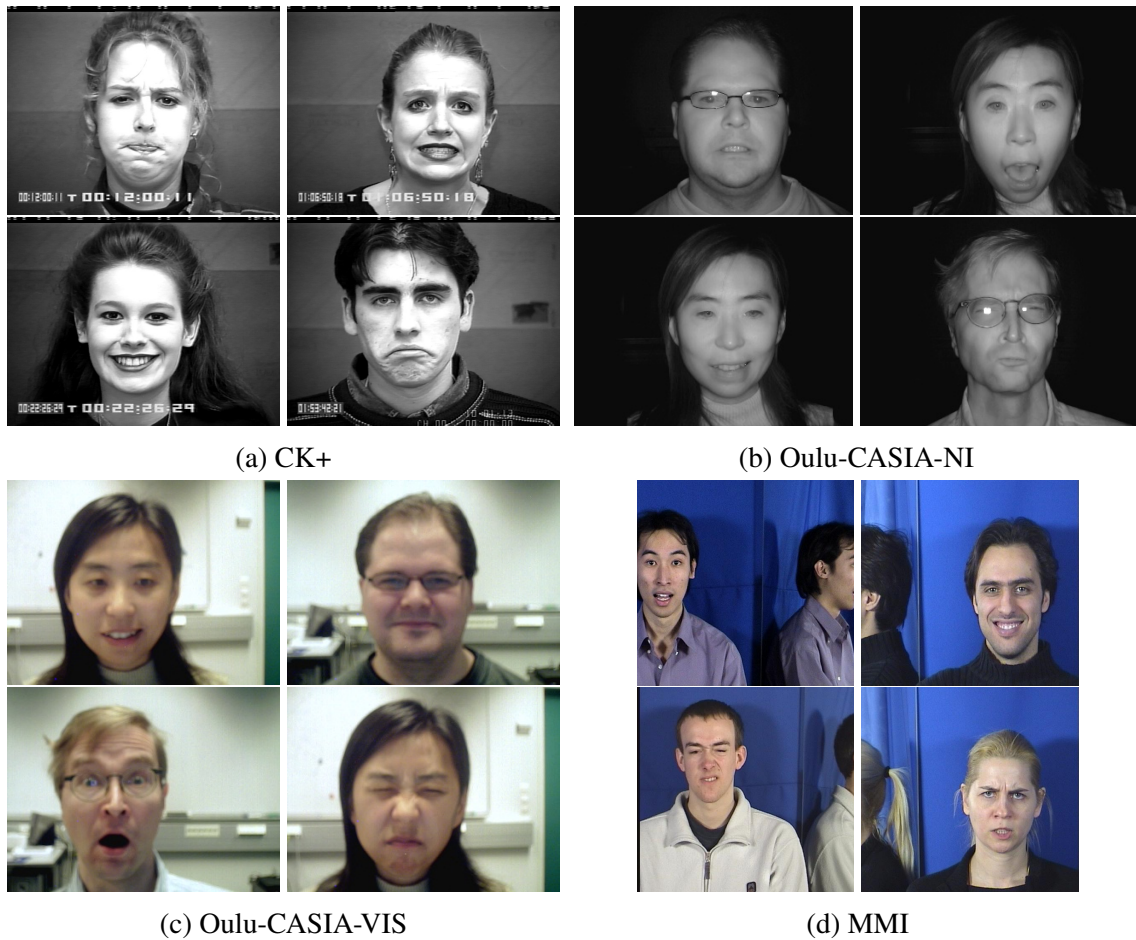


Figure 2.1: Collage of images in datasets which are used in macro-expression experiments. (a) Cohn-Kanade (CK+) dataset (Source: Lucey et al., 2010). (b) Oulu-CASIA dataset Near Infrared (NI) variation (Source: Zhao et al., 2011). (c) Oulu-CASIA dataset Visible Light (VIS) variation (Source: Zhao et al., 2011). (d) MMI dataset, only subjects that have sequential data is used (Source: Pantic et al., 2005).

portray one or multiple facial behavior patterns, starting with a neutral facial expression, followed by an expressive one, and ending with another neutral expression. The database features 19 distinct faces, belonging to both male and female students and research staff members, with an ethnic background of either European, Asian, or South American. The total number of female faces is 4400. The ages of the participants range from 19 to 62 years old. They were directed by a FACS coder on how to execute 79 different series of expressions and were asked to include a brief neutral state at the beginning and end of each expression. Onset apex and offset phases can be studied for this database (Pantic et al., 2005; Valstar, Pantic, et al., 2010).

2.3.3. Feature Creation Algorithm

Facial landmarks that we use in this paper provide a basis for deriving geometric features. These landmarks should be accurately positioned, and detection should be fast enough to achieve real-time performance for facial expression recognition tasks. For these reasons, MediaPipe FaceMesh solution is used as a facial landmark detector. MediaPipe FaceMesh is a facial landmark detection solution developed by Google’s MediaPipe team. It uses machine learning to identify and track 478 facial landmarks on a person’s face, including the eyes, eyebrows, nose, mouth, and jawline (Kartynnik et al., 2019). MediaPipe utilizes a lightweight and very fast, 200-1000 FPS on mobile GPUs, face detector, which is called BlazeFace (Bazarevsky et al., 2019). The face landmark model is a neural network-based model that estimates 478 landmarks with 3D coordinates. It uses a single camera output frame as an input to the model. This model is lightweight and applicable for real-time tasks with 100-1000 FPS on mobile GPUs [25]. Attention mesh is an optional step that applies attention to the eye, iris, and lip regions. As a result, estimated landmarks are more accurate on these regions (Grishchenko et al., 2020).

The first step of the algorithm shown in Figure 2.2 is to calculate facial landmarks of each camera frame. Camera frames are sequentially fed into FaceMesh algorithm, and resulting landmark coordinates are stored to be processed by the feature creation algorithm in the second step. Facial landmarks of the current frame and neutral frame are input to the feature creation step. The feature creation algorithm generates all features belonging to the current frame by calculating the Euclidean distance and angle of each landmark pair for the current frame and neutral frame. Equations 2.1 and 2.2 show the calculation of Euclidean distance and angle, respectively, for two landmark points, i and j . For each landmark pair, the calculated distance and angle values of the current frame are subtracted from the respective values of the neutral frame. Resulted values give the distance and angle features of that landmark pair.

$$d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (2.1)$$

$$\theta = \arctan\left(\frac{x_i - x_j}{y_i - y_j}\right) \quad (2.2)$$

By default, landmark pair count is calculated by finding number of two combinations $C(n, 2)$ for total number of facial landmarks n . This count can be reduced by grouping facial landmarks and calculating two combinations inside each group and com-

binning them. Grouping landmarks ensures that there will be no landmark pair that has two landmark points belongs to two different groups.

$$C(n, 2) > \text{Unique}(C(a, 2) + C(b, 2) + C(c, 2) + C(d, 2) + C(e, 2))$$

where:

n = total landmark count

a, b, c, d, e = landmark counts for 5 different groups




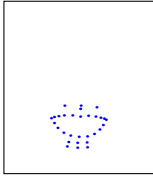

$a, b, c, d, e < n$

In this study, landmark grouping based on facial action coding system is implemented. This grouping is similar to the one presented in the paper by Buhari et. al. (Buhari et al., 2020). Table 2.3 and 2.2 show the implemented categories and respective action units (AU) that are represented by muscles residing at shown landmark positions. Action units 1,2,3,4 and 5 are activated by facial muscles in eye and eyebrow regions so category 1 is created with landmark points on that regions. For action unit 6 eye and nose regions are selected as category 2. For action units 7 and 9, category 3 is created that consists eye, eyebrow and nose landmarks. For action units 12,14,15,16,23 and 26, category 4 is created that consists nose, mouth and lower jaw landmarks. Lastly action unit 20 consists landmarks present in eye nose and mouth regions and category 5 is created.

Table 2.2: Relation between emotion, action units and category mapping. Correlation between action units and resulted emotion is proposed by (Source: Friesen, Ekman, et al., 1983)

Emotion	Action Units	Required Categories
Anger	4, 5, 7, 23	cat 1, 3, 4
Contempt	12, 14	cat 4
Disgust	9, 15, 16	cat 3, 4
Fear	1, 2, 4, 5, 7, 20, 26	cat 1, 3, 4, 5
Happiness	6, 12	cat 2, 4
Sadness	1, 4, 15	cat 1, 4
Surprise	1, 2, 5, 26	cat 1, 4

Table 2.3: Landmark grouping categories (Source: Buhari et al., 2020). Dividing facial landmarks into meaningful categories reduces total landmark pair counts that will be processed by feature creation algorithm.

Category	Landmarks	Region	Action Units
cat 1		Left Eye Left Eyebrow Right Eye Right Eyebrow	1, 2, 3, 4, 5
cat 2		Left Eye Right Eye Nose	6
cat 3		Left Eye Left Eyebrow Right Eye Right Eyebrow Nose	7, 9
cat 4		Nose Mouth Lower Jaw	12, 14, 15, 16, 23, 26
cat 5		Left Eye Right Eye Nose Mouth	20

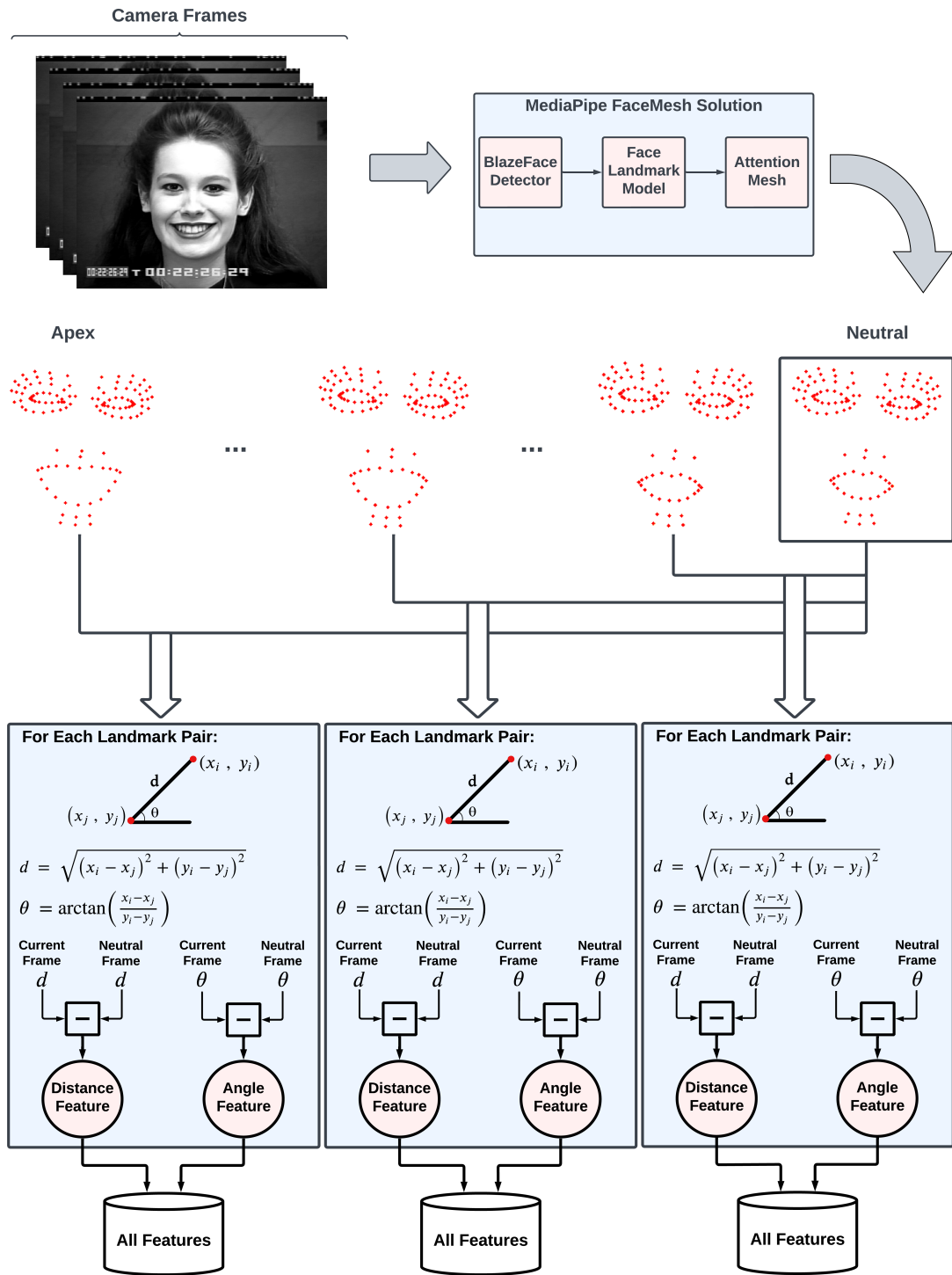


Figure 2.2: Feature creation algorithm flow for macro-expression experiments. Sequential camera frames are fed into MediaPipe FaceMesh Solution get extract facial landmark positions. For each selected landmark pair, Euclidean distance and angle θ features are calculated. Neutral frame features are taken as basis and at each frame after neutral frame, calculated features are subtracted with basis features. These two features are concatenated to create feature vector. Images of the subject are taken from CK+ dataset (Source: Lucey et al., 2010)

2.3.4. Classification Algorithm

In a study by Alvarez et al., it is concluded that for facial emotion recognition tasks, Multilayer Perceptron is the classifier that achieves the highest accuracy compared to SVM, Naïve Bayes, Decision Tree, Random Forest, and AdaBoost (Álvarez et al., 2018).

In this study, Multilayer Perceptron is selected as a classifier. Figure 2.3 shows the classification part of the proposed method after features are created.

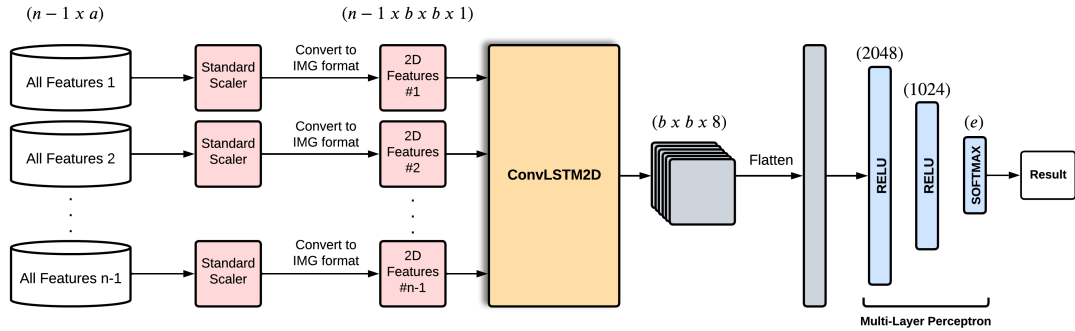


Figure 2.3: Classification algorithm for macro-expression experiments. First $N - 1$, A shaped array that holds feature vectors for whole sequence is scaled using standard scaler. Scaled 1D data is converted to image format that will be feed into ConvLSTM2D block. Output of ConvLSTM2D block is flattened and data is classified using multi-layer perceptron layers. Where, n : frame count, e : emotion count, a : feature count, b : $\text{ceil}(\sqrt{a})$

To extract information from temporal domain ConvLSTM is used. ConvLSTM is a type of neural network architecture that combines convolutional layers with LSTM (Long Short-Term Memory) layers. It is commonly used for sequence prediction tasks, such as video and image sequence processing, where both spatial and temporal dependencies need to be modeled.

In ConvLSTM, the input data is processed by convolutional layers to capture spatial features, and then the output of the convolutional layers is passed to LSTM layers, which capture temporal dependencies. The LSTM layers maintain an internal state that enables them to capture long-term dependencies in the sequence (Shi et al., 2015).

After features are created and stored in $N - 1$, A shaped array where N is the frame count and A is the total feature count, data should be scaled before classification. Standard scaler is used for this purpose.

Standard scaling is a preprocessing step in machine learning that scales the data so that each feature has zero mean and unit variance. This is important because many

algorithms assume that the data is normally distributed with zero mean and unit variance. Standard scaling can be applied to both training and test data and is particularly useful when dealing with features with different scales or units (Raju et al., 2020). It is implemented in many popular machine learning libraries, such as Scikit-learn in Python (Pedregosa et al., 2011).

$$z = \frac{x - u}{s} \quad (2.3)$$

where:

u = mean of the training samples

s = standard deviation of the training samples

Scaled 1D features are converted to image format which makes them 2D and an extra channel dimension is added to hold color information. Resulted scaled and converted feature vector is fed into ConvLSTM2D block with a kernel size 1, 1 and filter size 8. Output of the ConvLSTM2D is flattened and dense layers of 2048 and 1024 neurons respectively are used in multi-layer perceptron. Final classification layer has neuron count which is equal to emotion count to be classified and softmax activation function is used.

In the process of our study, we incorporated the use of Convolutional Long Short-Term Memory (ConvLSTM2D) networks, a variant of the traditional LSTM networks that are specially designed to handle spatiotemporal data. LSTM networks are a special kind of Recurrent Neural Networks (RNN) that have feedback connections, allowing them to process sequences of data (Hochreiter and Schmidhuber, 1997). They are capable of learning long-term dependencies, which makes them particularly effective for many sequential data tasks. The ConvLSTM is a type of LSTM that has convolutional structure in both the input-to-state and state-to-state transitions (Shi et al., 2015). This makes it uniquely suited to handle two-dimensional spatial data over time. Each unit of a ConvLSTM2D network maintains a cell state and multiple gating units, including an input gate, a forget gate, and an output gate, which control the flow of information into and out of the cell. The convolution operation is applied in the state transition and the gate activations, which allows the ConvLSTM2D to effectively capture the spatial dependencies in the data. The operations utilized within LSTM are reconfigured for ConvLSTM, as indicated in equations 2.4-2.8 below: in this context, the symbols " $*$ " and " \circ " correspond to the convolution operation and the Hadamard product, respectively. " x " stands for the input vector, which is the data that the network is receiving at a given time step. " h " stands for the hidden state vector, which represents the internal state of the network at

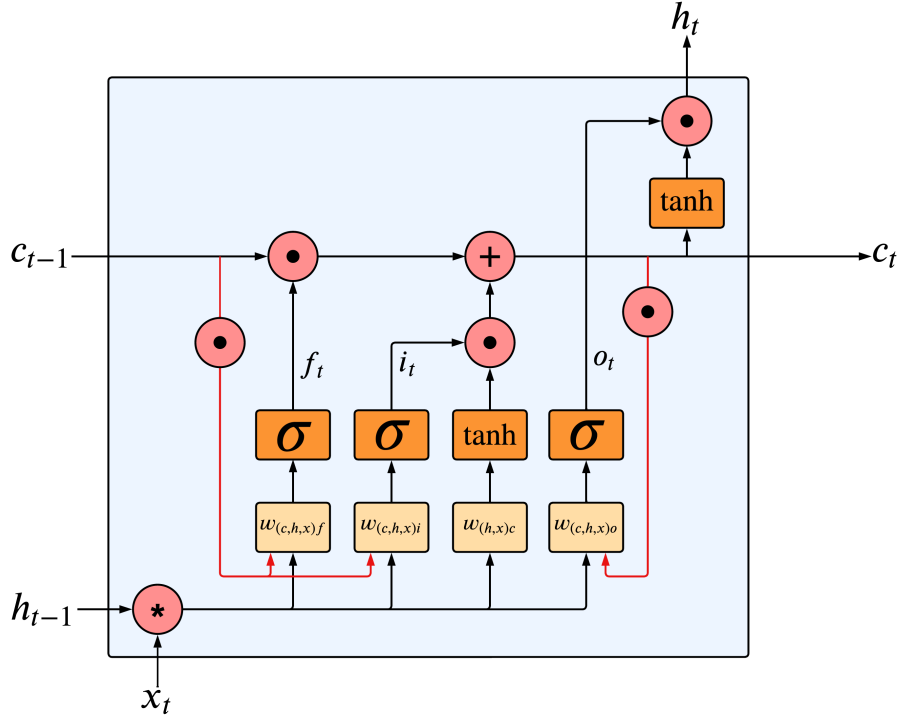


Figure 2.4: Inner structure of ConvLSTM cell

a given time step. It serves as the memory of the network, allowing it to keep track of relevant information from previous time steps and make predictions based on that information. "c" stands for cell state which allow the network to selectively remember or forget information based on its relevance to the current task. Figure 2.4 shows inner structure of ConvLSTM cell. The ConvLSTM2D network has proven to be effective in a wide range of applications, particularly those involving spatiotemporal data, such as video processing, weather forecasting, and traffic prediction (Di et al., 2019; Shi et al., 2015; Tariq, Lee, and Woo, 2020) The use of ConvLSTM2D in facial expression analysis allowed us to effectively capture the spatial and temporal dependencies in our data and provide valuable insights into the problem.

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \quad (2.4)$$

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \quad (2.5)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (2.6)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o) \quad (2.7)$$

$$h_t = o_t \circ \tanh(c_t) \quad (2.8)$$

2.4. Results and Discussion

In order to implement facial expression analysis methods in real-world scenarios, they need to satisfy several criteria. Among these, processing time emerges as a crucial consideration for real-time applications. The cumulative time required for preprocessing, feature creation, and classification stages should not surpass the interval between two consecutive frames captured by the camera. Moreover, models that incorporate the temporal dynamics of facial features can yield more robust models compared to those that rely solely on spatial features. Relying solely on a single static moment may lead to misinterpretations, particularly when a person's neutral state closely resembles a specific emotional expression. Facial expressions are dynamic in nature and continuously change over time. Thus, analyzing the entire sequence of expressions is more appropriate. This enables us to examine the onset, apex, and offset phases of emotion within the temporal domain, facilitating accurate detection and labeling of the entire sequence with the corresponding emotional tag. This approach is more precise and can provide valuable insights into the dynamics of facial expressions. In our study, we adopt a sequential approach to facial emotion recognition tasks, an advancement over traditional static methods. Our technique is designed to compare and differentiate all features of frames within an expression sequence from their respective neutral states. This methodology provides an automatic calibration to each subject's baseline, thereby enhancing the method's accuracy and reliability.

Apex is the most intense moment that an emotion can be observed in the face. In order to detect facial macro expression in real-time, determining the apex region is critical since the accuracy of prediction is higher than in other regions. In Figure 2.5 it can be observed that by tracking the mean value of distance features of frames, onset, apex, and offset regions can be detected. The distance feature is created by subtracting the Euclidean distance of landmark pairs for the current frame and neutral frame. Since coordinates of landmarks are expressed as pixels, the value shows pixel difference from the neutral state. In the figure, frame number 10 corresponds to the second image and can be considered as starting point of the apex phase. Frame number 40 corresponds to the fifth image and can be considered as the ending point of the apex phase.

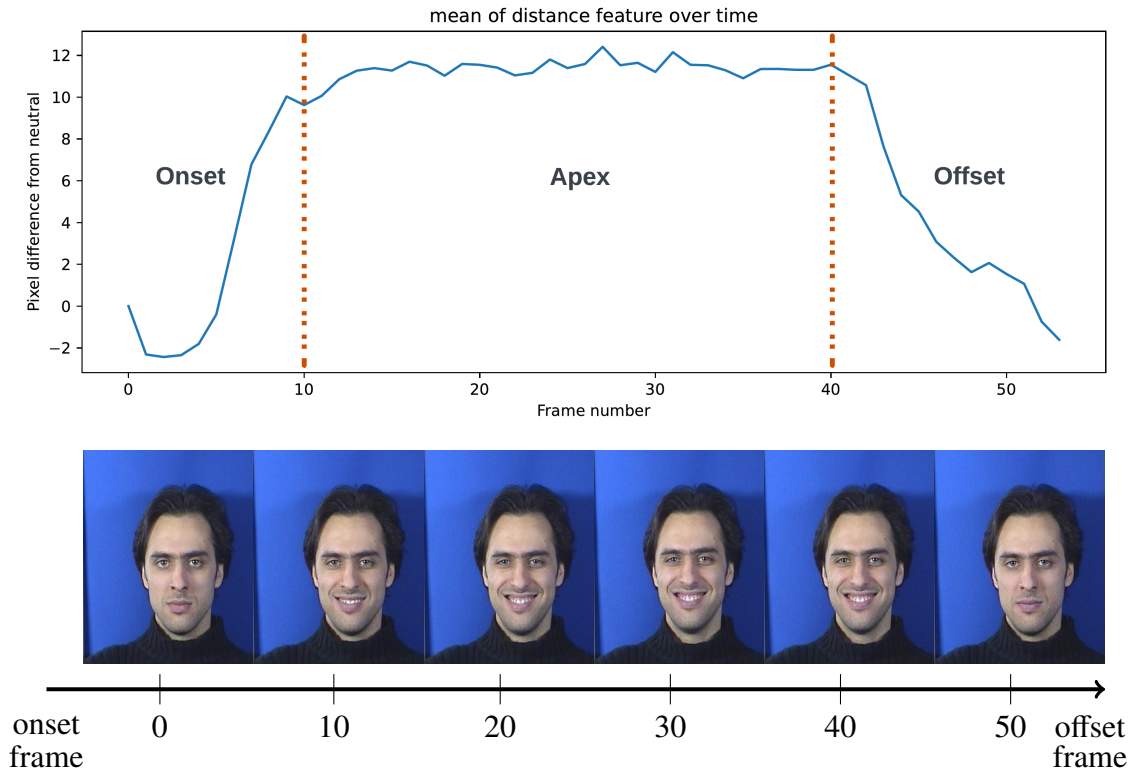


Figure 2.5: Tracking mean value of Euclidean distance features of frames. Y axis shows pixel difference from neutral state, x axis shows frame number starting from onset (frame number 0) till offset. Images below plot shows actual images that these features are extracted. Frames number in between 0 to 10 corresponds to onset phase, 10 to 40 corresponds to apex phase and after 40 offset phase starts. Images of the subject are taken from MMI dataset (Source: Pantic et al., 2005)

2.4.1. Experiments with Datasets Individually

In order to validate our model, 5-Fold cross validation is implemented to find average accuracies for CK+, Oulu-CASIA NIR & VIS and MMI datasets. Different experiments are conducted to expose correlation between accuracy and the number of facial landmarks, along with the number of created features. To find out the relation between accuracy and the number of facial landmarks, three different preset landmark counts are defined as 61, 122, and 250. Those landmarks seen in Figure 2.6 are selected manually from 478 landmarks of FaceMesh output. Landmarks are selected based on facial muscle locations on the face according to action units (AU). Main action units that are used to recognize emotions located on eye, eyebrow, mouth, nose, and chin regions on the face. To find out the relation between accuracy and the number of created features, a feature selection algorithm based on FACS by Buhari et al. (Buhari et al., 2020) is utilized

to reduce the number of generated features. So, in total six experiments are conducted for each dataset.

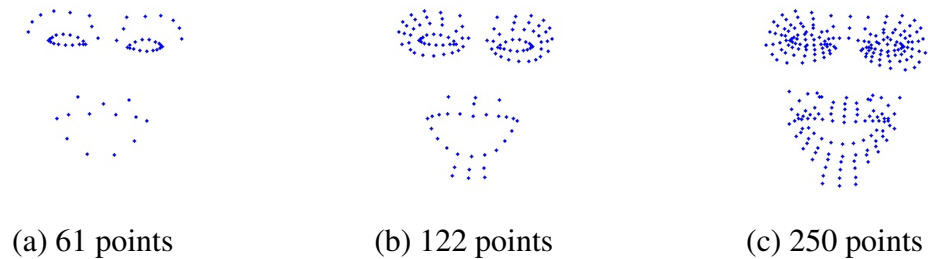


Figure 2.6: Selected facial landmark sets to be used in macro-expression experiments. 61, 122 and 250 landmark points are selected manually from 478 facial landmark.

In Figure 2.7, box plots of accuracies for every dataset are shown. For each plot, results for all six experiments, respectively 61, 122, and 250 points landmarks with AU grouping and without any grouping, are visible. The box plots display the mean accuracy values of all five folds, indicated by a red dashed line. The minimum and maximum values excluding outliers are represented by grey lines below and above the boxes, as well as grey dots if they are outliers. It can be deduced from conducted experiments that grouping facial landmarks based on FACS usually offers better results since it acts as a feature selection method that selects prominent features among all. Also, it can be stated that increasing landmark counts does not significantly increase the accuracies, and sometimes it even has negative effects like it is observed in the MMI dataset. The CK+ dataset achieved its highest mean accuracy of 93% in experiment 250 landmarks with AU grouping. For the MMI dataset, the highest mean accuracy of 68% was recorded in experiment 61 landmarks without grouping. In the case of the Oulu-CASIA VIS dataset, the results showed that experiment 250 landmarks without grouping had the highest mean accuracy of 79%. However, the Oulu-CASIA NIR experiment yielded slightly lower results compared to the VIS version. The experiment with 250 landmarks and AU grouping had the highest mean accuracy of 77%.

It's helpful to use confusion matrices to better understand how accurately emotions are predicted and which emotions are often confused with others. In Figure 2.8, it can be observed that for CK+ dataset, contempt is commonly confused with fear and sadness, while fear is often confused with surprise and sadness for the MMI and Oulu-CASIA NIR datasets. In contrast, sadness is the most difficult emotion to predict accurately for the Oulu-CASIA VIS dataset. Furthermore, happiness and surprise emotions tend to have the highest prediction accuracy across all datasets.

Table 2.4 presents the processing times for creating features in all six experiments.

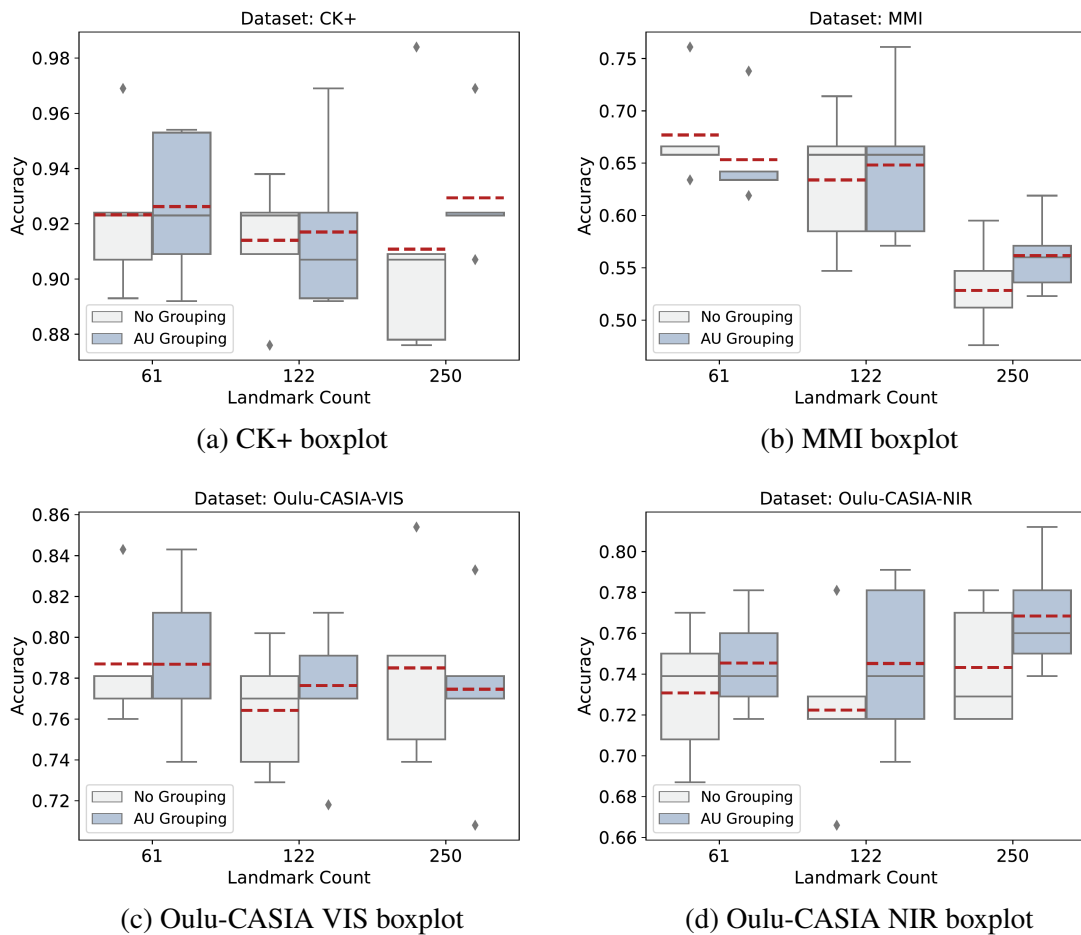


Figure 2.7: Accuracy box-plots of datasets used in macro-expression experiments. Results for all six experiments that are combinations of 61, 122, 250 point landmarks with AU grouping and without any grouping. Red dashed line shows mean value of accuracy for 5-fold cross validation.

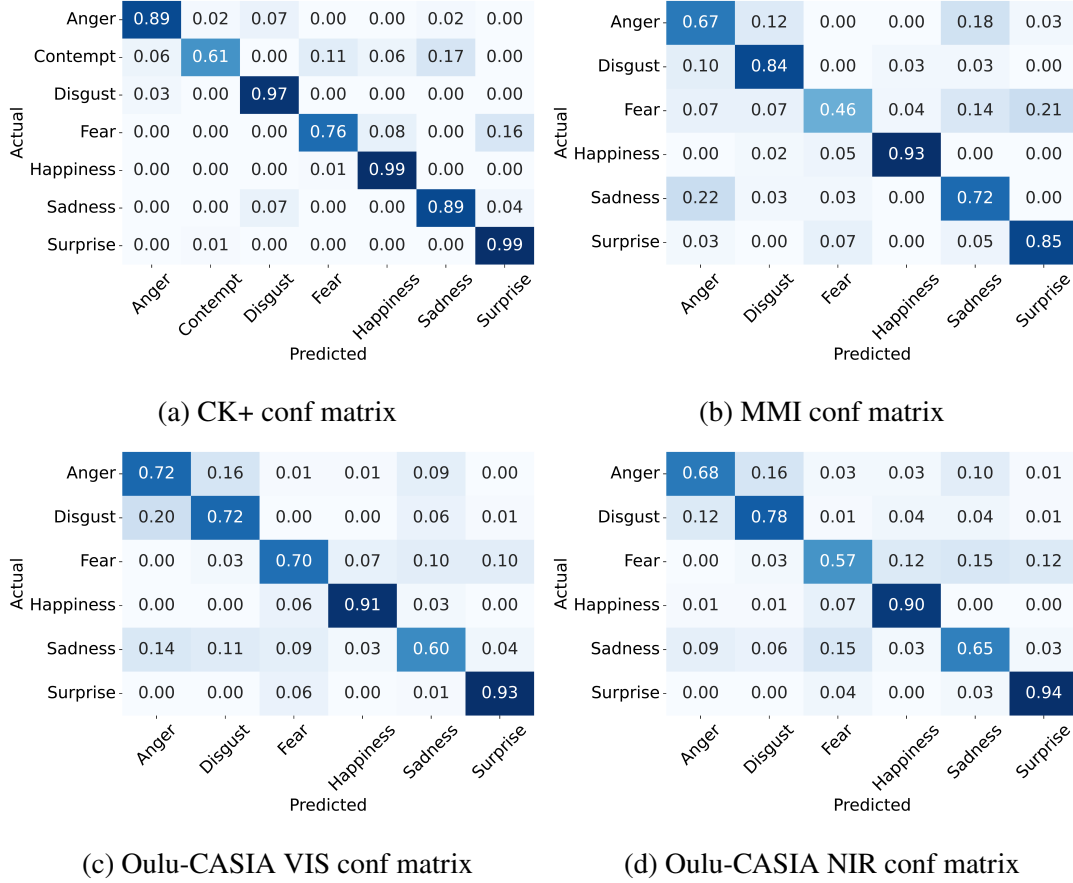


Figure 2.8: Confusion matrices of datasets used in macro-expression experiments. Results are created with average values of 5-fold cross validation.

It’s worth noting that using a GPU to create features is 12 times faster than using a CPU with our hardware setup. The fastest processing time recorded is 0.45ms, which was achieved by creating features for 61 landmarks with AU grouping.

We used MediaPipe’s FaceMesh solution to extract facial landmarks, which has superior performance on GPUs, making it suitable for real-time tasks. The time required to extract all landmarks with our hardware setup is measured as 5.6 ms, regardless of the landmark count used in the experiment. This is significantly faster than dlib’s landmark detection algorithm, which takes 100 ms to process. In the second phase, the time taken to create features depends on selected landmark point pairs. Table 2.4 provides the measured time to create features for different landmark point counts and categories. For 61 landmark points with AU grouping, the total processing time, from capturing camera frames to creating all features for each frame, is 6.05 ms on Nvidia RTX 3060 with 8.6 compute capability. This value indicates that the method can support approximately 165 fps video processing in real-time. This means that the system can analyze video frames at a high speed, allowing for efficient and timely recognition of facial expressions. Additionally, the method was evaluated using the Oulu-CASIA dataset, which consists of

Table 2.4: Comparison of processing times to create features for GPU and CPU execution. Full means using all landmark pairs and AU means using selected landmark pairs based on FACS based method.

	61 LM		122 LM		250 LM	
	Full	AU	Full	AU	Full	AU
GPU (ms)	0.65	0.45	2.60	1.65	11.01	8.08
CPU (ms)	7.74	5.17	31.13	21.53	130.67	91.20

frames captured under both visible light (VIS) and near-infrared light (NIR) conditions. This evaluation aimed to demonstrate the robustness of the method across different ambient light conditions without the need for additional preprocessing. By successfully classifying facial expressions under varying lighting conditions, the method proves its versatility and suitability for real-world applications.

Table 2.5: Comparison of accuracy of different methods in the literature which use only geometric features

Paper	Accuracy(%)		
	CK+	Oulu-CASIA	MMI
(Jung et al., 2015)	92.35	74.17	59.02
(Choi and Song, 2020)	92.60	-	-
(Qiu and Wan, 2019)	92.00	-	-
(Rohith Raj et al., 2020)	89.00	-	-
(Álvarez et al., 2018)	89.00	-	-
Proposed method	93.00	79.00	68.00

The accuracies of other geometric-based methods in the literature, which operate on the datasets used in this study, are presented in Table 2.5. While there are not many geometric-based methods that have been tested on the Oulu-CASIA and MMI datasets, our proposed method surpasses the performance of the mentioned methods in terms of recognition accuracy.

2.4.2. Composite Dataset Experiments

In this section, a single composite dataset is created by merging multiple datasets, and experiments are conducted using this composite dataset. The first experiment involves training the proposed method on three datasets and validating it on a fourth dataset that

is not included in the composite dataset. The second experiment combines all the macro datasets and performs internal validation using a 5-fold cross-validation approach. The third experiment utilizes the composite dataset trained in the first experiment and validates it on micro expression datasets.

In the first experiment, the CK+, Oulu-CASIA NI, and Oulu-CASIA VIS datasets are combined by considering only six basic emotions: anger, disgust, fear, happiness, sadness, and surprise, which are common to all datasets. Emotions such as contempt and others are removed to enable the merging of the datasets into a single composite dataset. The composite dataset consists of the following sequence counts for each emotion: anger (205), disgust (219), fear (185), happiness (229), sadness (188), and surprise (243). The resulting dataset is then trained using the proposed method and validated using the MMI dataset, which is not included in the composite dataset. The test accuracy of the composite dataset is achieved as 81.30% using 5-fold cross-validation. The validation accuracy on the MMI dataset is achieved as 69.70%, which is slightly higher than the value (68%) obtained during the individual training of the MMI dataset.

In the second experiment, all datasets, including CK+, Oulu-CASIA NI, Oulu-CASIA VIS, and MMI, are combined to create a composite dataset. This composite dataset comprises 476 sequences for anger, 500 sequences for disgust, 426 sequences for fear, 540 sequences for happiness, 440 sequences for sadness, and 566 sequences for surprise emotions. The model is validated using a 5-fold cross-validation approach, resulting in an accuracy of 82.10%.

In the third experiment, the model trained in first experiment is used to predict features created using micro expression dataset SAMM. It is observed that using original SAMM dataset without applying PBVM processing, 15.70% accuracy is achieved and for SAMM with PBVM applied this value raised to 23.96%.

Based on the composite dataset experiments, it can be concluded that our model learns informative features that are not only specific to the dataset but also generic, allowing its application to other macro expressions that were not included in the training. However, it is important to note that the model trained with macro expressions is not suitable for accurate predictions of micro expressions. Although there is an approximate 8% increase in accuracy when the micro expressions are magnified using PBVM processing, the model still falls short in accurately predicting micro expressions. Therefore, it is evident that there are significant differences between macro and micro expressions that require distinct modeling approaches.

2.4.3. Real-time Implementation

For the real-time implementation of the proposed method, a specific modification is introduced to simplify the classification algorithm and address potential frame count mismatches between the trained model and the video being predicted. In this modification, only the last value of the convlstm block is used, which carries the most informative feature representation of the entire sequence. The last value of the convlstm block corresponds to the difference between the euclidean distance and slope features extracted from the apex frame and their corresponding features from the neutral frame. By utilizing this single value, the complexity of the classification algorithm is significantly reduced. To prepare the feature for classification, it is scaled using standard scaler. Once the feature is scaled, it is directly fed into the multi-layer perceptron (MLP) block for further processing and classification. By employing this modification, the real-time implementation of the proposed method becomes more efficient and streamlined. It allows for a simplified classification algorithm that focuses on the most informative feature representation while mitigating potential frame count mismatches between the trained model and the video being predicted.

Model is trained with the composite dataset that is created for first experiment of section 2.4.1. Due to the modification explained in previous paragraph, only neutral and apex frames are used to create feature vector. Validation accuracy is calculated as 80% which is slightly less than full sequential method proposed in section 2.3.3. After the training, model and scaler are saved as an h5 file. Pretrained model is used to predict the emotion class of validation data while prefitted scaler is used to scale features of validation data.

To validate the real-time performance of the model, a specific video is selected that displays happiness macro expression and starts with a neutral state. This video is sourced from the MMI dataset and was not included in the training phase of the model. To conduct the validation, a test script is created. The script reads the video and considers the first frame as the neutral frame. The facial landmarks in each subsequent frame are extracted using MediaPipe Face Landmarker. From these landmarks, the Euclidean distances and slopes are calculated. These calculated distances and slopes are then subtracted from their corresponding neutral state values, which were stored in a variable beforehand, resulting in a feature vector for each frame. The feature vectors are first scaled using a pre-fitted scaler, ensuring they are in a suitable format for classification. The pretrained model is then used to predict the emotion class for each frame based on the scaled feature vector. The results of this real-time validation process can be observed in Figure 2.9.



Figure 2.9: Real time prediction results. First frame is taken in between neutral and onset phases. Second frame shows onset phase where facial expression is started. Third frame shows apex phase where emotion is predicted with 99.97% accuracy. Image sequence is taken from MMI dataset (Source: Pantic et al., 2005)

The real-time prediction results show the progression of emotion throughout the frames of the video. In the first frame, which is captured between the neutral and onset phases, the facial expression is not yet fully developed. In the second frame, the onset phase is observed, indicating the beginning of the facial expression. The model detects the changes in facial landmarks and starts to recognize the emerging emotion. In the third frame, the apex phase is reached, where the emotion is fully expressed. The model predicts the emotion with a high accuracy of 99.97%. This indicates that the model successfully captures the features and patterns specific to the expressed emotion, leading to a very accurate prediction.

2.5. Conclusion

In this study, we proposed a deep learning based sequential macro-expression recognition method by detecting facial landmarks using MediaPipe's FaceMesh solution which is significantly faster than the popular dlib facial landmark detection algorithm. While creating geometric features from facial landmarks we considered the difference of Euclidean distance and angle features with respect to neutral state of subjects. Unlike emotion recognition using a static snapshot of a subject, this approach provides auto calibration to the baseline of subjects which is changing from person to person. Also, it is shown that by tracking mean value of difference of distance features over time, onset, apex and offset phases of an emotion can be detected. In our experiments we observed that increasing landmark count does not necessarily improve accuracy and sometimes it can have negative effects. Experiments with FACS based landmark grouping method show that selecting useful features using a feature reduction algorithm often increases classification accuracy. With the proposed method we achieved competitive mean accuracy values among the landmark based methods in the literature using 5-fold cross validation technique. We tested the proposed method with CK+, Oulu-CASIA VIS & NIR and MMI datasets and achieved following accuracy results respectively; 93%, 79%, 77%, 68%. Composite dataset experiments involved merging multiple datasets to observe the generalization of the proposed model. The real-time implementation of the model was validated using a video displaying happiness emotion, achieving a remarkable prediction accuracy of 99.97%. These results demonstrate the model's ability to generalize across datasets and accurately predict emotions in real-time scenarios.

Despite the advancements, achieving a robust, accurate, and real-time solution for facial expression recognition remains an ongoing challenge. Given the escalating prevalence of human-computer interaction, it is foreseeable that many applications across

diverse areas will require capabilities to detect human emotions. To be viable for real-world adoption, the proposed system must not only be fast but also resilient against various challenges, including changes in illumination, face rotation, facial accessories, and other potential distortion factors. Our study paves the way for continued exploration in this field, contributing to the pursuit of a versatile, precise, and user-adaptive solution for facial expression recognition.

CHAPTER 3

SEQUENTIAL MICRO EXPRESSION RECOGNITION USING GEOMETRIC FEATURES

3.1. Introduction

In this chapter, we present sequential micro expression recognition method using geometric features extracted from facial landmarks. We begin with a literature review of existing methods for micro facial expression recognition, with a particular focus on geometric features. We then describe our methodology for sequential micro expression recognition, which includes facial landmark detection, feature extraction, and classification using machine learning techniques. We present experimental results and a detailed discussion of our approach's performance in terms of recognition accuracy, speed. Finally, we conclude with a summary of our contributions and future directions for research in this area.

3.2. Literature Review

The table 3.1 provides a comparison of facial micro expression recognition models experimented on two datasets, SAMM and CASME II. The table includes methods from five different papers and our proposed method, each with its own landmark detection method and feature type. The accuracy of each model is reported for both datasets if available.

In the first paper by Choi et al. (Choi and Song, 2020), the landmark detection method proposed by Dong et al. (Dong et al., 2018) was employed to create landmark points and Euclidean Distance was used as feature type. This study adopted a sequential approach to represent facial features, using facial landmarks to calculate the distances between all points and derive their differences for consecutive frames, producing construct termed as Landmark Feature Maps (LFM). These LFMs were normalized to a range of 0-255, resulting in LFM images. A VGG13-based Convolutional Neural Network (CNN) was then used to extract facial features for each LFM, with the final layer incorporating Long Short-Term Memory (LSTM) and Multilayer Perceptron (MLP) for classification

Table 3.1: Comparison of accuracy of different methods in the literature for micro-expression recognition which use only geometric features

Paper	Landmark Detection Method	Feature Type	Accuracy(%)	
			SAMM	CASME II
(Choi and Song, 2020)	(Dong et al., 2018)	Distance	-	73.98
(Buhari et al., 2020)	Dlib	Distance & Slope	87.33	75.04
(Beh and Goh, 2019)	Dlib	Ratio of Distances	-	82.00
(Buhari et al., 2022)	Dlib	Landmark Based Facial Graph	94.72	94.78
(Xia et al., 2019)	(Xia et al., 2016)	STRCN-G	78.60	80.30
Proposed method	MediaPipe Face Landmarker	Distance	84.21	92.94

(Choi and Song, 2020). The accuracy for SAMM is not reported, while the accuracy for CASME II is 73.98%. In the second paper by Buhari et al. (Buhari et al., 2020) the Dlib landmark detection method was used to create Euclidean distance and slope features of static image. For a static image, features extracted from full face landmarks and landmarks belongs to regions which are created based on Facial Action Coding System proposed by Paul Ekman, were experimented separately. Features were normalized and classified using SVM classifier (Buhari et al., 2020). The accuracy achieved was 87.33% for SAMM and 75.04% for CASME II. Beh et al. (Beh and Goh, 2019) used Dlib landmark detection method and Ratio of Distances as the feature type. First, face alignment was applied to ensure that the eyes were horizontally aligned, the size of the detected face was consistent across frames and the position of the face was centered in the video frame. Then 12 landmark pairs were selected manually and ratios of euclidean distances for selected landmark pairs were calculated. Also in order to reduce effect of jittering, thresholding was applied to the calculated features, only features above neutral state plus threshold value were used. Although the accuracy for SAMM is not reported, the model achieved 82% accuracy on the CASME II dataset. In the fourth paper by Buhari et al. (Buhari et al., 2022), the Dlib landmark detection method was used with Landmark Based Facial Graph as the feature type. First, a magnification vector consisting of integer values that represent the pixel point changes between the onset-frame and the apex-frame was created. This vector was then amplified using an integer scalar value. Finally, the magnified pixel change vector was added to neutral frame landmark position to obtain new apex frame landmarks with exaggerated landmark positions. Using these enhanced facial landmark positions, facial graph features were created to be trained with SVM. This model achieved the highest accuracy among all models on both datasets, with 94.72% accuracy for SAMM

and 94.78% accuracy for CASME II. The fifth paper by Xia et al. (Xia et al., 2019) used ASM landmark detection method proposed by Xia et al. (Xia et al., 2016) and STRCN-G as the feature type. The ASM face model was employed to portray the facial geometry shape, and the Procrustes analysis technique was applied to align the points of this shape. Once the alignment was completed, small changes in geometric shape were determined and these alterations in facial geometry were utilized as features for the Adaboost model training (Xia et al., 2019). The accuracy achieved was 78.60% for SAMM and 80.30% for CASME II.

3.3. Methodology

3.3.1. Computational Setup

Our experiments are conducted in the environments with following specs: Ubuntu OS, Intel i5-12600K CPU, 64 GB RAM, NVIDIA GeForce RTX 3060 12 GB GPU. Training of the proposed framework and the preprocessing operations have functioned using the Python programming language. We implemented our model in the Keras backend of the TensorFlow 2.1 framework. The categorical cross-entropy loss function and the Adam optimizer with default settings are used during the training. The batch size and number of epochs are selected as 32 and 200.

3.3.2. Datasets

Given the focus of this study on the spatio-temporal features, only sequential micro expression datasets are employed for the analysis. Figure 3.1 shows example images of subjects from the used datasets.

SAMM (Spontaneous Action, Multimodal, and Micro-expression) is a dataset that is commonly used for facial expression recognition research. The SAMM dataset includes 159 videos of 32 participants who were selected from the Manchester Metropolitan University, with a total duration of approximately 6 hours. On average, their age was 33.24 years. The participants belonged to various ethnic backgrounds, including 17 White British, three Chinese, two Arab, two Malay, and one each from African, Afro-Caribbean, Black British, White British/Arab, Indian, Nepalese, Pakistani, and Spanish backgrounds. The gender distribution was evenly split, with 16 male and 16 female participants. The

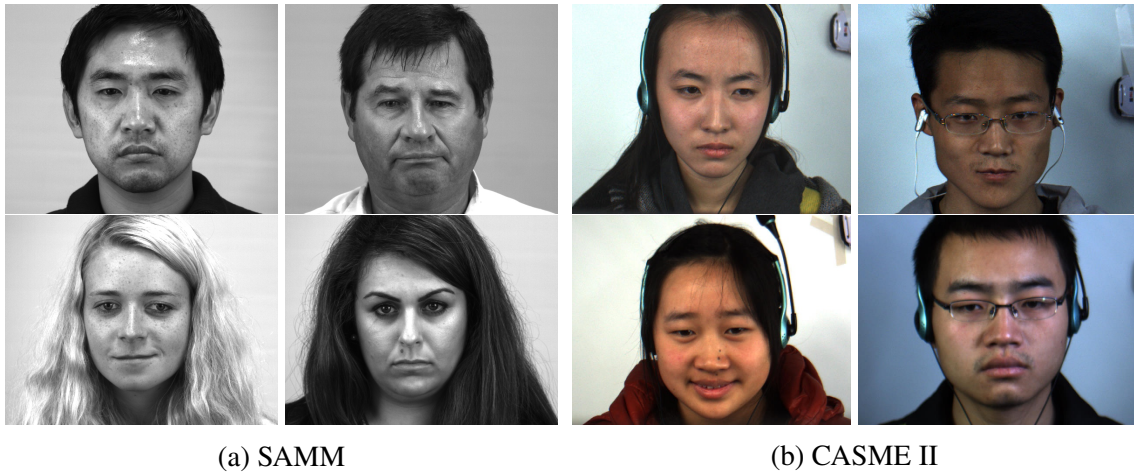


Figure 3.1: Collage of images in datasets which are used in micro-expression experiments. (a) SAMM dataset (Source: Davison et al., 2016). (b) CASME II dataset (Source: Yan et al., 2014).

videos were captured at 200 frames per second and at a resolution of 2040 1088 pixels. The dataset includes a variety of spontaneous facial expressions, including micro-expressions which are particularly challenging to detect and classify. The sequences in the SAMM dataset are annotated with frame-level labels indicating the onset, apex and offset times of each facial expression and detected emotion label of the overall sequence (Davison et al., 2016; Davison, Merghani, and Yap, 2018).

CASME II (Chinese Academy of Sciences Micro-expression) is another dataset that is commonly used for research on micro-expression recognition. The CASME II dataset includes 247 video clips from 26 participants with a mean age of 22.03 years and a total duration of approximately 30 minutes. The videos were captured at 200 frames per second and at a resolution of 640 x 480 pixels. The dataset includes a variety of micro-expressions. The sequences in the CASME II dataset are annotated with frame-level labels indicating the onset, apex and offset times of each facial expression and detected emotion label of the overall sequence (Yan et al., 2014).

3.3.3. Feature Creation Algorithm

The detection and classification of micro expressions pose greater challenges compared to macro expressions due to their brief duration and subtle intensity. Consequently, it became necessary to modify the methodology by introducing Phase-Based Video Motion (PBVM) Processing. This technique enhances the visibility of micro facial expressions, allowing for improved analysis and interpretation.

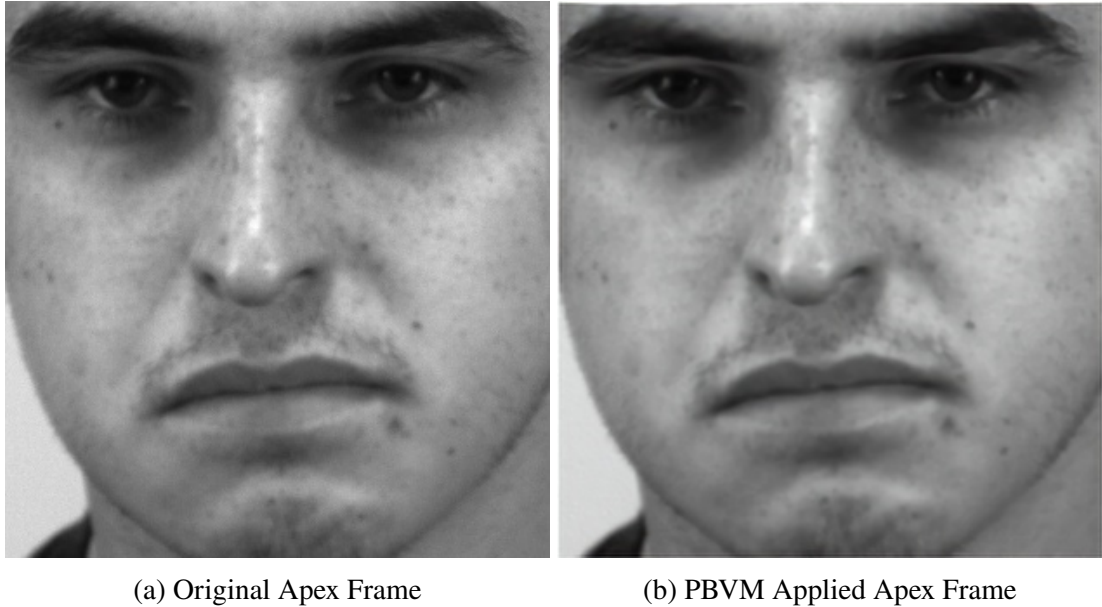


Figure 3.2: PBVM is applied for subject 006_1_2 in SAMM dataset. Subject demonstrates anger emotion and Brow Lowerer (AU 4) movement is enhanced using PBVM. Image of the subject is taken from SAMM dataset (Source: Davison et al., 2016)

In the experiments involving macro expressions, a subtraction operation was performed between the Euclidean distance features of each subject and their corresponding neutral state value. This approach yielded superior accuracy when compared to directly using the raw features for each frame individually. However, when dealing with micro-expressions, even with the application of PBVM processing, it was observed that there was insufficient change in the positions of facial landmarks. Despite the fact that PBVM processing amplifies the motion of the expression, making it potentially visible to the observer as depicted in Figure 3.2, it does not have a significant effect on the positions of facial landmarks. One possible explanation for this phenomenon is the presence of jittering in the landmark positions. During the calculation of landmarks for each sequential image, a small amount of jittering occurs, resulting in a shift in the positions of the landmarks compared to the previous consecutive image. Consequently, subtracting the Euclidean distance features from their corresponding neutral state did not yield promising results in terms of accuracy for micro-expression recognition. Hence, in the case of micro-expressions, only the raw features extracted from individual frames are employed. Figure 3.3 illustrates the algorithm flow for generating features from a single image. This flow is applied to the entire sequence of images, and the resulting features are appended to an array, which is then fed into the ConvLSTM1D block for further processing.

MediaPipe Face Mesh solution that is used in macro-expression experiments was upgraded in May 2023 to a new solution named as “Face Landmarker”. In this new version

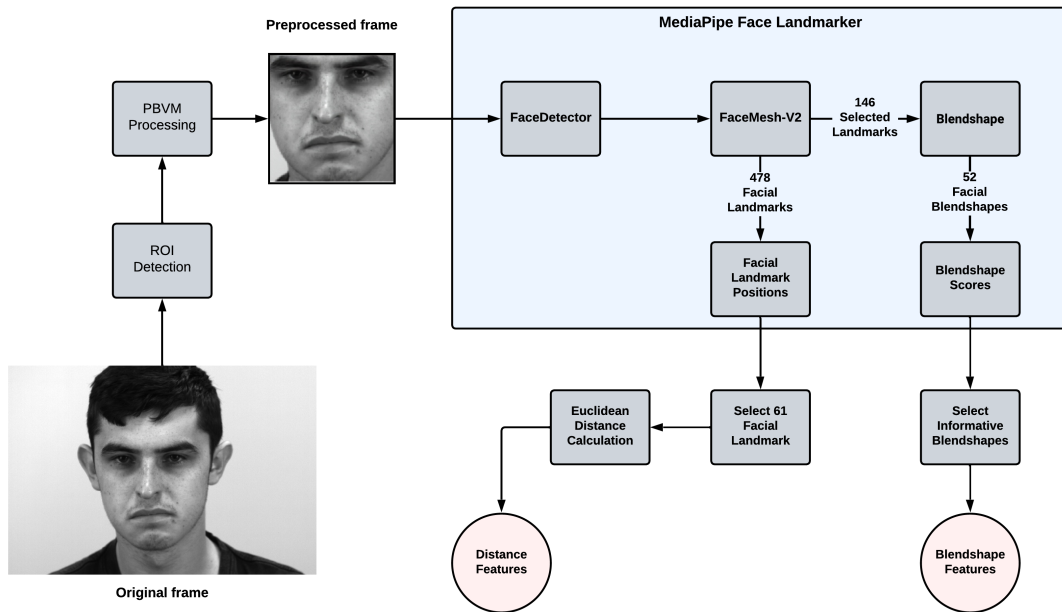


Figure 3.3: The algorithm flow for micro-expression experiments involves preprocessing each frame and feeding it into MediaPipe Face Landmarker to gather facial landmark positions and blendshape scores. Out of all the facial landmark positions, 61 are selected and Euclidean distance features are created from them. Informative blendshape scores are selected and blendshape features are created from them. These Euclidean distance features and blendshape features are then used as input for classification model. Image of the subject is taken from SAMM dataset (Source: Davison et al., 2016)

apart from facial landmark position a new feature called blendshape scores was introduced. Blendshapes are widely used in the digital production industry to create realistic facial animations (Anjyo, 2018). Each blendshape feature represents a specific facial expression or muscle action, linear weighted sum of these features creates blendshape model of the real subject (Anjyo, 2018).

Blendshape model inside Face Landmarker solution uses 146 landmarks, a subset of the 478 landmarks generated by FaceMesh model. Output of the blendshape model comprises 52 blendshape scores, represented as floating-point values within the range [0, 1] (Grishchenko et al., 2022). Predicted blendshapes can be seen in Table 3.2. Some of the predicted blendshapes are not helpful and can be misleading about emotion recognition such as eyeBlink, eyeLook and tongueOut features since these features are not related with any action units defined in FACS coding. So, all 11 features belonging to these groups are ignored and not included in further layers.

Table 3.2: List of predicted blendshapes

List of Predicted Blendshapes			
browDownLeft	eyeLookInRight	mouthClose	mouthRollLower
browDownRight	eyeLookOutLeft	mouthDimpleLeft	mouthRollUpper
browInnerUp	eyeLookOutRight	mouthDimpleRight	mouthShrugLower
browOuterUpLeft	eyeLookUpLeft	mouthFrownLeft	mouthShrugUpper
browOuterUpRight	eyeLookUpRight	mouthFrownRight	mouthSmileLeft
cheekPuff	eyeSquintLeft	mouthFunnel	mouthSmileRight
cheekSquintLeft	eyeSquintRight	mouthLeft	mouthStretchLeft
cheekSquintRight	eyeWideLeft	mouthLowerDownLeft	mouthStretchRight
eyeBlinkLeft	eyeWideRight	mouthLowerDownRight	mouthUpperUpLeft
eyeBlinkRight	jawForward	mouthPressLeft	mouthUpperUpRight
eyeLookDownLeft	jawLeft	mouthPressRight	noseSneerLeft
eyeLookDownRight	jawOpen	mouthPucker	noseSneerRight
eyeLookInLeft	jawRight	mouthRight	tongueOut

3.3.3.1. Phase-Based Video Motion Processing

The main idea behind Phase-Based Video Motion (PBVM) Processing is to analyze the phase information of video frames to detect and track motion. PBVM processing operates by first decomposing the video frames into their respective phase and magnitude components using a complex steerable pyramid. The phase component represents the local orientation of the image patterns, while the magnitude component reflects the strength of the patterns (Wadhwa et al., 2013).

After the decomposition, PBVM processing calculates the phase difference between consecutive frames to obtain the motion information. By using the phase difference, PBVM processing can detect motion even when the magnitude component is small, such as in regions with low contrast or texture. Additionally, PBVM processing can track both rigid and non-rigid motion, which is useful in applications such as video surveillance and motion analysis (Wadhwa et al., 2013). Figure 3.4 shows algorithm flow of PBVM processing.

One of the advantages of PBVM processing is its robustness to noise and occlusions. Since it operates on the phase component, which is less sensitive to noise and clutter, PBVM processing can still detect motion even when the magnitude component is heavily corrupted. Furthermore, PBVM processing can handle partial occlusions, where only a portion of the moving object is visible, by tracking the motion of the visible portion and propagating it to the occluded regions (Wadhwa et al., 2013).

The Python implementation of Phase-Based Video Motion Processing method is done in the article ‘Learning-based Video Motion Magnification’ by Oh et al. with

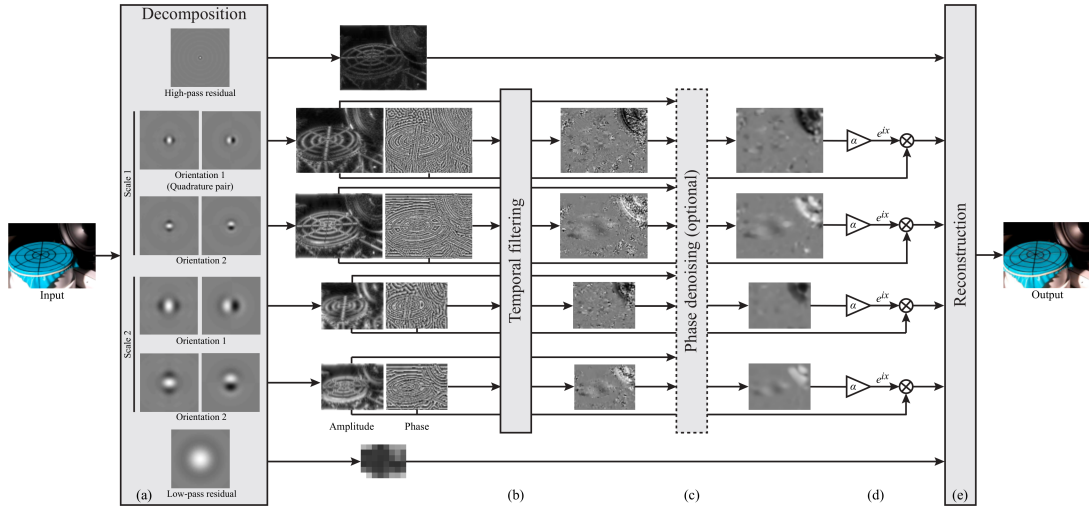


Figure 3.4: PBVM method involves analyzing local phase signals over time in different spatial scales and orientations using complex steerable pyramids. The amplitude of local wavelets is separated from their phase, and the phases are temporally filtered independently at each location, orientation, and scale. Spatial smoothing can be applied to increase the phase signal-to-noise ratio, which improves the results. The temporally-bandpassed phases are then amplified or attenuated, and the video is reconstructed (Source: Wadhwa et al., 2013)

a name ‘temporal filtering based processing’ (Oh et al., 2018). In our study we used this implementation to generate motion magnified images. Required inputs for temporal filtering based processing are amplification factor, low cut-off frequency, high cut-off frequency, sampling rate of video and filter type. There are 3 different filter types available first one is the FIR filter design using the window method, second one is Nth-order digital Butterworth filter and third one is difference of IIR which designs two lowpass filter for given low and high cutoff frequencies and creates band pass IIR filter (Oh et al., 2018).

Since the SAMM and CASME II datasets are captured at 200 frames per second, sampling rate variable is set to 200 for both datasets. For the filter type difference of IIR, a low cut-off frequency of 0.001 and a high cut-off frequency of 0.002 are chosen. Amplification factor is selected as 40. The selection of the amplification factor and cut-off frequencies is based on empirical considerations, aiming to achieve adequate magnification while avoiding excessive blurring of the image. These values are determined through experimentation to strike a balance between enhancing the desired features and maintaining visual clarity.

Selecting low and high cut-off frequencies for filter requires knowledge about motion to be magnified. If bandpass filter is selected too wide there can be unwanted magnification that corrupts the image. If it is selected too narrow, desired motion frequencies can be missed.

To avoid magnification and excessive blurring of unrelated parts of the image, a region of interest (ROI) capturing only the face region was introduced as seen in Figure 3.5. Phase-Based Video Motion Processing was applied to this ROI only. For the detection of the ROI, important facial landmarks from the original image were extracted using MediaPipe's Face Landmarker. The landmarks were selected from the eye, eyebrow, mouth, nose, and jaw regions of the face. After the landmark positions were obtained, a bounding box was drawn around the face region using OpenCV's 'boundingRect' function.

Figure 3.6 and 3.7 show blendshape scores of subject 006_1_2 that is seen in Figure 3.2 before and after PBVM processing is applied. Since the subject demonstrates action unit 4 which is brow lowerer movement, it can be observed that browDownLeft and browDownRight feature scores are increased after PBVM processing is applied.

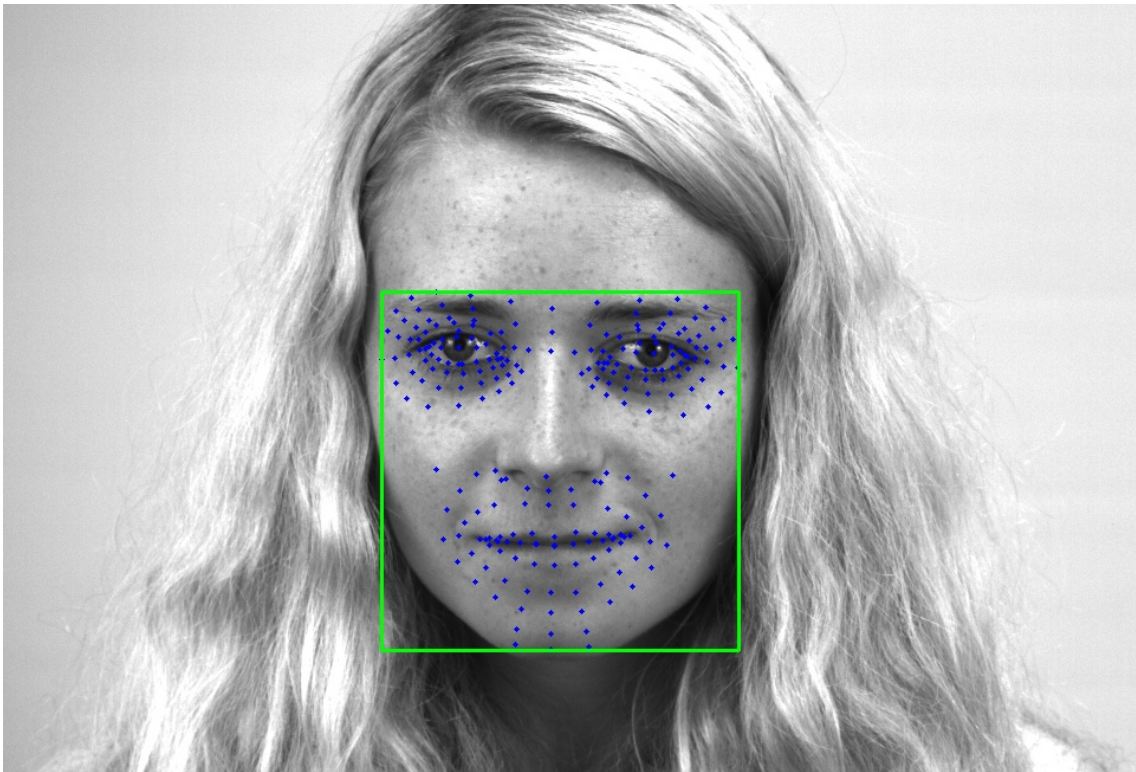


Figure 3.5: ROI extraction using facial landmarks and minimum bounding rectangle function of OpenCV. Image of the subject is taken from SAMM dataset (Source: Davison et al., 2016)

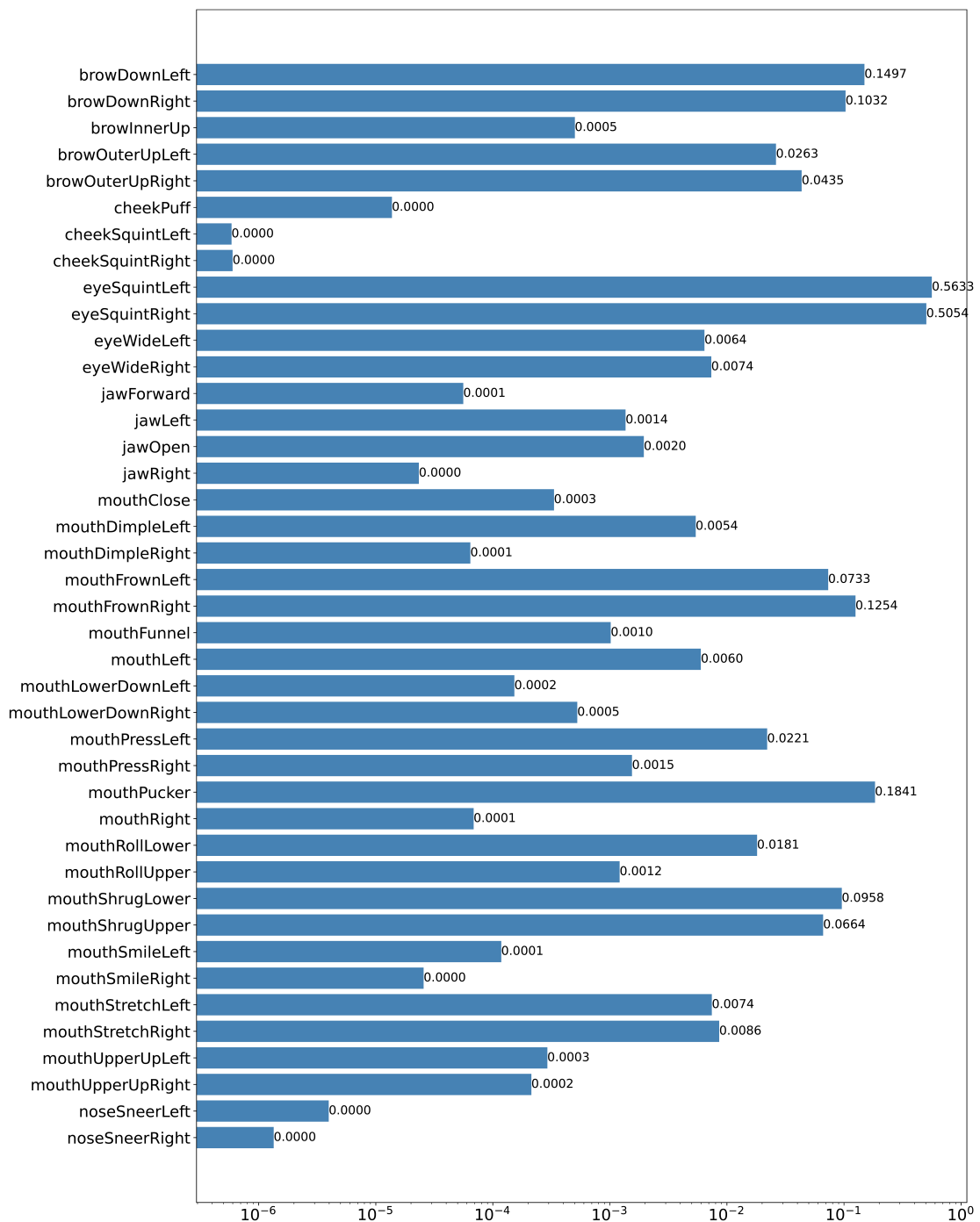


Figure 3.6: Mediapipe blendshape scores for original apex frame of the subject seen in Figure 3.2 (a)

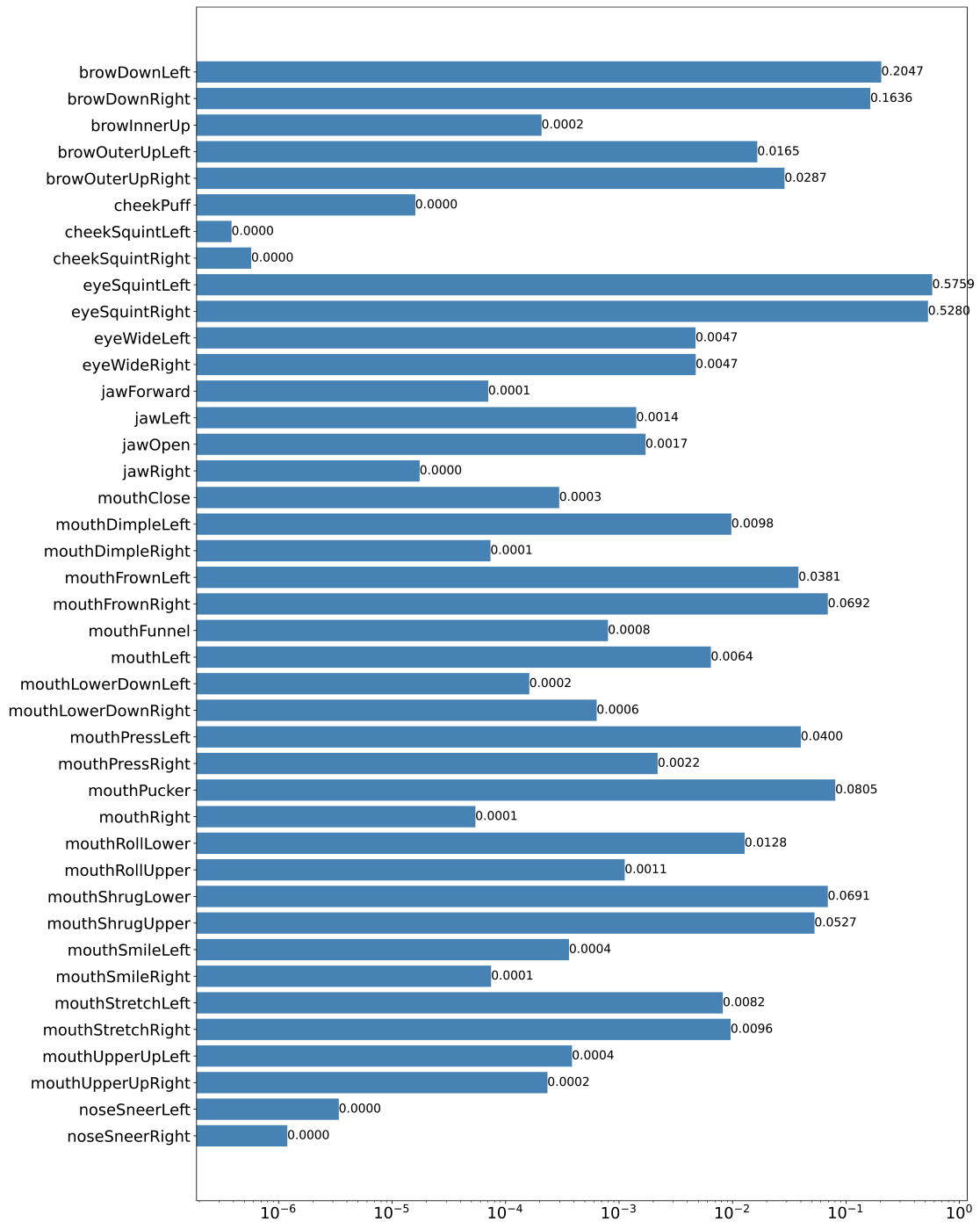


Figure 3.7: Mediapipe blendshape scores for PBVM applied apex frame of the subject seen in Figure 3.2 (b)

3.3.4. Classification Algorithm

Classification layers for micro-expression experiments consist of a ConvLSTM1D layer followed by multi-layer perceptron block which has 5 dense layers with sizes 4096, 2048, 2048, 1024, 1024 and finally a dense layer with softmax activation function as it is seen in figure 3.8.

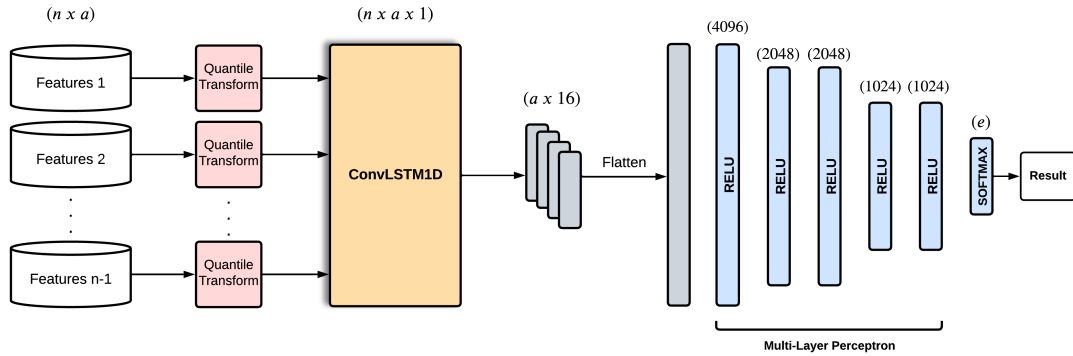


Figure 3.8: Classification algorithm for micro-expression experiments. First (n, a) shaped array that holds feature vectors for whole sequence is scaled using quantile transformer. Scaled 1D data is fed into ConvLSTM1D block. Output of ConvLSTM1D block is flattened and data is classified using multi-layer perceptron layers. Where, n : frame count, e : emotion count, a : feature count

Before classifying extracted features they are scaled using quantile transformation. Quantile transformation is a data transformation technique used in machine learning to map the probability distribution of a given dataset to a uniform or a normal distribution. It is a non-linear transformation, which means that it does not preserve the rank or order of the original data. (Pedregosa et al., 2011)

The quantile transformer works by estimating the cumulative distribution function (CDF) of the input data and then mapping it to a standard normal distribution (with a mean of 0 and a standard deviation of 1) or a uniform distribution (with values between 0 and 1). This transformation is useful for various machine learning tasks where the input features are expected to have a specific distribution, such as in linear regression or neural network models. (Pedregosa et al., 2011)

One of the primary advantages of the quantile transformer is that for a given feature, this transformation tends to spread out the most frequent values. It also reduces the impact of (marginal) outliers. It is also useful in cases where the input data has a nonlinear relationship with the target variable. (Pedregosa et al., 2011)

In scikit-learn, a popular Python machine learning library, the QuantileTransformer class can be used to perform the quantile transformation on the input data. (Pedregosa et al., 2011)

3.4. Results and Discussion

In Table 3.3, accuracies for all conducted experiments of micro-expression recognition is shown. Leave One Subject Out Cross Validation (LOSO CV) technique is used to validate proposed model. Four different experiments were conducted for each dataset, consisting of combinations of whether PBVM was applied or not, and whether Blendshape scores or Euclidean distance were used as the feature type. The results indicate that applying PBVM generally leads to a slight increase in accuracy, with one exception being the experiment conducted on the CASME II dataset using Euclidean distance features, where the accuracy did not change. Furthermore, the findings reveal that Euclidean distance features outperformed Blendshape score features, resulting in 3% to 5% more accurate predictions. Additionally, the processing times for creating Blendshape score and Euclidean distance features after an image is fed into the MediaPipe framework were measured as 17.8 ms and 20.9 ms respectively.

Table 3.3: Accuracy table for micro-expression experiments

Dataset	PBVM	Feature Type	Accuracy (LOSO CV)
SAMM	Yes	Blendshape Score	79.69%
SAMM	No	Blendshape Score	77.44%
CASME II	Yes	Blendshape Score	89.10%
CASME II	No	Blendshape Score	87.82%
SAMM	Yes	Euclidean Distance	84.21%
SAMM	No	Euclidean Distance	80.45%
CASME II	Yes	Euclidean Distance	92.94%
CASME II	No	Euclidean Distance	92.94%

Confusion matrices for SAMM and CASME II datasets can be visible in Figure 3.9, from these figures it can be observed that, classification accuracy of some emotions like fear and sadness gives poor results. The reason for that is the unbalanced distribution of number of subjects per emotion in datasets. In Figure 3.10, the bar plot of number of subjects per emotion is shown. Also, for SAMM dataset most of the wrong predictions by the model are confused with anger since the highest percentage of training data consist of

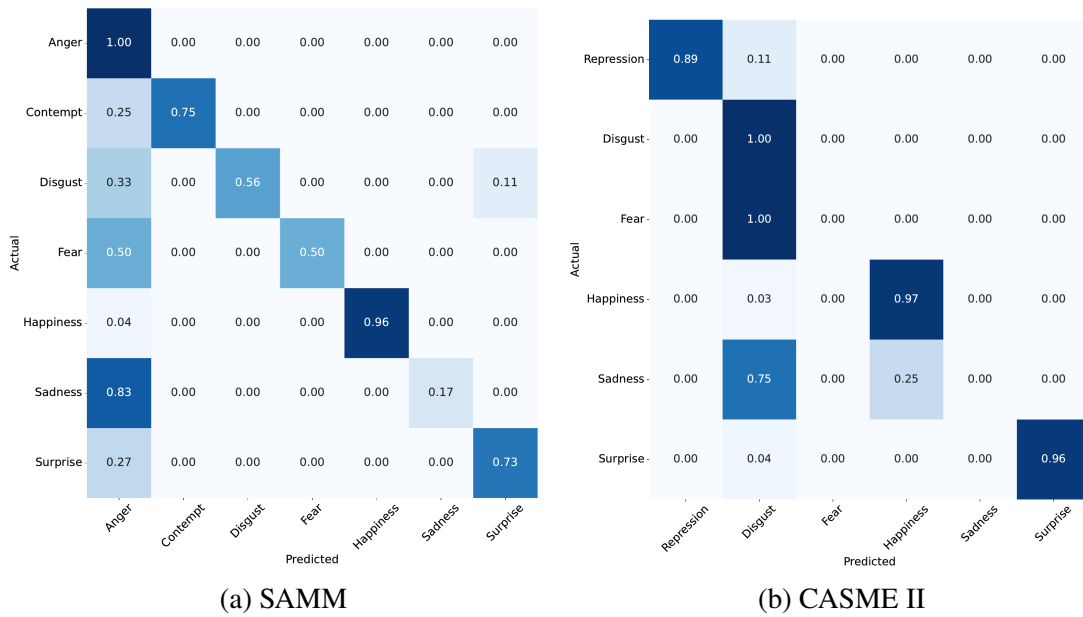


Figure 3.9: Confusion matrices for micro-expression experiments using Euclidean distance features.

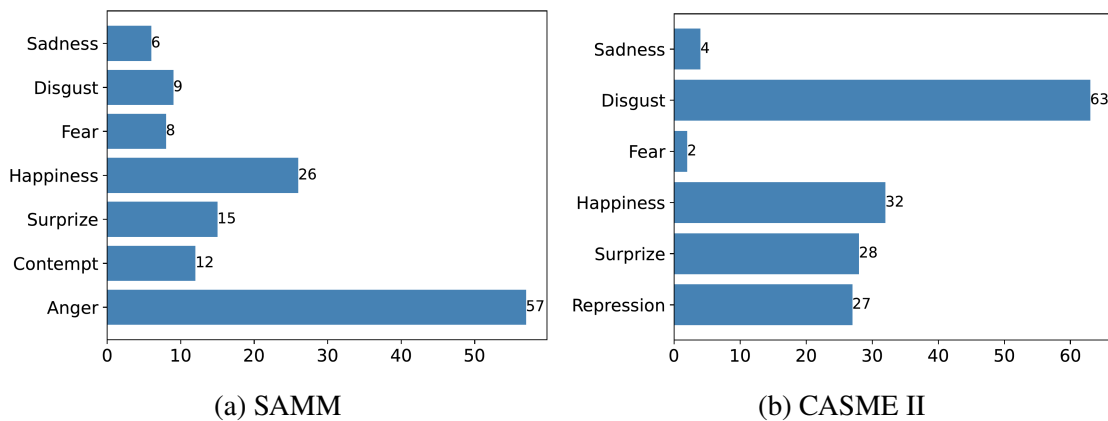


Figure 3.10: Emotions and number of subjects mapping

anger emotion. For CASME II dataset this dominant emotion is disgust.

3.5. Conclusion

In this study, we proposed a deep learning based sequential micro-expression recognition method by detecting facial landmarks using MediaPipe's Face Landmarker solution which is significantly faster than the popular Dlib facial landmark detection algorithm. We used Euclidean distances and blendshape scores as geometric features.

Since positional change of facial muscles are subtle, we applied PBVM processing in order to magnify landmark position change. It was observed that, even though PBVM processing magnifies the motion of expression it does not have significant effect on changes of landmark positions. With the proposed method we achieved competitive mean accuracy values among the landmark based methods in the literature using leave one subject out cross validation technique. We tested the proposed method with SAMM and CASME II datasets and achieved following maximum accuracy results respectively; 84.21%, 92.94%.

CHAPTER 4

CONCLUSION

In this thesis, we conducted two research to recognize emotions from macro and micro facial expressions. These two types of expressions require different perspectives for analysis due to their distinct characteristics. However, the proposed methods for both studies share some common properties. First common property is that only facial geometric features are used and they are created based on facial landmarks. To extract facial landmarks of a face in real-time we have used MediaPipe's FaceMesh, also as known as Face Landmarker solution in newer versions, which is significantly faster than the popular Dlib facial landmark detection algorithm. Second common property is that to have temporal information of a facial expression, unlike using a static snapshot of a person's face, sequential images from neutral till apex are used.

For macro expression study, to create geometrical features from facial landmarks we considered the difference of Euclidean distance of landmark pairs and angles with respect to neutral state of subjects. This approach provides auto calibration to the baseline of subjects which is changing from person to person. Also, it is shown that by tracking mean value of difference of distance features along the time, onset, apex and offset phases of an emotion can be detected. In our experiments we observed that increasing landmark count does not necessarily improve accuracy and sometimes it can have negative effects. Experiments with FACS based landmark grouping method show that selecting useful features using a feature reduction algorithm often increases classification accuracy. With the proposed method we achieved competitive mean accuracy values among the landmark based methods in the literature using 5-fold cross validation technique. We tested the proposed method with CK+, Oulu-CASIA VIS & NIR and MMI datasets and achieved following maximum accuracy results respectively; 93%, 79%, 77%, 68%.

In the study of micro expressions, where the duration of expressions is very short and their intensity is relatively weak compared to macro expressions, we applied PBVM processing to the image sequences to enhance the visibility of facial micro expressions. In addition to utilizing facial landmark positions to create Euclidean distance features, we also conducted a separate experiment using blendshape scores provided by MediaPipe's Face Landmarker. In total, we conducted four experiments per dataset: one with PBVM processing applied, one without PBVM processing, one using Euclidean distance features, and one using blendshape score features. Throughout our experiments, we observed that the effect of PBVM processing on facial landmark positions may not be significant

enough, depending on the dataset. Moreover, we found that Euclidean distance features yielded higher accuracy compared to blendshape scores, although the processing time to create them was longer compared to processing blendshapes. To evaluate the proposed method, we tested it on the SAMM and CASME II datasets using the Leave One Subject Out Cross Validation (LOSOCV) technique. The maximum accuracy results achieved for each dataset were 84.21% and 92.94% respectively, demonstrating the effectiveness of our approach in accurately recognizing micro expressions.

For facial expression recognition, finding accurate, robust and real-time solution still remains as a challenge. With increasing human-computer interaction, it is not hard to predict that many applications from different areas will desire to detect emotion of a human. Proposed system should be fast enough and robust against different illumination variations, rotation of face, facial accessories and any other distortion factors to be able to adopted by commercial applications.

REFERENCES

Adyapady, R Rashmi, and B Annappa. 2023. “A comprehensive review of facial expression recognition techniques.” *Multimedia Systems* 29 (1): 73–103.

Aloysius, Neena, and M Geetha. 2017. “A review on deep convolutional neural networks.” In *2017 international conference on communication and signal processing (ICCSP)*, 0588–0592. IEEE.

Álvarez, Victor M, Claudia N Sánchez, Sebastián Gutiérrez, Julieta Domínguez-Soberanes, and Ramiro Velázquez. 2018. “Facial emotion recognition: a comparison of different landmark-based classifiers.” In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, 1–4. IEEE.

Anjyo, Ken. 2018. “Blendshape Facial Animation.” In *Handbook of Human Motion*, 2145–2155. Cham: Springer International Publishing. ISBN: 978-3-319-14418-4. https://doi.org/10.1007/978-3-319-14418-4_2. https://doi.org/10.1007/978-3-319-14418-4_2.

Barrett, Lisa Feldman, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. “Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements.” PMID: 31313636, *Psychological Science in the Public Interest* 20 (1): 1–68. <https://doi.org/10.1177/1529100619832930>. eprint: <https://doi.org/10.1177/1529100619832930>. <https://doi.org/10.1177/1529100619832930>.

Bazarevsky, Valentin, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. “Blazeface: Sub-millisecond neural face detection on mobile gpus.” *arXiv preprint arXiv:1907.05047*.

Beh, Kai Xin, and Kam Meng Goh. 2019. “Micro-expression spotting using facial landmarks.” In *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*, 192–197. IEEE.

Buhari, Adamu Muhammad, Chee-Pun Ooi, Vishnu Monn Baskaran, Raphael CW Phan, KokSheik Wong, and Wooi-Haw Tan. 2022. “Invisible emotion magnification algorithm (IEMA) for real-time micro-expression recognition with graph-based features.” *Multimedia Tools and Applications* 81 (7): 9151–9176.

Buhari, Adamu Muhammad, Chee-Pun Ooi, Vishnu Monn Baskaran, Raphaël CW Phan, KokSheik Wong, and Wooi-Haw Tan. 2020. “Facs-based graph features for real-time micro-expression recognition.” *Journal of Imaging* 6 (12): 130.

- Choi, Dong Yoon, and Byung Cheol Song. 2020. "Facial micro-expression recognition using two-dimensional landmark feature maps." *IEEE Access* 8:121549–121563.
- Cohn, Jeffrey F, Zara Ambadar, and Paul Ekman. 2007. "Observer-based measurement of facial expression with the Facial Action Coding System." *The handbook of emotion elicitation and assessment* 1 (3): 203–221.
- Davison, Adrian K, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2016. "Samm: A spontaneous micro-facial movement dataset." *IEEE transactions on affective computing* 9 (1): 116–129.
- Davison, Adrian K, Walied Merghani, and Moi Hoon Yap. 2018. "Objective classes for micro-facial expression recognition." *Journal of Imaging* 4 (10): 119.
- Di, Xiaolei, Yu Xiao, Chao Zhu, Yang Deng, Qinpei Zhao, and Weixiong Rao. 2019. "Traffic congestion prediction by spatiotemporal propagation patterns." In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, 298–303. IEEE.
- Dong, Xuanyi, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. 2018. "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 360–368.
- Ekman, Paul. 1992a. "Are there basic emotions?" *Psychological Review* 99 (3).
- Ekman, Paul. 1992b. "Facial expressions of emotion: an old controversy and new findings." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335 (1273): 63–69.
- Friesen, Wallace V, Paul Ekman, et al. 1983. "EMFACS-7: Emotional facial action coding system." *Unpublished manuscript, University of California at San Francisco* 2 (36): 1.
- Grishchenko, Ivan, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. 2020. "Attention mesh: High-fidelity face mesh prediction in real-time." *arXiv preprint arXiv:2006.10962*.
- Grishchenko, Ivan, Geng Yan, Andrei Zanfir, and Eduard Gabriel Bazavan. 2022. *Model Card Mediapipe Blendshape V2*, November. <https://storage.googleapis.com/mediapipe-assets/Model%5C%20Card%5C%20Blendshape%5C%20V2.pdf>.
- Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long short-term memory." *Neural computation* 9 (8): 1735–1780.

- Jung, Heechul, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. "Joint fine-tuning in deep neural networks for facial expression recognition." In *Proceedings of the IEEE international conference on computer vision*, 2983–2991.
- Kanade, Takeo, Jeffrey F Cohn, and Yingli Tian. 2000. "Comprehensive database for facial expression analysis." In *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*, 46–53. IEEE.
- Kartynnik, Yury, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. *Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs*. arXiv: 1907.06724 [cs.CV].
- Khan, Fuzail. 2018. "Facial expression recognition using facial landmark detection and feature extraction via neural networks." *arXiv preprint arXiv:1812.04510*.
- Kim, Nayeon, Sukhee Cho, Chung Hyun Ahn, and Byungjun Bae. 2021. "Facial Micro-Expression Recognition in Video using Squeezed Landmark Feature Maps." In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 1107–1110. IEEE.
- Kumari, Jyoti, Reghunadhan Rajesh, and KM Pooja. 2015. "Facial expression recognition: A survey." *Procedia computer science* 58:486–491.
- Li, Shan, and Weihong Deng. 2020. "Deep facial expression recognition: A survey." *IEEE transactions on affective computing* 13 (3): 1195–1215.
- Lucey, Patrick, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, 94–101. IEEE.
- Mollahosseini, Ali, David Chan, and Mohammad H Mahoor. 2016. "Going deeper in facial expression recognition using deep neural networks." In *2016 IEEE Winter conference on applications of computer vision (WACV)*, 1–10. IEEE.
- Oh, Tae-Hyun, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. 2018. "Learning-based video motion magnification." In *Proceedings of the European Conference on Computer Vision (ECCV)*, 633–648.
- Pantic, Maja, and Ioannis Patras. 2006. "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36 (2): 433–449.

- Pantic, Maja, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. "Web-based database for facial expression analysis." In *2005 IEEE international conference on multimedia and Expo*, 5–pp. IEEE.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.
- Qiu, Yinghong, and Yi Wan. 2019. "Facial expression recognition based on landmarks." In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 1:1356–1360. IEEE.
- Raju, VN Ganapathi, K Prasanna Lakshmi, Vinod Mahesh Jain, Archana Kalidindi, and V Padma. 2020. "Study the influence of normalization/transformation process on the accuracy of supervised classification." In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735. IEEE.
- Rohith Raj, S, D Pratiba, P Ramakanth Kumar, and Dummy Name. 2020. "Facial expression recognition using facial landmarks: a novel approach." *ASETS J.* 5:24–28.
- Sharma, Garima, Latika Singh, and Sumanlata Gautam. 2019. "Automatic facial expression recognition using combined geometric features." *3D Research* 10:1–9.
- Shi, Xingjian, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." *Advances in neural information processing systems* 28.
- Tariq, Shahroz, Sangyup Lee, and Simon S Woo. 2020. "A convolutional lstm based residual network for deepfake video detection." *arXiv preprint arXiv:2009.07480*.
- Valstar, Michel, Maja Pantic, et al. 2010. "Induced disgust, happiness and surprise: an addition to the mmi facial expression database." In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 10:65. Paris, France.
- Wadhwa, Neal, Michael Rubinstein, Frédo Durand, and William T Freeman. 2013. "Phase-based video motion processing." *ACM Transactions on Graphics (TOG)* 32 (4): 1–10.
- Wu, Chung-Hsien, Jen-Chun Lin, and Wen-Li Wei. 2014. "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies." *APSIPA transactions on signal and information processing* 3:e12.

- Xia, Zhaoqiang, Xiaoyi Feng, Jinye Peng, Xianlin Peng, and Guoying Zhao. 2016. "Spontaneous micro-expression spotting via geometric deformation modeling." *Computer Vision and Image Understanding* 147:87–94.
- Xia, Zhaoqiang, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. 2019. "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions." *IEEE Transactions on Multimedia* 22 (3): 626–640.
- Xie, Hong-Xia, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2022. "An overview of facial micro-expression analysis: Data, methodology and challenge." *IEEE Transactions on Affective Computing*.
- Yan, Wen-Jing, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. "CASME II: An improved spontaneous micro-expression database and the baseline evaluation." *PloS one* 9 (1): e86041.
- Zhang, Ligang, and Dian Tjondronegoro. 2011. "Facial expression recognition using facial movement features." *IEEE transactions on affective computing* 2 (4): 219–229.
- Zhao, Guoying, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen. 2011. "Facial expression recognition from near-infrared videos." *Image and vision computing* 29 (9): 607–619.

APPENDIX A

LICENSE AGREEMENT OF CK+ DATASET

CK and CK+ DATABASE USER AGREEMENT

All requests must be made by a faculty member at a university or college.

CK+ may be used for non-commercial research that is not subject to US export controls. To obtain a copy of the database, please complete the following agreement and return it to Megan Ritter meri60@pitt.edu.

Once the signed agreement is received and approved, you will receive instructions to download the database via Box hosted at the University of Pittsburgh. The database remains the property of Dr. Jeffrey Cohn. Use is subject to the following terms. For questions, please contact Megan Ritter at the address above.

By signing this agreement, you agree:

- To cite the following publications in any paper of yours or your collaborators that makes any use of the database.
 - Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Grenoble, France, 46-53.
 - Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, San Francisco, USA, 94-101.
- To use the images for non-commercial research purposes only.
- Not to provide any portion of the database to other parties.
- In any publications, print, electronic, or other media to use images from only the following subjects and to include notice of copyright (©Jeffrey Cohn):
 - S52, S55, S74, S106, S111, S113, S121, S124, S125, S130, S132

All requests must include the following information.

Faculty member's name: Abdurrahman Gumus

Faculty member's official title: (e.g., Assistant professor) Assistant Professor

Faculty member's university email address: abdurrahmangumus@iyte.edu.tr

Faculty member's signature:

Name of any additional requestor: Talha Enes Koksall

Other requestor's official title (e.g., student or postdoc): Graduate Student

Other requestor's university email address: talhakoksall@iyte.edu.tr

Other requestor's signature:

APPENDIX B

LICENSE AGREEMENT OF OULU-CASIA DATASET

License Agreement
"Oulu-CASIA NIR&VIS facial expression database"
v.05.11.2018

By signing this document the user, intended as who will make use of the database, agrees to the following terms.

1. Commercial use

The user may not use the database for any commercial purpose. Commercial purposes include, but are not limited to:

- proving the efficiency of commercial systems,
- testing commercial systems,
- using screenshots of subjects from the database in advertisements,
- selling data from the database,
- broadcasting data from the database.

2. Distribution

The user may not distribute or broadcast the database in any form. Small portions (i.e.; sample images) may be used in academic publications and presentations by strictly abiding the following conditions:

- A. showing with necessity: only show the real face photo if a drawing or fake face won't be enough for the illustration purpose.
- B. only use the least amount of face photos for you need.
- C. never link the used face photo with any possible trait of personal info (e.g., city, university, etc.)
- D. not all subjects agreed that their data can be cited in papers or presentations, sample images should be chosen only from these subject ID: **P002, P023** and **P038**.

3. Access

The user may only use the database after this LA has been signed by his/her group leader or professor (who has fixed position in the data requiring institution), and returned to the Center for Machine Vision and Signal Analysis (CMVS). The user must send a scanned copy of the signed and dated LA by email, in PDF format to: guoying.zhao@oulu.fi, and if a student is sending the email, the email should be copied to the leader or professor who signed the LA.

With the group leader or professor signing the LA, we agree that all students and researchers that reported to the signer are eligible of using the data. The signer should get familiar with the latest EU General Data Protection Regulation (GDPR), and is responsible for controlling who in his/her group gets access to the data, and how the data will be stored and used properly according to GDPR. The signer will be fully responsible if any issue rises in relation with the GDPR about the usage of shared data in his/her group.

4. Publications

Publications include not only research papers, but also presentations for conferences or educational purposes.

All documents and research papers that report on research that use the “**Oulu-CASIA NIR&VIS facial expression database**” should include a citation to:

G. Zhao, X. Huang, M. Taini, S.Z. Li & M. Pietikäinen (2011): **Facial expression recognition from near-infrared videos**. *Image and Vision Computing*, 29(9):607-619.

The user will send an electronic copy of all papers that reference the database to: guoying.zhao@oulu.fi

5. Research

The user may only use the database for scientific research, but should not be used for commercial purpose in any form (e.g., for testing, training models which could be embedded in software or products that related to commercial plans, etc.)


6. Changes

Both The Center for Machine Vision and Signal Analysis (CMVS) and the Institute of Automation, Chinese Academy of Science (CASIA) are allowed to change this LA at any time; users will be informed about changes beforehand and given the choice to opt out of the new LA. Opting out will render the previous LA void.

7. Warranty

The database comes without any warranty, neither CMVS nor CASIA can be held accountable for any damage (physical, financial or otherwise) caused by the use of the database. CMVS and CASIA will try to prevent any damage by keeping the database virus free.

If you read and agree with this LA, please sign here:

Name Abdurrahman Gumus	Title Assistant Professor
Affiliation (Institute, University, ...) Electrical and Electronics Engineering Izmir Institute of Technology, Turkey	
Work email (at your affiliation) abdurrahmangumus@iyte.edu.tr	
Detail Address Izmir Institute of Technology Gulbahce Mah. Urla, Izmir, 35430, Turkey	
Signature 	Date Aug 08 2022

APPENDIX C

LICENSE AGREEMENT OF MMI DATASET

End User License Agreement MMI Facial Expression Database (<http://www.mmifacedb.com>)

By signing this document the user, he or she who will make use of the database or the database interface, agrees to the following terms.

With database, we denote both the actual data as well as the interface to the database.

1. Commercial use

The user may not use the database for any non-academic purpose. Non-academic purposes include, but are not limited to:

- proving the efficiency of commercial systems
- training or testing of commercial systems
- using screenshots of subjects from the dataset in advertisements
- selling data from the dataset
- creating military applications
- developing governmental systems used in public spaces

2. Responsibility

This document must be signed by a person with a permanent position at an academic institute (the signee). Up to five other researchers affiliated with the same institute for whom the signee is responsible may be named at the end of this document which will allow them to work with this dataset.

3. Distribution

The user may not distribute the database or portions thereof in any way, with the exception of using small portions of data for the exclusive purpose of clarifying academic publications or presentations. **Only data from subjects who gave consent to have their data used in publications and presentations may be used for this purpose.** Note that publications will have to comply with the terms stated in article 5.

4. Access

The user may only use the database after this End User License Agreement (EULA) has been signed and returned to the iBUG Group at Imperial College London. The signed EULA should be returned in digital format by uploading it to the website when requesting account at:

<http://www.mmifacedb.com/accounts/register/>

Only if the user is not capable of requesting an account in this manner, accounts may be requested by sending the signed EULA via traditional mail to:

Prof. Maja Pantic
Imperial College London
Department of Computing
180 Queen's Gate
London SW7 2A U.K.

The user may not grant anyone access to the database by giving out their user name and password.

5. Publications

Publications include not only papers, but also presentations for conferences or educational purposes.

The user may only use data of subjects in publications if that particular subject has explicitly granted permission for this. This is specified with every database element.

All documents and papers that report on research that use any of the MMI Facial Expression Database will acknowledge this as follows:

“(Portions of) the research in this paper uses the MMI-Facial Expression Database collected by Valstar and Pantic”

And include a citation to:

M.F. Valstar, M. Pantic, “*Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database*”, Proceedings of the International Language Resources and Evaluation Conference, Malta, May 2010

In addition, any user that uses Part I of the database in their report shall also cite the following paper:

M. Pantic, M.F. Valstar, R. Rademaker and L. Maat, “*Web-based database for facial expression analysis*”, Proc. IEEE Int'l Conf. on Multimedia and Expo (ICME'05), Amsterdam, The Netherlands, July 2005

A description of the different parts is given in the first paper (Valstar & Pantic 2010).

The user will send a copy of any document or papers that reports on research that uses the MMI Facial Expression Database to Dr. Maja Pantic or to <mmi_face_db@mahnob-db.eu>.

6. Academic research

The user may only use the database for academic research.

7. Warranty

The database comes without any warranty. The iBUG Group at Imperial College London can not be held accountable for any damage (physical, financial or otherwise) caused by the use of the database. The iBUG Group at Imperial College London will try to prevent any damage by keeping the database virus free.

8. Misuse

If at any point, the administrators of MMI Facial Expression database and/or iBUG Group at Imperial College have a reasonable doubt that the user does not act in accordance to this EULA, he/she will be notified of this and will immediately be declined the access to the database.

User: Abdurrahman Gumus
User's Affiliation: Izmir Institute of Technology, Turkey
User's address: Gulbahce Mah. IYTE Kampusu Urla Izmir 35430
User's e-mail: abdurrahmangumus@iyte.edu.tr
Additional Researcher 1: Talha Enes Koksal
Additional Researcher 2: _____
Additional Researcher 3: _____
Additional Researcher 4: _____
Additional Researcher 5: _____

Signature:

Date/place

November 29 2022

2/2

APPENDIX D

LICENSE AGREEMENT OF CASME II DATASET

License Agreement

Chinese Academy of Sciences Micro-expression Database II (CASME2)

I agree

- to use the video or images for research purposes only.
- not to provide the video or images to second parties.
- if I reproduce images and video in electronic or print media, to use only in **scientific journals** and include notice of copyright (©Xiaolan Fu). Images and video from sub12 and sub22 should not be published.

Signature: 

Name: Abdurrahman Gumus

Title: Assistant Professor

Institution: Izmir Institute of Technology, Turkey

Date: December 07, 2021

Email address: abdurrahmangumus@iyte.edu.tr
eee.iyte.edu.tr

APPENDIX E

LICENSE AGREEMENT OF SAMM DATASET

SAMM Dataset: Release Agreement

Introduction:

The goal of the Spontaneous Activity and Micro-Movements (SAMM) Dataset is to develop new techniques, technology, and algorithms for the automatic interpretation and analysis of micro-movements. Manchester Metropolitan University ("MMU") and the Emotional Intelligence Academy is involved in an ongoing effort to develop this dataset of high-speed video data. The dataset is meant to aid research efforts in the general area of developing, testing and evaluating algorithms for facial micro-movement analysis. MMU has copyright on the data and is the principal distributor of the SAMM Dataset.

Release of the dataset:

To advance the state-of-the-art in micro-movement analysis, this dataset is made available to the research community. All other uses of the dataset will be considered on written application to the MMU, on the case-by-case basis. To receive access to the dataset, for non-commercial research into micro-movement analysis and other branches of related research, you must sign this document agreeing to the conditions and restrictions listed below:

Consent:

I/We agree to the following conditions and restrictions of access and use of the SAMM Dataset:

1. **Redistribution:** Without prior written approval from MMU, the SAMM Dataset, will not be further distributed, published, copied, or disseminated in any way or form whatsoever, in whole or in part, whether for profit or not. This includes further distributing, copying or disseminating to a different facility, department or organisational unit within this university, organisation, or company.
2. **Modification and Commercial Use:** Without prior written approval from MMU, the SAMM Dataset, in whole or in part, may not be modified or used for commercial purposes. If any unauthorised use is made of the Dataset and such use is attributable to your act or default then, without prejudice to MMU's other rights and remedies, MMU can terminate this Agreement with immediate effect by serving you with written notice.
3. **Requests for the SAMM Dataset:** All requests for the SAMM Dataset will be forwarded to MMU Principal Investigator(s).
4. **Publication Requirements:** Where permitted to publish by agreement of MMU, publication

will be restricted to paper, web-based data and image data, for scientific purposes only, in summary forms. Images of participants should not be over-used in their original form in publication to avoid participant embarrassment or mental anguish. Participants who have not agreed to have their images published will be outlined in the dataset information sheet. The images of these participants are expressly prohibited to be shown in any form of publication. Their data still may be used in summary forms (i.e. tables and diagrams only, not images) that prevents these participants from being identified.

5. **Intellectual Property Rights Ownership:** You acknowledge that all Intellectual Property Rights in the SAMM Dataset are the property of MMU or its licensors, as the case may be; and you shall have no rights in or to the SAMM Dataset other than the right to use it in accordance with the express terms of this Agreement. You assign to MMU, and shall assign to it, with full title guarantee all Intellectual Property Rights arising from the use of the Dataset, by way of future assignment. You shall use all reasonable endeavours to procure that any necessary third party shall, at MMU's cost, promptly execute such documents and perform such acts as may reasonably be required for the purpose of giving full effect to this Agreement. The Intellectual Property Rights assigned to MMU shall be deemed to be included in the Release of the Dataset from the date when such rights arise.

6. **Citation/Reference:** All documents and papers that report on research that uses the SAMM Dataset will acknowledge the use of the dataset by including an appropriate citation to the following:

A. K. Davison; C. Lansley; N. Costen; K. Tan; M. H. Yap (2018), "SAMM: A Spontaneous Micro-Facial Movement Dataset," in IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 116-129. doi: 10.1109/TAFFC.2016.2573832

For users who report on micro-expression recognition on emotion classes and/or objective classes should include a citation to:

A.K. Davison, W. Merghani and M.H. Yap (2018), "Objective classes for micro-facial expression recognition", Journal of Imaging, 4(10), p.119.

For users who report on SAMM Long Videos micro- and macro-expressions recognition and spotting should include a citation to:

C.H. Yap, C. Kendrick and M.H. Yap (2019), "SAMM Long Videos: A Spontaneous Facial Micro- and Macro-Expressions Dataset", <https://arxiv.org/abs/1911.01519>.

7. **Publications to MMU:** A copy of all reports and papers that are for public or general release that use the SAMM Dataset should be forwarded immediately upon release or publication to MMU Principal Investigator(s).

8. **Destruction of the Dataset:** The SAMM Dataset has a limited life span. Subject to earlier termination of this Agreement in accordance with clauses 2 and 9, you must delete all your copies of the raw data not later than 30th January 2026. If the Agreement is subject to earlier termination, you must delete all your copies of the raw data not later than the date of

termination. This does not apply to publications or reports which have been forwarded to MMU Principal Investigator(s) in accordance with the preceding clause. You shall permit MMU and its third party representatives, on reasonable notice, but without notice in case of any reasonably suspected breach of this clause 8, to gain (physical and remote electronic) access to your systems to ensure the destruction of the Dataset. Such audit rights shall continue for three years after termination of this Agreement. You shall give all necessary assistance to the conduct of such audits during the term of this Agreement and for a period of three years after termination of this Agreement.

9. **Termination:** MMU may terminate this Agreement (wholly or in part) at any time with immediate effect. Either party may terminate this Agreement with immediate effect by giving written notice to the other party if the other party commits a material breach of any term of this Agreement and (if that breach is remediable) fails to remedy that breach within a period of 30 days after being notified in writing to do so.
10. **Assignment:** This Agreement is personal to you and you shall not assign, transfer, mortgage, charge, sub-contract, declare a trust of or deal in any other manner with any of its rights and obligations under this Agreement without the prior written consent of the MMU. You confirm that you are acting on your own behalf and not for the benefit of any other person. MMU may at any time assign, transfer, mortgage, charge, sub-contract, declare a trust of or deal in any other manner with any of its rights and obligations under this Agreement without your consent.
11. **No Warranty:** THE PROVIDER OF THE DATA MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE MATERIAL WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADE- MARK, OR OTHER PROPRIETARY RIGHTS.
12. **Governing Law:** This Agreement and any dispute or claim arising out of or in connection with it or its subject matter or formation (including non-contractual disputes or claims) shall be governed by and construed in accordance with the law of England and Wales.
13. **Jurisdiction:** Each party irrevocably agrees that the courts of England and Wales shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this agreement or its subject matter or formation (including non-contractual disputes or claims).
14. **Access:** The user may only use the datasets after this License Agreement has been signed by his/her group leader or professor (who has fixed position in the data requiring institution). If a student is sending the email, the email should be copied to the leader or professor who has signed the License Agreement.

By signing this license agreement, I agree that all students and researchers that reported to me are eligible of using the data. I am familiar with the latest EU General Data Protection Regulation (GDPR), and I am responsible for controlling my research group gets access to the data, how the data will be stored and used properly according to GDPR. I am fully responsible if any issue rises in relation to the GDPR on the usage of shared data in my group.

Signature: _____
Name (please print): Abdurrahman Gumus
Role: Assistant Professor
Date: December 07, 2021
Email: abdurrahmangumus@iyte.edu.tr,
eee.iyte.edu.tr
Organisation: Izmir Institute of Technology, Turkey
Address: İZTECH Electric and Electronics
Engineering Department
Gulbahce Urla 35430 İzmir, Turkey

Please email a scanned signed copy to the SAMM Dataset Principal Investigator - Moi Hoon Yap at M.Yap@mmu.ac.uk.

Address: Manchester Metropolitan University School of Computing, Mathematics and Digital Technology, John Dalton Building, Chester Street, Manchester M1 5GD.