Spontaneous Facial Micro Expression Recognition and Analysis using Varying Resolutions

by

Pratikshya Sharma

Thesis submitted for the degree of

Doctor of Philosophy (PhD)

Intelligent Systems Research Centre
School of Computing, Engineering & Intelligent Systems
Faculty of Computing, Engineering and the
Built Environment
Ulster University

April 2022

I confirm that the word count of this thesis is less than 100,000 words

ABSTRACT

During the early years of facial expression research, works have mostly employed macro expressions which are easily identifiable. In contrast, in recent years utilizing facial micro expression has gained more acknowledgement in facial analysis due to stronger genuineness of its attributes. Subsequently, emotion analysis through facial micro expression has higher acceptability especially in psychology, autism, pain assessment, security, criminal investigations, and similar circumstances that demand critical decision making. Owing to its cross-discipline application, today micro expression analysis using facial images remains an active research field. Due to extreme minuteness of these expressions, they are often missed during observations however, studies show with the introduction of computer vison and machine/deep learning algorithms they have a higher chance of being identified. Therefore, the focus of this thesis is to conduct thorough investigations and design novel approaches for micro expression analysis employing suitable methods.

Most of the existing literature has overlooked the phase information while describing image patterns specifically to achieve micro expression recognition. Consequently, this thesis investigates the effectiveness of employing phase information for micro expression analysis. Furthermore, interpolation and video magnification are also introduced in later experiments to aid the extraction method. Additionally, the literature also highlighted the absence of adequate work examining the impact of resolutions and image quality for micro expression analysis. Therefore, the aim of this thesis is to explore micro expression to design a pipeline capable of boosting the expression recognition performance. Moreover, this thesis establishes threefold contributions to address the research gap: firstly, a pipeline that exploits interpolation, phase and temporal information in a non-cross database environment is utilized. Secondly, influence of video magnification is examined to improve expression recognition within this pipeline. Third, a novel pipeline to employ low quality micro expression images is developed by reconstructing such images using deep learning and generative adversarial networks.

In this thesis, to verify the suitability of combining phase, temporal, and magnification methods for micro expression, experiments are conducted on seven spontaneous micro expression databases. Results obtained clearly indicate the approach is as competitive as any other existing traditional methods. Furthermore, the experimental results obtained after introducing deep learning and generative adversarial networks into the second novel pipeline clearly highlight the significance of image reconstruction in achieving recognition boost even when the quality of input is compromised. Therefore, this thesis establishes significant progress towards the development of techniques for micro expression recognition that can be collaborated with medical/security and similar fields to assist in identifying vital cues.

ACKNOWLEDGEMENTS

This momentous experience certainly wouldn’t have happened without the support I received from many people throughout my research. First and foremost, I am extremely grateful to my supervisors Professor Sonya Coleman, Dr. Pratheepan Yogarajah and Dr. Laurence Taggart for giving me this excellent research opportunity at Ulster University, UK. The constant support, encouragement, patience, and knowledge from Professor Sonya has helped me cultivate my scientific research ability. Invaluable insight on how to conduct research, helpful advice along with personal support received from Dr. Pratheepan has helped me profoundly in solving research problems throughout this period. Working closely with Dr. Laurence has been very fruitful in grasping the required understanding in the field of autism. By sharing his expertise during several discussion sessions has helped me immensely in conceptualizing my research.

I extend my special thanks to external supervisor Dr. Pradeepa Samarsinghee, Sri Lanka Institute of Information Technology for her support and cooperation in all stages of my conference and journal paper publications. Her expert technical assistance through assessments and suggestions continuously guided me in strengthening my technical writing competency.

I wish to express my gratitude to Ulster University for the financial support I received through Vice-Chancellor’s Research Scholarship (VCRS) to conduct this research at Intelligent Systems Research Centre (ISRC).

I offer a special thanks to my colleagues and friends for making my research experience very enjoyable.

Lastly, I am forever grateful to my parents for their constant encouragement that helped me embark on this journey. The unconditional love and moral support from them, my siblings, my nephews, and every other member of my family helped me manage stressful situations throughout these years.

Contents

List of Figures

Figure Page

List of Tables

List of Abbreviations

3D	Three Dimensional
ASM	Active Shape Modelling
ASD	Autism Spectrum Disorder
AU	Action Unit
AU-IGAN	AU Intensity Controllable GAN
Bi-WOOF	Bi-Weighted Oriented Optical Flow
BN	Batch Normalization
CASME	Chinese Academy of Sciences Micro-expression
CBP-TOP	Centralized Binary Pattern on Three Orthogonal Planes
CDMER	Cross Database Micro Expression Recognition
CLM	Constrained Local Model
CM	Contiguous Memory
CNN	Convolutional Neural Network
CV	Computer Vision
DFF	Dense Feature Fusion
DFT	Discrete Fourier Transform
DL	Deep Learning
DRMF	Discriminative response map fitting
EAI	Emotion Avatar Image
EDSR	Enhanced Deep Super Resolution Network
ELM	Extreme Learning Machine
EMM	Eulerian Motion Magnification
ESRGAN	Enhanced Super Resolution Generative Adversarial Network
EVM	Eulerian Video Magnification
FACS	Facial Action Coding System
FCNN	Fully Convolutional Neural Network
FE	Facial Expression
FER	Facial Expression Recognition
fps	Frames per second
FSRCNN	Fast Super Resolution Convolutional Neural Network
GAN	Generative Adversarial Network
GFF	Global Feature Fusion
GLMM	Global Lagrangian Motion Magnification
GRL	Global Residual Learning
GSL	Group Sparse Learning
HIGO	Histogram of Image Gradient Oriented on Three Orthogonal Planes
HOG	Histogram of Gradient
HOG-TOP	Histogram of Gradient on Three Orthogonal Plane
HOOF	Histogram of Optical Flow Orientation
HR	High Resolution
HS	High Speed
ICE-GAN	Identity-aware and Capsule Enhanced Generative Adversarial Network
IIR	Infinite Impulse Response
ISR	Image Super Resolution
kNN	k-Nearest Neighbour

LBP	Local Binary Pattern
LBP-MOP	Local Binary Pattern - Mean Orthogonal Plane
LBP-SIP	Local Binary Pattern with Six Intersection Points
LBP-TOP	Local Binary Pattern on Three Orthogonal Plane
LFF	Local Feature Fusion
LPQ	Local Phase Quantisation
LPQ-TOP	Local Phase Quantisation on Three Orthogonal Plane
LR	Low Resolution
LReLU	Leaky Rectified Linear Unit
LRL	Local Residual Learning
LSTM	Long Short-Term Memory model
LWM	Local Weighted Mean
MDMO	Main Direction Mean Optical Flow
ME	Micro Expression
MER	Micro Expression Recognition
METT	Micro Expression Training Tool
MKL	Multiple Kernel Learning
ML	Machine Learning
MMEW	Micro and Macro Expression Warehouse
nESRGAN+	Further Improved Enhanced Super Resolution Generative Adversarial Network
NIR	Near Infra-red
PCA	Principal Component Analysis
PI	Perceptual Index
PReLU	Parametric Rectified Linear Unit
PSNR	Peak Signal to Noise Ratio
RBF	Radial Basis Function
RDB	Residual Dense Block
RDN	Residual Dense Network
ReLU	Rectified Linear Unit
RRDB	Residual in Residual Dense Block
RRDRB	Residual-in-Residual Dense-Residual-Block
SAMM	Spontaneous Micro-Facial Movement
SMIC	Spontaneous Micro-expression database
SR	Super Resolution
SRCNN	Super Resolution Convolutional Neural Network
SRGAN	Super Resolution Generative Adversarial Network
SSIM	Structural Similarity Index
STFT	Short Term Fourier Transform
STLBP-IP	Spatio Temporal Local Binary Pattern with Integral Projection
STTM	Spatio temporal Texture Map
SVM	Support Vector Machine
TD	Typically Developed
TIM	Temporal Interpolation Model
VGG	Visual Geometry Group
VIS	Visual
York DDT	York Deception Detection Test

List of Publications

Conference Publications

1. Sharma, P., Coleman, S., Yogarajah, P. & Laurence, T. (2019). Micro expression classification accuracy assessment. IMVIP 2019: Irish Machine Vision & Image Processing, Technological University Dublin, Dublin, Ireland, August 28-30. doi:10.21427/kbny-0a41.

2. Sharma,P., Coleman, S., Yogarajah, P. , Taggart, L. and Samarasinghe P., (2021) "Magnifying Spontaneous Facial Micro Expressions for Improved Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 7930-7936, doi: 10.1109/ICPR48806.2021.9412585.

3. Sharma, P., Coleman, S., Yogarajah, P., Taggart, L. and Samarasinghe, P. (2022), Evaluation of Generative Adversarial Network Generated Super Resolution Images for Micro Expression Recognition. In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - ICPRAM, ISBN 978-989-758-549-4, pages 560-569. DOI: 10.5220/0010820100003122.

Journal Publication

4. Sharma, P., Coleman, S., Yogarajah, P. et al. Comparative analysis of super-resolution reconstructed images for micro-expression recognition. Adv. in Comp. Int. 2, 24 (2022). https://doi.org/10.1007/s43674-022-00035-x.

Chapter1

Introduction

1.1 Background

Over the years, use of digital images has significantly increased due to steady growth in availability of high end, yet cost effective, imaging devices. Subsequently, facial image analysis has evolved as an extensively researched multidisciplinary field with substantial work leading to exceptional outcomes and continued progress. Facial appearance is the first visible feature perceived by the human eye and is a prime source of information, vital to multiple aspects of our everyday life [1]. Making judgements based on analysis of an individual's facial appearance comes naturally to humans. Meanwhile, building methods that enable machines to imitate similar skills is a challenging task. Nevertheless, with the rapid boom of technology, numerous automated methods have emerged over several years. Facial cues are crucial to performing various analysis including age estimation [2-3], gender prediction [2], face recognition [4], blink detection [5] and concentration level estimation [6]. In the medical field, cues drawn from facial images are often used for pain assessment [7], psychological analysis [8], emotion estimation [9], monitoring mental health [10], and more. Entertainment [11] and ecommerce [12] are two other areas where facial image analysis finds application. Such diverse applications rightly demonstrate the usefulness as well as significance of performing facial expression analysis.

Facial expressions (FE) are a result of facial muscle movements, stimulated by facial nerves. These muscles surround facial components like ears, nose, eyes, and mouth, and also span across the neck area and skull [13]. This unique placement and association of facial muscles with other facial components facilitate muscle movements leading to formation of expressions perceived as smile, surprise, angry etc., as they appear on the human face. Extensive research by [14] suggested that three components of effective communication are verbal, facial gestures, and vocal, with 7%, 55% and 38% contribution respectively. Tone, intonation, and pause are some of the attributes constituting vocal components whereas messages or words spoken constitute verbal components. Evidently facial cues seem to be a major contributor for enriched communication and, therefore, remains one of the most actively pursued research fields within facial image analysis and is commonly known as facial expression recognition (FER). Muscle changes that appear on a face due to an individual's current emotional state are commonly referred as FE, thus are often seen as one of the strongest emotional indicators. Applying computer vision (CV) techniques to analyse such facial expressions has fuelled numerous smarter inventions promoting quality social information exchange. For instance, by reading a user's emotion a music application can tailor the playlist [15]. Estimating a student's engagement level through FE is another instance highlighting its usage [6,16]. Another breakthrough in FER research is the development of advanced techniques where analysis is no longer limited to still images, rather 3D image, video, varying pose, dynamic image etc., are employed. Since FE has a dominant role in conveying an effective message, its absence or limited availability in real life will make the process of drawing fundamental cues more challenging. Real life scenarios where individuals exhibit reduced FE are commonly associated with schizophrenia [17], depression [18], brain injury [19], autism [20], partially conscious patients, partial face paralysis [21] to name a few. Consequently, research involving images containing reduced FE to draw cues that aid decision making in the medical field is significantly growing. Developing assistive applications that can automatically analyse FE can significantly reduce the overall decision-making process, scale down workload for health professionals, and speed up the diagnosis/detection phase which ultimately accelerates access to medical aid for patients [22]. The research scope of this thesis is within a particular category of FE known as micro expression (ME). Since manual methods for ME recognition involve extensive training and yet are unable to yield satisfactory recognition accuracy, a shift towards automated systems have risen considerably [23]. Availability of better imaging devices, processors, and machine/deep learning (DL) techniques have boosted the computational power, thereby favouring this shift. In addition, contactless data acquisition methods using better quality equipment has led to an increase in the number of publicly available ME databases. It is notable that the availability of these databases supports the incremental increase in interest for ME research and analysis. Today, several works exist to confirm the significant contributions made in the field of ME solely by exploiting ME data contained within these databases. Further, during creation of these databases, ME were acquired in either controlled or in the wild scenarios, therefore both categories of ME databases are available for research. At present ME analysis can be performed by training algorithms or trained individuals on two ME data formats i.e., image or video. Evidently, automated ME recognition systems have achieved significantly improved results [23] over manual methods.

Therefore, in line with the current trend, this thesis focuses on the problem of automatic facial ME recognition, classification and analysis. Accurately classifying such expressions from image/video sequences to draw useful interpretations will comprise a significant segment of this thesis.

1.2 Importance of Micro Expression Recognition in Autism

CV seeks to design methods to acquire information from regions of interest within an image and assist computers to derive useful analysis. Use of CV techniques to automate medical examination processes has been steadily growing over the past decade. A systematic review by [24] suggested that analysis drawn by employing CV techniques can provide useful information that can help further autism-based research. The main motive behind introducing CV techniques in autism-based research is to overcome existing limitations that occur due to manual processes. These limitations include higher costs for medical examination, time-consuming processes, frequent clinical visits, scalability issues, and human errors to name a few [22]. In [25] it was suggested that almost 30 different medical conditions could be pre-diagnosed by automatically detecting the appropriate symptoms using computer vision. Simultaneous growth in application of CV techniques for autism spectrum disorder analysis is also noticed in recent years. The word autism was derived by a Swiss psychiatrist, Paul Eugen Bleuler, from a Greek word “autos” which means self [26]. He used this term initially to define certain characteristics of schizophrenia around 1912. Later it was used by Leo Kanner in 1943 to describe symptoms of autism that are commonly accepted nowadays [27]. Autism spectrum disorder commonly abbreviated as ASD is said to affect a child right from early childhood to adulthood and beyond. It is a developmental disorder particularly in the central nervous system hence often referred to as a neurodevelopmental disorder. Some of the early developmental deficiencies are usually visible in terms of non-verbal interactions and social behaviours. Some distinctive social behaviours include restricted body gestures, limited eye contact, reduced facial expressions and reduced range of overall activities [28]. A systematic review [24] pointed out the usefulness of CV based analysis for autism. In order to examine the expression production between ASD and typically developed (TD) children, [29] utilized machine learning (ML) and CV techniques to estimate facial action unit (AU) intensities. The results indicated that the proposed automated method was successful in analysing the expression production. In [20], videos of ASD adults were recorded, and facial analysis was performed using software named FaceReader 6.1, by estimating expression intensity for neutral and six other categories of facial expression. In order to assess attention and atypical behaviours in infants with ASD, [30] employed a CV method where a tablet was used to present the video stimuli and its camera was used to record a video of the infant. To test the feasibility of the CV approach for analysing emotion in children with ASD, [31] utilized automated FE analysis. Using a web camera, the recordings of the children viewing stimuli videos were taken. The faces acquired thereafter were used for expression examination. Through experiments it was determined that CV based methods could reliably capture atypical attention usually associated with infants having ASD. Therefore, these works indicate the success and suitability of CV based methods for automatic analysis in autism-based research.

When a child undergoes an ASD screening process, the first step involves analysis of expressions displayed on the face [32]. Through experimental observations, [33] found that individuals with ASD smiled less during conversation. In work by [34], while examining children with ASD it was established that facial gestures exhibited by them were less complex. Observations also suggested that regions located near the eyes were significantly different when compared with children without ASD. Moreover, the ability to exhibit negative emotions like sadness and disgust varied significantly between ASD and TD children. In an exhaustive meta-analysis by [35], some characteristic observations made in regard to facial expressions exhibited by children with ASD were (a) they last for a brief period only, (b) limited display of expressions, (c) lower frequency, (d) lower quality. Another meta-analysis stated that facial expressions exhibited by people with ASD are not as explicit enough as compared to those in a TD individual [36]. Consequently, it is understood that individuals with ASD potentially either lack or have very little capability to be expressive through facial expressions. The work in [36] once again pointed out that those facial expressions in ASD have a short duration whereas [37] concluded that such faces tend to have minimal facial muscle movements. Another survey presented in [38] further affirms the advantage of utilizing CV for healthcare. Such facial characteristics common to individuals with autism resembles ME therefore, this thesis considers examining ME in TD adults closely using computer vision techniques which can be used as a reference point for making comparisons and useful analysis about ME in individuals with ASD in the future.

1.3 Motivation

Several human attributes (e.g., body movements, posture, heart rate etc.) exist that can be used to derive emotion related information during a social interaction. Nonetheless, communication is an essential component for information exchange during social interaction and since FE are a substantial contributor, analysing them to make useful inferences seems feasible and more relevant. However, ASD in an individual seems to affect one's ability to have effective communication. Reduced ability to express facial gestures is one characteristic very common in an individual with ASD. Therefore, gathering information about the current state of mind using facial expressions seems challenging.

Due to high prominence and ability to be manipulated, acquiring emotional state using macro expressions may not be reliable during the decision-making process. On the other hand, due to feeble intensity, spotting ME on one's face is laborious. However, owing to its genuineness analysing ME to derive relevant information seems more valid especially during high stake situations. Continued advances in facial ME clearly demonstrate its ability to be utilized in a variety of situations to solve interdisciplinary human centred problem. Therefore, this thesis considers analysing expression by employing facial ME in TD adults.

Emerging superior video acquisition methods, as well as CV algorithms, promises development of a more effective and powerful automated facial micro expression recognition (MER) workflow. As demonstrated by extensive research, a human's ability to spot and decipher ME is far behind state-of -the-art results achieved using automated approaches even with trained human experts. Besides, manually keeping track of the types of ME identified during examination can be extremely tedious. Therefore, using an automated MER system can increasingly offer better processing capability, suitable for real world applications employing ML/DL techniques. Moreover, little research exists that has explored the impact of image resolution during MER and analysis, as such this thesis presents a novel pipeline for a MER system to address this point. Though ME analysis is now a well-established research field with notable success, its contributions still fall short in comparison to normal FE, providing abundant room for continued progress.

To summarize, the following points highlight the motivation behind development of an automated MER pipeline in this thesis.

1. The ME appears when expression leaks over the face while an individual attempts to hide one's feelings and are difficult to manipulate [39]. Due to this characteristic, information extracted from these expression finds its application for solving crime and addressing security issues [40].
2. With the ability to provide cues to identify and monitor a number of health concerns

like autism, depression, facial paralysis etc., MER can significantly supplement decision making process in the health sector.

3. In a more realistic outlook, authors in [41] rightly pointed out that ME acquired in real world scenarios are often prone to unfavourable settings. Consequently, it is expected that such data will be of much inferior quality which raises the need to have suitable approaches to utilize it. Taking this notion forward this thesis examines lower quality ME data (degraded and low resolution) in addition to usual HR data.
4. With evidence from previous research implying better recognition accuracy achieved using automated methods [23], this thesis follows a similar trend and incorporates ML/DL techniques to mobilize interdisciplinary research.

1.4 Problem Statement

Several feature extraction methods have been used to describe ME patterns in the spatial as well as temporal domain, and the approaches to recognize and classify them employ machine, as well as DL techniques. The core problem faced while analysing ME is how to extract the changes that are extremely minute in nature. While performing such tasks one often stumbles upon certain obstacles like low intensity movements, imbalanced distribution of data per class, non-uniform class labels among the available databases, varying frame rates of the captured data within a database as well as across different databases, poor inter-class discriminative features, and limited availability of databases. Further, none of the databases take real world situations into consideration due to which there is an absence of low-resolution ME databases.

1.5 Aim and Objectives

The aim of this thesis is to utilise feature extraction techniques and evaluate their suitability for ME, and also apply ML algorithms on both high resolution (HR) as well as low resolution (LR) images to develop a novel automated ME classification pipeline. To achieve this, the thesis considers the objectives as outlined below:

1. To investigate the efficiency of recognizing different classes of ME from the available databases, employing features from the temporal domain and training models using ML algorithms.
2. To explore the advantage of utilizing video magnification with temporal interpolation and phase quantization technique and build a MER pipeline by conducting suitable experiments.
3. To explore and compare two feature extraction approaches i.e., LPQ-TOP and LBP-TOP for describing the ME pattern.
4. To conduct research using LR micro expression images for expression recognition and investigate the impact of introducing super resolution techniques into MER pipeline.
5. To evaluate the performance of the proposed pipeline mentioned above (2 & 4) on several standard databases.

1.6 Contributions

The research undertaken in this thesis makes contributions towards the development of an automatic MER system. The main contributions are outlined below:

1. The ME frames are unified using a temporal interpolation model (TIM), its texture patterns are described using local phase quantization on three orthogonal planes (LPQ-

TOP), and tested in a non-cross database environment i.e., where images employed for the testing and classification processes belong to the same database.

2. A novel MER pipeline employing Eulerian video magnification (EVM), TIM and LPQ-TOP is proposed.
3. The proposed methods are evaluated on seven different ME databases i.e., SMIC-HS, SMIC-VIS, SMIC-NIR, CASME, CAS(ME)^2, CASMEII and SAMM.
4. To deal with low resolution ME images, another novel pipeline is proposed which exploits DL and a generative adversarial network (GAN) to reconstruct super resolution ME images.
5. The proposed pipeline is evaluated on three different databases i.e., CASMEII, SMIC-HS and SMIC-VIS.

1.7 Thesis Structure

The thesis consists of seven chapters of which the introductory work presented in the current chapter structurally appears as the first. The overall structure of the remainder of this thesis is organised as follows:

- Chapter 2 presents a review of the literature relating to MER techniques. It introduces the concept of ME along with the specifications of various ME databases used in research. It includes a brief explanation of key components of MER systems with reference to the experiments performed in this thesis. Challenges faced when dealing with low quality ME images and existing solutions are also discussed. This is followed by discussion on some notable applications of ME highlighting its influence over cross-discipline research areas.

- Chapter 3 highlights the importance of resolutions for ME analysis followed by the contribution of DL for super resolution. Thereafter it provides technical information on the various techniques explored for performing image super resolution. This includes algorithms utilizing residual dense networks, generative adversarial networks and bicubic interpolation. It also provides a short description on how image degradation can be achieved. All theories and methods described in this chapter provide the foundation on which the contribution Chapter 6 will be based upon.

- Chapter 4 explores the working principles of LPQ-TOP feature extraction algorithm and investigates its use for describing micro expression patterns on the spatial and temporal domains. Further, the classification of the images is realised by using SVM. The chapter explores the suitability of LPQ-TOP as a micro expression feature extraction method along with TIM, and realises the first contribution listed in Section 1.6.

- Chapter 5 addresses the limitation of discriminative features inherent with micro expressions by introducing video magnification. It explores the muscle amplification process before initiating feature extraction on all seven databases using the LPQ-TOP method. Further it examines the combined contribution of magnification, interpolation, and LPQ-TOP to achieve improved classification. This chapter realises second, and third contributions listed in Section 1.6.

- Chapter 6 presents the newly developed pipeline to perform micro expression classification using low quality images. To mimic low quality images, 64x64 and 32x32 sized images have been considered throughout the experiments. Five different super resolution methods are exploited individually to construct super resolution images and are then utilized for recognition. An exhaustive performance analysis of the proposed pipeline is presented over three different ME databases. This chapter realises the last two contributions listed in Section 1.6.

- Chapter 7 concludes this thesis by providing a summary of contributions, the limitations of the research presented in the thesis in the field of micro expression recognition along with direction and suggestions for future research.

Chapter 2

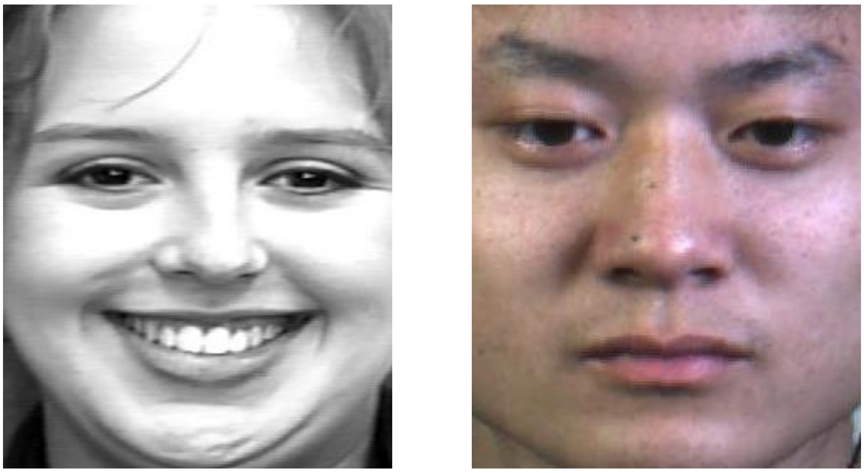
Facial Micro Expression: A Literature Review

2.1 Introduction

For the past few decades there is a growing interest in understanding human emotions and feelings, particularly for measuring health and well-being; communication is considered as a distinguished and competent tool for doing so. Classically, communication is divided into two categories, verbal and non-verbal. The ability to identify and draw appropriate implication from non-verbal cues is one of the most challenging tasks demanding extensive research from various disciplines, particularly from social science, medical science, psychology, and technological sciences, for more than two decades. In addition, some of the common non-verbal cues that have attracted interest from research scholars of diverse backgrounds include body language, facial expressions, facial emotion, eye movements, eye contact, hand gestures etc. Analysing such cues are imperative since they are sourced directly from the emotional brain [42] and are manifested largely as most authentic expressions. As rightly stated by the proverb “face is the index of mind”- FE are major contributors in decoding facial emotion thereby facilitating health professionals in deriving one’s psychological status. Conventional facial movements that occur due to contraction of facial muscles in certain order result in exhibition of expressions on faces. Lines, wrinkles, and folds are by-products of such muscle contraction which alters the positions of facial landmarks. It is universally accepted that almost all varieties of FE generally involve brow movements [43]. These movements could construe either raising or lowering brows and have high visibility. Brow lowering is usually associated with negative emotions identified as sadness, fear, anger etc. On the contrary, brow raising may imply expression of positive emotions like happiness, surprise etc.

Research has shown that FE are effective in comprehending a rich source of information essential for rendering facial emotion and are broadly categorized into macro and micro expressions. Both these expressions are highly informative non-verbal cues in facial emotion analysis and hence examining such expressions has gained immense popularity over several decades. Facial macro expressions are typically very prominent and can be determined very easily since it lasts > 0.5 second and <4 seconds [44]. Extensive research with high end results has been successfully achieved for macro expressions and the research continues. Sometimes it is debated that macro expressions might not always be the best measure for decoding a person’s real emotion and psychological status mainly due to its voluntary and conscious characteristics. Consequently, a person displaying such expressions is aware of it and has full control over them, therefore, can potentially display manipulated expression, diverging from their real emotion. In such circumstances these facial macro expressions cannot be a measure for determining legitimate cue.

On the other hand, research work on ME appeared much later and is still evolving. Though research on this field has increased in recent years yet, it still remains an area where much work is needed. ME are believed to expose legitimate emotion since they are natural, authentic, genuine, and honest expressions. Such ME appear on a person’s face in an involuntary manner; hence a person has minimal control over them. It is believed that in an



(a) (b)

Figure 2.1. (a) Happy macro expression (CK database) [46], (b) Happy micro expression (CASMEII database) [47].

attempt to stifle one's true emotion consciously or unconsciously, sometimes these expressions leak on one's face [39]. These are more natural and genuine than macro expressions and are thus perceived to be useful cues. Few characteristics that distinguish these expressions from one another include duration, facial segments forming these expressions and intensity. However, it has been claimed that duration is the exclusive characteristic that sets these expressions apart, dismissing intensity as a potential classification factor [45]. The work in [45] also notes that expressions with life span stretching beyond ME duration eventually qualify as macro expressions. It is a well-known fact that ME is signified by rapid muscle movement which lasts only for a short span of time, ideally for less than half a second. These works have also revealed that ME have comparably very low intensity than macro expressions which makes recognizing such expressions more arduous. Some of the terms that have been commonly used to describe the nature of ME include short duration, low intensity, faint, rapid movements, subtle, split-second movement, fleeting expression etc., [39] [47]. Almost 80% of the facial components contribute to forming a macro expression whereas a significantly smaller number of facial segments take part in forming a ME. Carefully observing Figure 2.1(a) high prominence and clear visibility of macro expression is evident, whereas subtle and minuteness of ME which seem almost undetectable are evident in Figure 2.1(b). The formation of lines around the mouth and nose region due to facial activity is clearly visible for the macro expression but is extremely faint for ME. These ME have a wide range of applications in the real world like criminal investigation, autism, psychology, schizophrenia, lie detection, business negotiations and mental diseases [48][49]. Some popular applications of ME will be discussed in later sections.

MER systems can be devised using both manual as well as automated methods. For recognizing ME manually, observations are made by highly trained individuals and has acquired an accuracy of only 47% [50]. In [51] a psychological experiment was conducted and an average recognition rate of 50% was achieved. The Micro Expression Training Tool (METT) is an instance of a tool developed for manual ME detection by Ekman [52]. Such manual approaches are often tedious and time consuming therefore, [53] designed an automatic MER system and tested it using the METT videos. From the performance obtained, the automated approach clearly worked better in comparison to the trained human approach. Moreover, with the availability of a number of ME databases, attention from computer science researchers has increased significantly. Hence CV methods are being successfully explored for ME and are continuously improving recognition performance. This accelerating success of automated approaches is reflected by accuracies as high as 75.3% achieved for MER [49] compared with accuracy not exceeding beyond 50% obtained so far for manual approaches. Such results endorse the fact that by using automated MER techniques, we can achieve superior performance compared with manual approaches. These positive outcomes have contributed to automated approaches gaining more prominence than manual approaches for identifying such fleeting expressions.

The focus of this thesis is designing an effective automated MER system utilizing ML and DL algorithms. By building such a MER system we intend to provide groundwork which can be extended and utilized for future applications like estimating emotion of individuals with autism using these ME.

This chapter will provide a systematic overview on ME along with discussion on several topics related to this field of research. Section 2.2 introduces ME, then, in Section 2.3 a short description of different databases used for evaluation of ME analysis algorithms is provided. This is followed by Section 2.4 which outlines core components of an automatic system that recognizes ME. Taking a real-life scenario into account Section 2.5 discusses recognizing ME from images with low resolution. Areas where MER systems can be applied are discussed in Section 2.6, with final concluding remarks presented in Section 2.7.

2.2 Micro Expression

According to [54] FE are a constant negotiation between two neurological pathways, pyramidal and extrapyramidal tract, which are sourced from two different sections of the brain. Facial movements that are voluntary in nature are caused by the pyramidal tract whereas involuntary ones are the result of the extrapyramidal tract. In a high-stake situation when a person tries to control their expressions, both these tracts get activated which creates a neural conflict leading to a quick leakage of expressions, called ME [54]. ME were first spotted by Haggard and Isaacs while reviewing motion pictures captured for psychotherapy sessions in 1966 [55]. The concept was further expanded by Ekman with his subsequent works [39][43]. Through these works, ME were believed to be indicators of emotional state. Research suggests that the discriminative characteristics of facial ME are its exceptionally short duration lasting approximately between 0.04 to 0.2 seconds [45] and remarkably low intensity muscle movements [39-45]. The highest acceptable duration limit is 0.5 seconds [47]. Since facial muscle contractions last for a brief amount of time these facial movements are perceived to be almost invisible to the naked eye and are likely to be missed using traditional observation methods. The inherently imperceptible characteristics makes recognizing ME considerably challenging and effortful. Based on muscle contraction patterns, expressions are generally categorized as happy, sad, fear, surprise, disgust etc. [56]

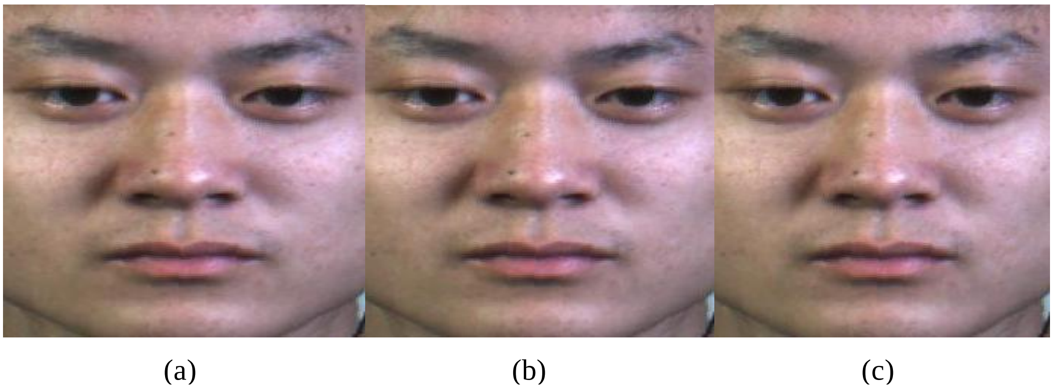


Figure 2.2. Three stages of micro expression: (a) onset, (b) apex (c) offset [47].

The lifetime of a ME is generally divided into three stages commonly termed as onset, apex and offset [45] (see Figure 2.2). These three stages are based on the strength of intensity for a given expression which acts as markers for choosing the correct frames during spotting. Facial Action Coding System (FACS) [57] plays a crucial role in building a bond between the two entities: muscle changes and emotional states. This system was crafted to point out the precise time of occurrence and ending of an AU. During the initial stage when an AU is first perceptible and has the lowest intensity of facial motion, it is termed as onset stage (see Figure 2.2 (a)). In this stage the muscle starts contracting with stronger changes in various facial regions along with increased visibility. As they progress the AU becomes fully visible at its peak and the intensity is at the highest, this stage is known as apex (see Figure 2.2 (b)) and is visibly the most expressive frame [45]. Further, as facial muscles start relaxing and AU disappears, it characterizes the offset stage (see Figure 2.2 (c)), with no intensity of facial motion [45]. This absence of motion intensity on facial muscles brings the face back to its neutral state. Thus, this sequential shift of facial muscle motion is in general a progression from neutral to onset, then to apex, followed by offset and, back to neutral state.

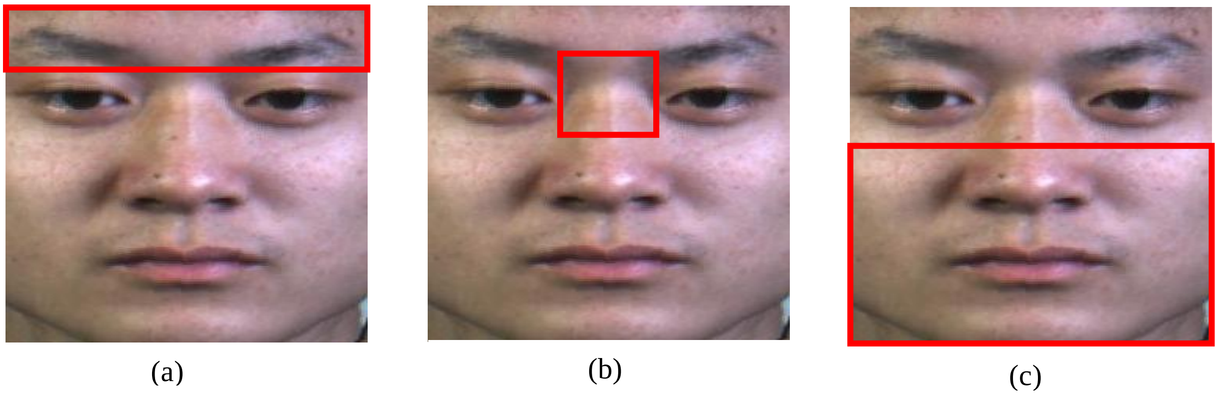


Figure 2.3. Three facial regions where expressions are generally visible [57]: (a) brow-forehead, (b) eyes-nose bridge (c) lower face.

Appearance of facial expressions is generally visible on three facial regions identified as brow-forehead, eyes-nose bridge, and lower face as illustrated in Figure 2.3. The lower face includes regions identified as cheek, nose, mouth, chin, and jaw. The presence of a particular FE can be established by examining the arrangement of features, degree of tension or relaxation, and existence, or absence of wrinkles. It is believed that any change in one of the facial regions will alter the appearance of other facial components present in that particular section. However, any movement in one of the facial regions may not always result in noticeable variation in terms of appearance in the other two regions. Considering universal expressions convention [57], a happy expression is generally identified by the appearance of raised lip corners, raised cheeks, appearance of wrinkles around the eyes along with tight muscles around them. Following the same convention, sadness is identified by raised inner corners of eyebrows, both lip corners drawn down, and muscles of eyelids loosened. A dilated pupil with open mouth, eyebrow, and eyelids both pulled up signifies a surprise expression. When the eyebrows, upper and lower eyelids all seemed pulled down with tightened lips, it is identified as an anger expression. A stretched mouth with eyebrows and eyelids pulled up determines muscle patterns for fear. Wrinkled nose with loose lips along with pulled up upper lip and pulled down eyebrow patterns implies presence of disgust expression. These muscle patterns for a variety of expressions are applicable to ME as well; however, due to feeble motion intensity as well as extremely short span, the visibility of such patterns may be almost imperceptible.

Though researchers have been working dedicatedly to improve the performance of recognition systems for these miniscule expressions, it is still behind results achieved for macro expression and provides ample scope for research and improvement. The two fundamental reasons behind this deficiency have been rightly pointed out as inadequate number of databases as well as demanding characteristics of ME itself [23,58]. However, recent work indicates both drawbacks have been addressed to some extent by various works utilizing different approaches which shall be discussed in Section 2.4.

2.3 Spontaneous Facial Micro Expression Datasets

To test and evaluate ME algorithms and carry out a meaningful comparison, standardised databases are essential. Deriving strengths and weaknesses of algorithms on such standardised databases helps to achieve fair and acceptable contributions. Most of the existing state-of-the-art MER systems have experimented on HR images to achieve high end results. This thesis has explored the area of ME with both high as well as LR image data. Since ME databases contain frontal views of the face, not much effort has been put into developing algorithms that cater to various face orientations for ME. To ease the task of recognition, ME databases contain image or video sequences recorded over a neutral background with good lighting conditions. For manual ME detection, a tool popularly known as METT was developed by [52]. Subsequently in later years databases were categorized into posed and spontaneous. The ME collected from those participants who deliberately produce the required expression or are asked to mimic a given expression constitutes a posed ME database. Polikovsky's database (PD) [59] and unsupervised segmentation fusion-high definition (USF-HD) [60] are two such posed ME databases.

On the other hand, ME gathered from participants without any prior knowledge of the type of expressions to be displayed for a given set of stimuli constitutes a spontaneous database. Spontaneous ME databases that are generally employed with automated approaches include:

- York Deception Detection Test (York DDT) [61],
- Chinese Academy of Sciences Micro-expression (CASME [62],
- CAS(ME)^2 [63],
- CASMEII [47]);
- Spontaneous Micro-expression database (SMIC-HS, SMIC-VIS, SMIC-NIR) [64] and
- Spontaneous Micro-Facial Movement (SAMM) Dataset [65-66].

Apart from these, "in the wild" ME database comprising of poker game videos downloaded from YouTube named as MEVIEW is also available [67]. Recently a new database known as micro and macro expression warehouse (MMEW) has been developed by [68]. In this thesis databases employed include SMIC-HS, SMIC-VIS, SMIC-NIR, CASME, CAS(ME)^2, CASMEII and SAMM at various stages. ME can be classified into universally categorized emotion classes namely, happiness, surprise, contempt, disgust, sadness, fear, and anger. Capturing and eliciting such ME is extremely difficult, consequently existing datasets do not have a complete set of all these classes. Some MEs like happy can easily be elicited, while some expressions like fear are difficult, this has resulted in uneven distribution of the data collected. Class labels provided for various ME databases are not uniform either. For instance, SMIC database has only three categories i.e., positive, negative and surprise whereas CASME database has seven categories of expressions. Detailed information about these databases relevant to this thesis has been listed in Table 2.1. along with a sample image from each of these databases in Figure 2.4.

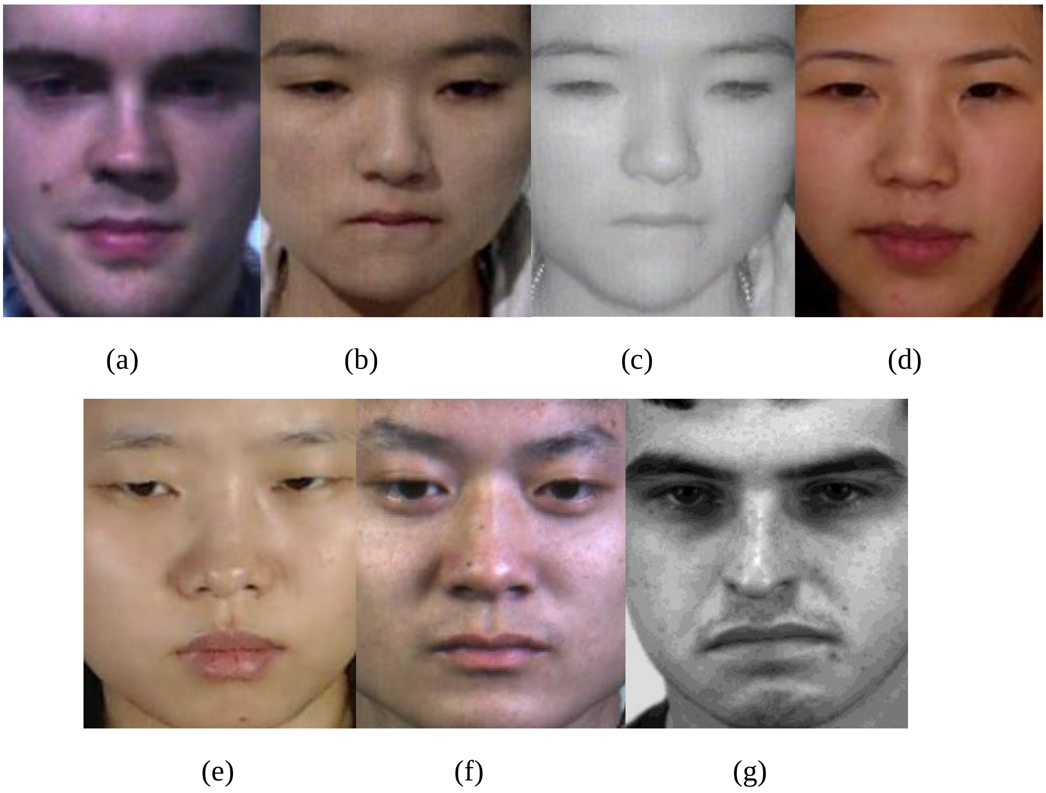


Figure 2.4. Sample micro expression images from various datasets. (a) SMIC-HS (b) SMIC-VIS (c) SMIC-NIR (d) CASME (e) CAS(ME)^2 (f) CASMEII (g) SAMM

Table 2.1. Spontaneous micro expression database summary.

Database	Total participants	Total ethnic groups	Total data	Total emotion class	FPS	Facial resolution	Label names & distribution	
							Label	Distribution
SMIC-HS [64]	16	3	164	3	100	190 x 300	Positive	51
							Negative	70
							Surprise	43
SMIC-VIS [64]	08	3	71	3	25	130x160	Positive	28
							Negative	23
							Surprise	20
SMIC-NIR [64]	08	3	71	3	25	190 x 230	Positive	28
							Negative	23
							Surprise	20
CASME [62]	35	1	195	8	60	150 x 190	Surprise	20
							Happiness	9
							Disgust	88
							Fear	2
							Sadness	6
							Tense	28
							Contempt	2
							Repression	40
CAS(ME)^2 [63]	22	1	53	4	30	NA	Positive	6
							Negative	19
							Surprise	9
							Others	19
CASMEII [47]	26	1	246	5	200	280 x 340	Happiness	32
							Surprise	25
							Disgust	63
							Repression	27
							Others	99
SAMM [65-66]	32	13	133	7	200	400 x 400	Happiness	26
							Disgust	9
							Surprise	15
							Fear	8
							Contempt	12
							Anger	57
							Sadness	6

2.4 Micro Expression Recognition System

MER is a research area which deals with classifying human facial emotions through extremely minute and fine expressions on one's face. Given its significant role for facial emotion estimation and analysis, automating the MER would be an extremely beneficial step. This thesis investigates the advancement of technologies for MER and develops an automated system by employing CV with the possibility of extending it for real world applications. As highlighted in Section 2.1, the manual method for MER is very laborious since it involves rigorous individual training. For manually recognizing such expressions using videos, experts need to carefully inspect the video, frame by frame, often pausing in between to ensure these expressions are not missed. During such inspections if any interruptions occur then the viewing may have to be restarted which makes this method very tedious and laborious. This process has achieved much less accuracy so far and hence successful automated MER systems are still being sought, thus the thesis focus remains on building a successful automated approach. Some early works that attempted to automatically identify spontaneous ME include [64][69-70]. In almost all these works, focus was laid on automatically recognizing ME using images or videos picturing frontal-view of faces. These expressions were collected from various participants who displayed expressions in response to a particular stimulus. As highlighted earlier, ME can be posed or spontaneous, however the work of this thesis focuses on spontaneous ME.

This section aims to provide a comprehensive literature review of recent methods in the field of ME and focuses exclusively on its automatic recognition. A general approach to such an automated framework for spontaneous MER usually consists of three primary stages namely, pre-processing, micro facial feature extraction and feature classification (Figure 2.5). For a better understanding, each of these components are systematically reviewed in this section.

Additionally, the existing spontaneous ME databases employed in this work has already been summarised in Section 2.3.

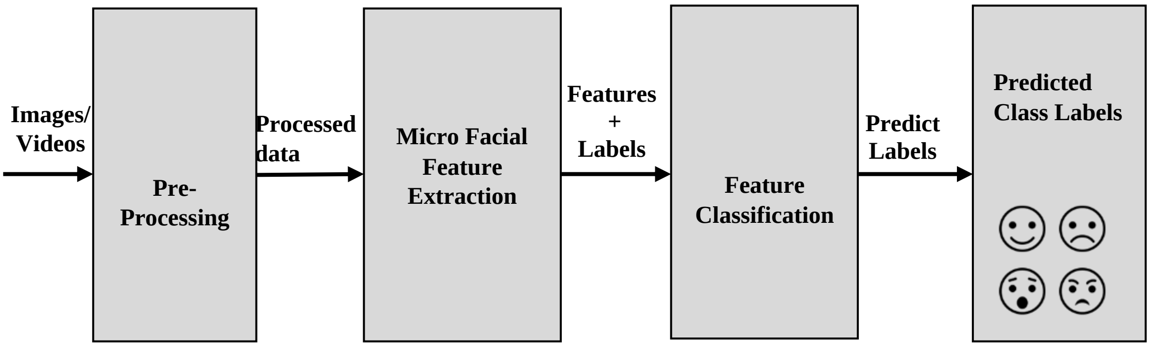


Figure 2.5. A basic framework for microexpression recognition [23].

2.4.1 Pre-processing

Image quality is generally affected by several factors while undergoing the acquisition process, for example, pose, illumination etc. This has an adverse impact on recognition accuracy and requires attention. To neutralise such influence, it is essential to perform pre-processing on the raw images. This pre-processing stage is the initial stage of the recognition process that primarily focuses on data preparation. Here, the available databases are processed to generate improved data suitable for recognition purposes. Such improvements are commonly achieved by eliminating or suppressing redundant and unwanted attributes of the input data. Different pre-processing methodologies relevant to MER include face detection, face alignment and segmentation, frame normalization, motion magnification and data augmentation. Face detection and tracking is an active research field with much literature available and is beyond the scope of this

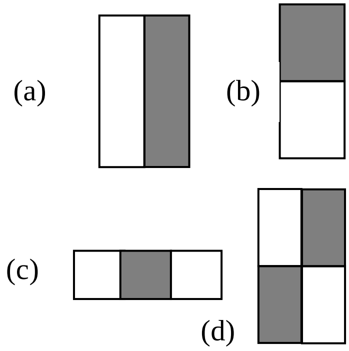


Figure 2.6. Illustration of Haar features [71].

thesis. Hence, in this section we discuss different methodologies of the pre-processing step which are popularly employed for ME analysis and highlight their contributions.

The most widely used face detector for face analysis is the Viola and Jones (V & J) detector [71]. It processes input in the form of image or image sequences to automatically find the face region by employing Haar features. The edge, linear, centre and diagonal are four Haar features used by the algorithm as illustrated in Figure 2.6 (a), (b), (c) and (d) respectively. The edge features are used for detecting facial patterns that appear as edges, like eyebrows which generally appear darker. Line features are used for detecting facial features that appear as lines, like nose bridge. Similarly, diagonal features are used for picking diagonal facial features like wrinkles, jaw, chin etc. For frontal-view face images with neutral background this method is sufficient for face detection. The raw images are likely to be occupied by other objects along with a face, due to which face detection becomes necessary. It helps to crop out the face and therefore becomes the only object occupying the whole image. The step-by-step procedure it follows to achieve detection is illustrated in Figure 2.7 [71].

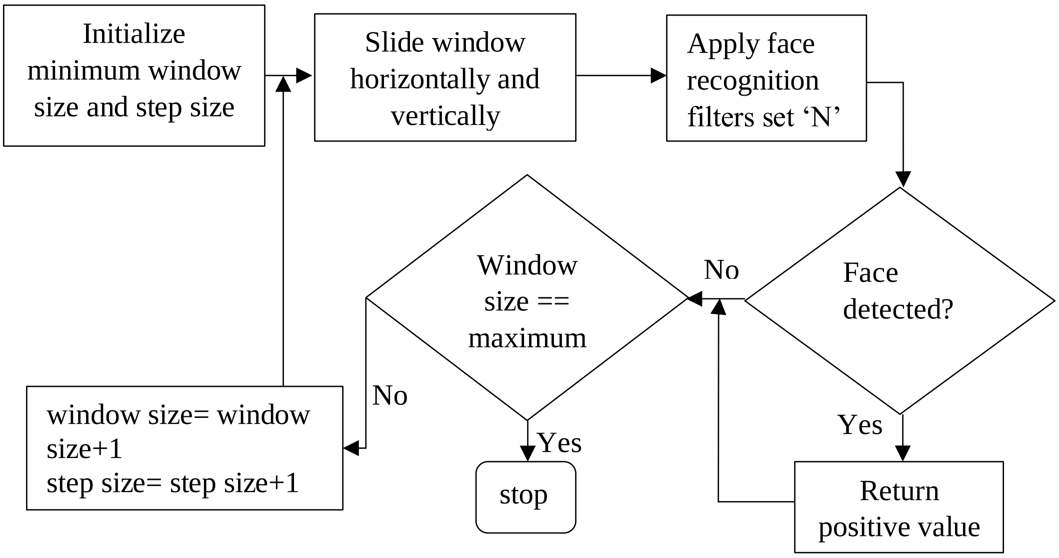

Initialize minimum window size and step size
Slide window horizontally and vertically
Apply face recognition filters set 'N'
Face detected?
No
Yes
Return positive value
Window size == maximum
No
Yes
window size= window size+1
step size= step size+1
stop

Figure 2.7. Illustration of face detection procedure [71].

For MER, inclusion of a sequence of images rather than a single image is vital. Thus, variances in pose or scale between images need to be removed to eliminate any adverse influence over recognition performance. Algorithms that are capable of aligning faces are helpful in dealing with such situations.

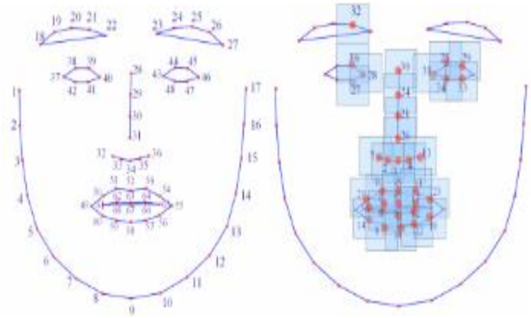
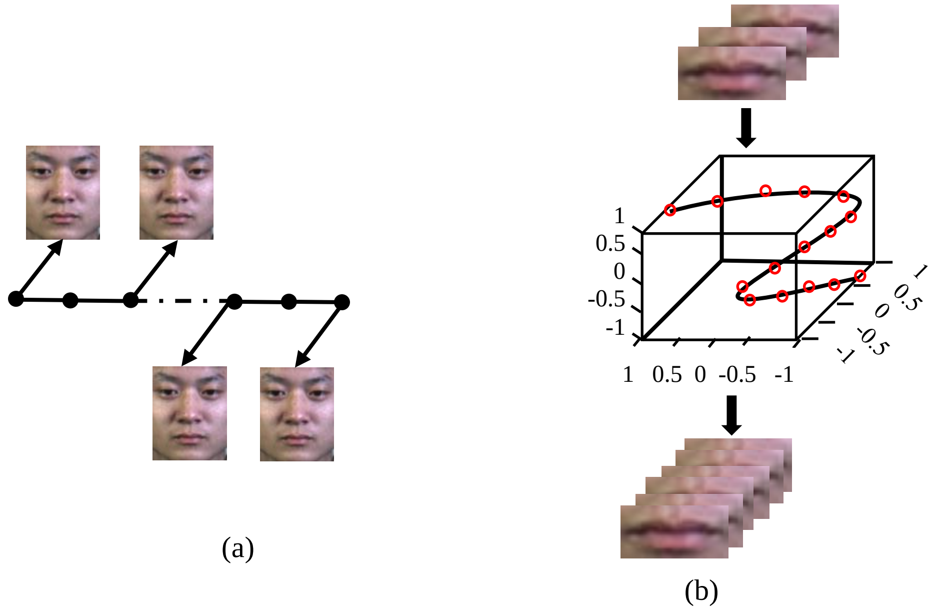


Figure 2.8. Modelling face using ASM by identifying landmarks and regions [72].

Active Shape Model (ASM) [72] is widely used as a face alignment technique which employs a group of patches to represent a facial appearance. Alternating between response map construction and shape fitting, the alignment process is fine-tuned at every iteration. An instance of face modelling using ASM with landmarks and regions is depicted in Figure 2.8.

Discriminative response map fitting (DRMF) [73] is another texture-based face alignment technique that can effectively detect 68 feature points, if given a facial region as input. By employing a template, another technique commonly known as constrained local model (CLM) [74] learns shape layout and texture variations for modelling faces. To preserve shape attributes of a face, the integral projection [75] method employs image differences while computing horizontal and vertical projections. Another popular face alignment technique is local weighted mean (LWM) [76] which has been extensively used for ME [41] [47]. Databases with ME videos contain varying frame lengths, hence, to achieve uniform alignment of frames while employing these videos, the temporal interpolation method (TIM) [77] has been widely accepted for ME analysis [70]. The method has been employed in this thesis therefore its working principle is briefed here.

The basic principle of TIM is to build a continuous function following a curve trajectory using the set of images that constitute a ME video. To achieve this, a graph representation is



(a)

(b)

Figure 2.9.(a) Illustrating an instance of micro expression with a graph representation [70].
(b) Abstract view of temporal interpolation method for mapping video against a curve [70].

generated where every instance of ME video is expressed in terms of a path graph consisting of ‘n’ vertices, denoted as P_n (see Figure 2.9). In this type of graph-based representation the edges represent adjacency matrix W whereas vertices represent video frames expressed as [70]:

$$W \in \{0,1\}^{nxn}; \qquad W_{i,j} \begin{cases} = 1, if\ |i-j| = 1 \\ \ \ 0, otherwise \end{cases} \tag{2.1}$$

To achieve generalization where all connected vertices have minimum distance between them, the graph path is mapped to a line, with minimization given by [70]:

$$\sum_{i,j} \left(y_i - y_j\right)^2 \ W_{i,j} \ ; where\ i,j = 1,2,\dots n \tag{2.2}$$

In equation (2.2) the map is denoted by $y = (y_1, y_2 \dots, y_n)^T$. It must be noted here that this minimization process is same as computing Laplacian graph constituting $\{y_1, y_2, \dots, y_{n-1}\}$ eigenvectors. Therefore, y_k can be assumed to be a set of points expressed by [70]:

$$f_k^n(t) = \sin(\pi kt + \pi(n-k)/(2n)), t \in [1/n, 1] \tag{2.3}$$

Here in equation (2.3) the points are sampled over time interval t, where $t = 1/n, 2/n, \dots, 1$

resulting in a curve expressed by [70]:

$$\mathcal{F}^n(t) = \begin{bmatrix} f_1^n(t) \\ f_2^n(t) \\ \vdots \\ f_{n-1}^n(t) \end{bmatrix} \tag{2.4}$$

Arbitrary positions within the ME images are then chosen to temporally interpolate them using this resultant curve. To establish a correlation between the image space and curve F^n, individual image frames are mapped to a set of points described by [70]:

$$\mathcal{F}^n(1/n), \mathcal{F}^n(2/n), \ldots.., \mathcal{F}^n(1) \tag{2.5}$$

In addition to this linear extension of graph, embedding is also utilized to gather knowledge about transformation vector denoted by w, which aims in minimizing the following [70]:

$$\sum_{i,j} \left(w^T x_i - w^T x_j\ \right)^{\mathbf{2}} W_{i,j}\ ; where\ i,j = 1,2,\ldots,n \tag{2.6}$$

Here in equation (2.6) x_i denotes a vector with its mean removed, given by $x_i = \xi_i - \xi'$, the vectorized image is denoted by ξ_i and ξ' denotes the mean. The resulting problem of eigen value is expressed as:

$$XLX^T\ w = \lambda' XX^T\ w \tag{2.7}$$

Utilizing singular value decomposition to further solve this problem where $X= U \sum V^T$ the new image ξ can be interpolated as described by equation (2.8) where square matrix is denoted by symbol M and ξ_i is assumed to be linearly independent:

$$\xi = UM\mathcal{F}^n\ (t) +\ \xi' \tag{2.8}$$

To summarise, the uniform frame length is achieved by interpolating frames at random locations using graph embedding technique (See Figure 2.9(a)). The interpolated frames at low dimensional space are mapped back to corresponding high dimensional space which ultimately results in a normalized sequence of frames (Figure 2.9(b)).

For dealing with the subtleties of ME, video magnification has been identified as an effective technique and this is substantiated with experiments reporting increase in recognition accuracy with its usage [40][49][78]. The technique essentially amplifies the feeble motion thereby increasing the visibility of these subtle expressions. Eulerian video magnification (EVM) was initially developed to boost the visibility of the human pulse and heartbeat rate in infants [79]. Both activities involve extremely faint movements that are almost invisible to human eyes. The method successfully intensified irregular and low magnitude movements thereby increasing their visibility. A detailed model of this magnification technique is discussed in Chapter 5. Global Lagrangian motion magnification (GLMM) is another motion-based technique explored by [80] which takes a different approach than EVM to achieve magnification. It considers global displacement between frames as opposed to local displacement considered in EVM. Experiments also demonstrate that TIM combined with video/motion magnification help to promote capabilities of existing MER systems [40][78].

To address insufficiency in availability of ME data, data augmentation has been applied for MER by various researchers. The most common practice for augmenting data involves flipping image/video data horizontally or vertically, rotating with varying angles, translating along *x* and *y* axis, using varied scaling factors etc. Using augmentation, [81] generated synthetic images and reported accuracy boost on both the CASME and CASMEII database. Similarly, to increase the volume of data, [82] exploited a generative adversarial network (GAN) and successfully generated fake ME images. It was also effective in eliminating class imbalance issue prevalent in the ME dataset. In [83] a new technique was proposed to deal with the unbalanced data issue in ME using both an AU and a GAN. The method was called an AU intensity controllable GAN (AU-IGAN). AU have intensities that vary according to the level of expressiveness, therefore by using its variants in different combinations, synthetic ME training samples were formed. For every ME category approximately 300 to 400 synthesised data were created using this technique. Moreover, every type of ME class was created in this new synthesized database for each subject. Employing these steps, a significant boost in the volume of data was achieved along with more balanced class distribution.

2.4.2 Micro Facial Feature Extraction

Primarily, the target of this phase is to obtain stable and discriminatory facial features. The quality of features extracted directly affects the performance of classifiers, hence recognition systems give much attention to methods that can extract optimal features. For effectively representing input information, feature extraction techniques generally convert pixel data into a reduced form. This helps in minimising discrepancy that may be caused by unwanted external conditions associated with the input such as motion blur, lightning condition etc. Additionally, extraction of such optimal features is critical to minimize variations within a class and simultaneously maximize inter-class variations. The goal is not only to extract desired features, but also to ensure that the scaled down data can effectively represent its source. Feature extraction methods can be broadly categorized into appearance based, geometric based or hybrid.

Appearance based approaches aim to capture texture patterns based on different appearance features such as image intensity, gradient, filter bank etc. In simple terms, it applies filter banks or image filters to extract required facial attributes. On the other hand, geometry-based approaches generally use geometric relationships of facial components defined by features such as location of facial landmarks, angle between certain points, Euclidean distance etc. Sometimes combinations of these approaches can be employed and are known as hybrid methods. Facial representation, when encoded using image sequences employing a frame-by-frame approach, is termed as a spatial based method. However, if encoding involves a temporal window for image sequences, then it is known as a spatio-temporal method. Various feature extraction techniques employed specifically for ME have emerged in recent years. To ensure minute changes of ME are successfully extracted, inclusion of temporal variations is essential. During implementation this is realized by considering a temporal window and examining the sequence of frames within it. Consequently, approaches that facilitate information extraction from spatio-temporal domains are popular among MER system, commonly identified as dynamic based approaches.

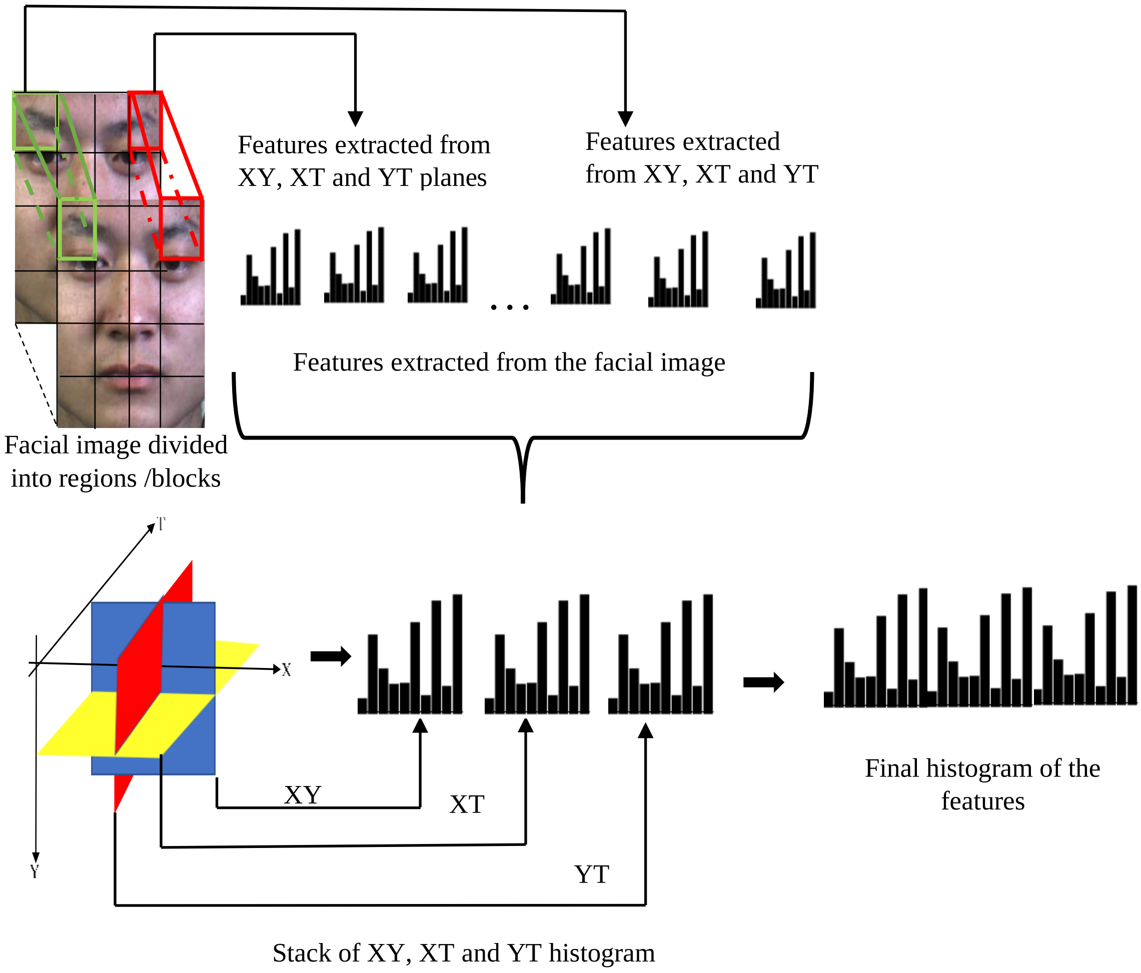

Features extracted from XY, XT and YT planes
Features extracted from XY, XT and YT
. . .
Features extracted from the facial image
Facial image divided into regions /blocks
T
X
Y
XY
XT
YT
Final histogram of the features
Stack of XY, XT and YT histogram

Figure 2.10. Description of facial image with LBP-TOP method [84].

Local binary pattern on three orthogonal planes (LBP-TOP) [84] is one such approach and remains one of the most widely employed spatio-temporal feature extraction techniques particularly for ME [47][49]. Its approach is to select a centre pixel and its neighbouring pixels, then compare their pixel values. This helps to describe texture variations within a circular region by applying binary codes. The method extracts required features from three planes then generates a histogram to represent the features (See Figure 2.10). Including planes corresponding to the time domain, helps in representing temporal variations of the subjects. This method was also used as a baseline evaluation for the CASMEII [47], SMIC [64] and SAMM [65-66] databases. The popularity of this method is due to its low computational complexity and favourable tolerance with varying illumination in the input. This method has been employed in this thesis therefore its working principle is briefed here.

The notation used to refer a Local Binary Pattern (LBP) operator with certain radius R and neighborhood sample points P is $LBP_{\mathrm{P,R}}$. For every $LBP_{\mathrm{P,R}}$, the total number of binary patterns generated is given by 2^p. Mathematically, for a center pixel, c with cordinates (x_c, y_c), with P neighbouring pixel at R radius, the LBP is computed using equation (2.9) and (2.10) [84].

$$LBP_{\mathrm{P,R}}(x_c, y_c) = \sum_{p=0}^{\mathrm{P}-1} s(\mathrm{g}_p - \mathrm{g}_\mathrm{c})2^p \tag{2.9}$$

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{2.10}$$

In equation (2.9), g_p and g_c denote grey values for the neighbour pixel and centre pixel respectively, 2^p represents the weight on the neighbouring pixel at p^{th} location where, $p = 0, \dots, \mathrm{P} - 1$ and $s(x)$ in equation (2.10) manages the sign issue. The binary pattern obtained is then arranged in either clockwise or anti-clockwise manner to obtain its corresponding decimal equivalent. Every occurrence of the LBP code generated throughout the image is then collected to build a histogram. This operation is performed on all three planes XY, XT and YT to generate three corresponding histograms. As a concluding step all three histograms are concatenated to form a final histogram.

A variation of LBP-TOP method was developed by [85] where the central pixel was used to compare with a pair of neighbouring pixels, thereby reducing the binary code length to half the size than that of LBP-TOP. The method was known as centralized binary pattern on three orthogonal planes (CBP-TOP). By employing an extreme learning machine (ELM) during classification, the combination of these two techniques was effective in improving the recognition. Subsequently a more condensed form of LBP-TOP was suggested by [86-87] which used only six unique points of three intersecting planes to represent the features known as local binary pattern with six intersection points (LBP-SIP) (see Figure 2.11). This method was able to reduce the space and time complexity by half compared with the LBP-TOP method. They also developed another variation of LBP-TOP which represented features in a super compact form by computing the average of each plane known as local binary pattern - mean orthogonal plane (LBP-MOP). This method drastically reduced the feature extraction time compared with the LBP-TOP method; specifically, it extracted features 38 times faster than the conventional LBP-TOP.

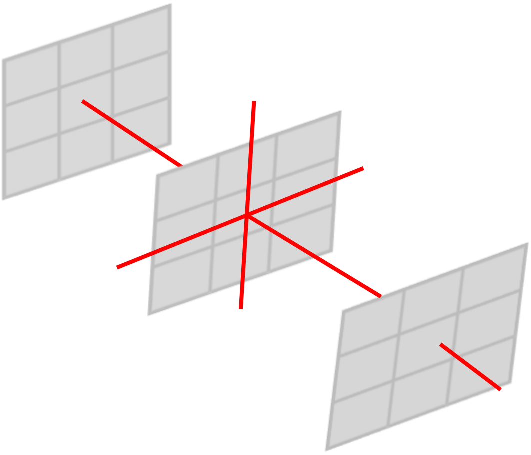


Figure 2.11. Illustration of six discrete neighbour points in LBP-SIP technique [86-87].

LPQ based methods have also been actively employed in macro expression research

[88] with good performance and now its usefulness can also be seen for ME too. An extension of this method, popularly known as LPQ-TOP was explored for ME in [89]. These works successfully demonstrate the robustness of LPQ-TOP as an extraction technique.

Another extraction technique based on integral projection with difference images known as spatio-temporal local binary pattern with integral projection (STLBP-IP) was proposed by [90]. The method successfully preserved facial images' shape attributes and was also effective in generating more discriminative facial features. To enhance the existing discriminative capability of the overall system, an improvisation of STLBP-IP was proposed by [91] which employed principal component analysis (PCA) and a feature selection technique based on the Laplacian method. By employing this approach an impressive increase in recognition rate by (approx.) 4% using the CASME II database and (approx.) 9% using the SMIC database was recorded compared with STLBP-IP. All these feature extraction techniques belong to the LBP family.

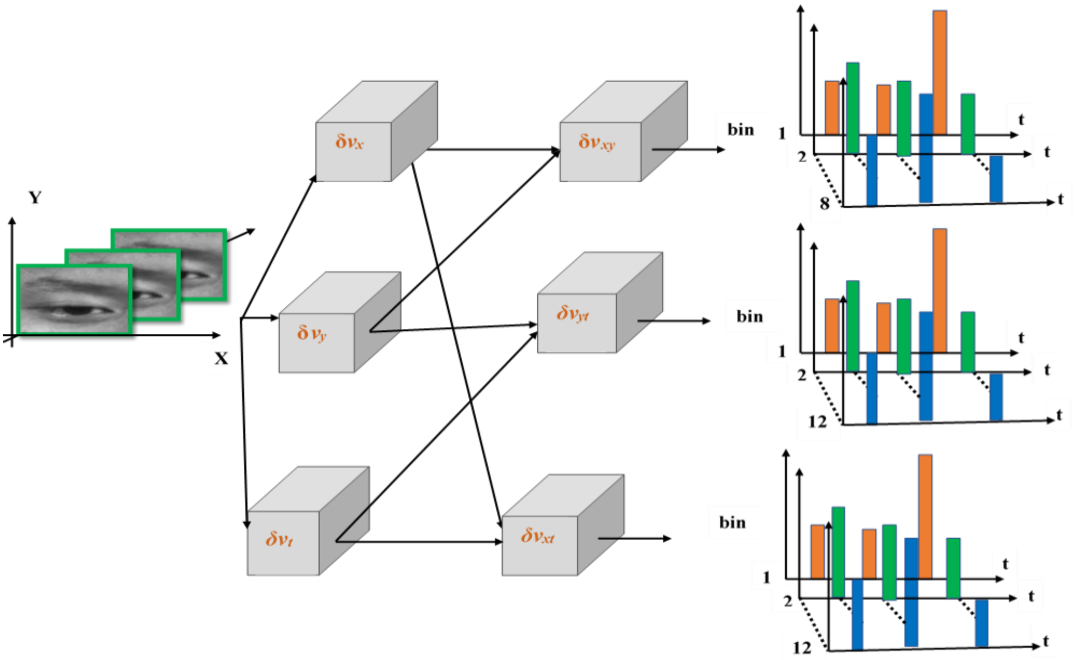


Figure 2.12. An outline for extraction process employing 3DHOG [59].

Apart from these LBP based methods, two gradient based methods that are popular for solving MER problems include histogram of oriented gradient with three orthogonal planes (HOG-TOP) and histogram of image-oriented gradient with three orthogonal planes (HIGO-TOP). Three-dimensional HOG (3DHOG) was proposed by [59] to describe the spatio-temporal structures of facial ME. By dividing the entire face into twelve regions, spatiotemporal blocks corresponding to each region was obtained. To construct a histogram for these regions the magnitude of gradient projections was computed in all three directions. An outline of this approach for extracting features is presented in Figure 2.12. In another research, the HOG technique was extended to calculate gradient magnitude and local gradient direction for each of its three planes known as histogram of image gradient oriented on three orthogonal planes (HIGO-TOP) [40]. The extraction process followed by HIGO is identical to HOG but ignores the magnitude factor. Both these methods have been extensively examined in [40] along with video magnification. The HIGO method augmented with magnification was able to achieve a remarkable recognition rate of 78.14% using the CASME II dataset. Besides LBP and gradient methods, optical flow-based approaches have also been widely examined for ME.

The bi-weighted oriented optical flow (Bi-WOOF) approach was proposed in [45] to extract optical features of ME from a single apex frame. An apex frame is believed to possess the most discriminative features in comparison to other frames hence this work performs recognition based on features extracted from a single frame only. The method was tested on the CASME II and SMIC datasets and produced results comparable with other methods with accuracies of 61% and 62% respectively. Optical flow-based methods generally estimate motion by examining the change in pixel intensities of frames over a period of time. A region of interest based main direction mean optical flow (MDMO) method was proposed in [92] for MER. The method was immune from the influence of translation, rotation, and illumination variance. Experimental results demonstrated that this method performed better than the baseline LBP-TOP. This method takes both spatial as well as local static motion information into account.

The 2D Gabor filter with sparse representation was employed by [93] and it was found to work better than HOOF and LBP-TOP used in earlier works. Similarly, exploiting Gabor filters further, [94] first magnified ME clips using EVM then employed spatio-temporal Gabor filters for feature extraction. The findings were similar to [93], which further strengthened its superiority over the other two methods. Alongside these methods, attempts have been made to apply DL based approaches which will be discussed in Section 2.4.3.

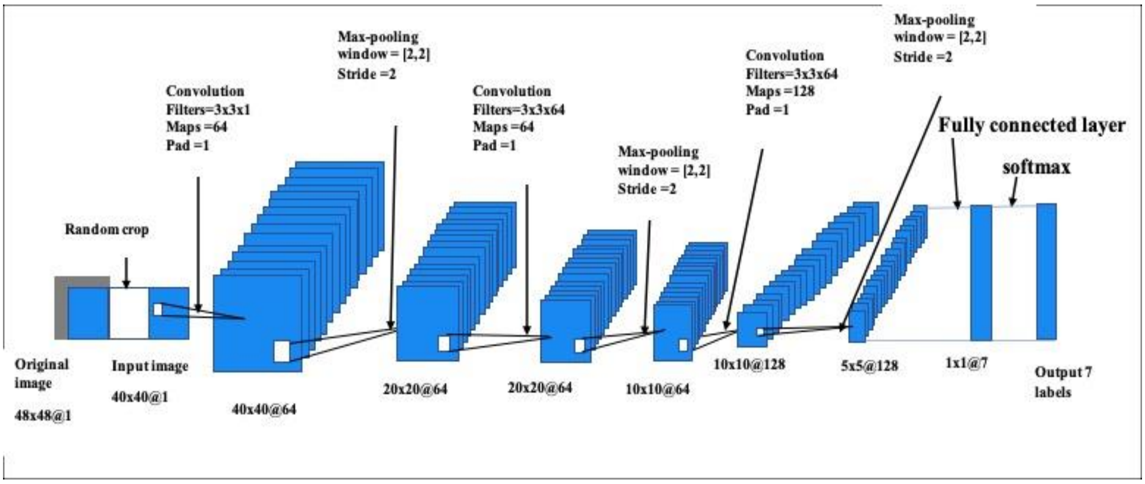


Figure 2.13. An outline of convolutional neural network employed for facial expressions [95].

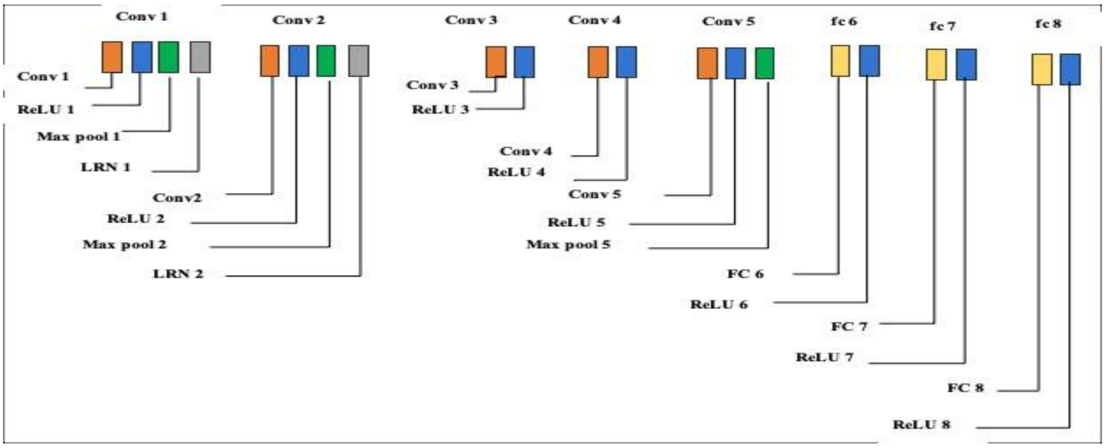


Figure 2.14. Network structure of convolutional neural network [81].

2.4.3 Deep Learning Based Micro Expression Recognition

In recent years, besides handcrafted methods, a variety of DL based MER systems have evolved with excellent outcome. Though the small-scale nature of available ME dataset continues to be a bottleneck for DL based methods, research shows emergence of various techniques to overcome this. The work in [95] was one of the earliest attempts to utilize the DL concept for MER which is based on the architecture given in Figure 2.13. To deal with data inadequacy, transfer learning was applied, where a convolutional neural network (CNN) model was trained on ImageNet and then transferred the appropriate features for further processing. To extract features from datasets dissimilar to ME, the layer positioned right beneath the layers that were fully connected was chosen. In contrast, for datasets that resembled ME, the layer positioned just prior to the final layer with full connection was chosen.

In [81] a CNN based MER system was explored and applied data augmentation to generate exhaustive data for training purposes. The components of its model consisted of a convolutional layer, a rectified linear unit (ReLU), and a pooling and fully connected layer for classification. Features that are useful for performing recognition were extracted by the convolution and pooling layers whereas the classification process was taken care of by a fully connected layer. Five convolutional layers were built to obtain a deep network where the last four layers contained 3x3 filters and the first layer contained 11x11 filters. With a kernel size of 3x3, a total of three max pooling layers were fitted into the architecture. Moreover, it also consisted of three layers which were fully connected. This configuration of the deep network employed is illustrated in Figure 2.14. Using this approach, a significantly improved accuracy of 75.57% was achieved on the newly created database consisting of synthetic images. In [96], a CNN with an optical flow method was proposed which used an apex frame and an onset frame to derive optical flow features which were then sent to a CNN model. The idea behind building the CNN model was to aid emotion class prediction by selecting features relevant for the classification process.

Relying on facial colour related features rather than on motion related features, [97] presented a unique framework for MER. The framework employed a popular recurrent neural network; long short-term memory (LSTM) model and facial colour features. Based on the experimental results obtained, the authors concluded that this colour-based method outperformed some of the previously used motion-based approaches with an accuracy of 66.6% reported on CASME II dataset and 70.5% on SMIC dataset. Thus, the method seemed to be as competent as any other state-of-the art techniques.

Similarly, in another CNN based work [98], simultaneous convolution of spatial and temporal data was performed to obtain motion as well as facial texture representations. Features extracted from 3DCNN were used as input to the long short-term memory (LSTM) model to further boost the temporal data. This method proved to be beneficial for ME cross-database based research. By fusing optical flow features with DL, [99] tested their approach using the CASME and CASMEII databases, achieving recognition rates of 57.80% and 58.03% respectively. Here, 68 facial landmarks were detected using the deep convolutional network, followed by a fused version of a deep CNN, FlowNet2.0, for extracting the optical flow information from the facial regions. To obtain more reliable features, an enhanced version of the optical flow technique was further applied before performing classification.

Exploiting GAN and graph-based techniques, [100] developed the identity-aware and capsule enhanced generative adversarial network (ICE-GAN) for ME framework. The method effectively achieved an increase of 12.9% accuracy compared with other methods. Another example of such work is [101] which has explored the concept of CapsuleNet for ME. In their approach as an initial step, the apex frame is identified by computing the absolute pixel differences between the current frame and onset, this is repeated for the offset frame also. It is then followed by a recognition process utilizing CapsuleNet on the identified apex frame. The method was able to surpass the baseline recognition performance.

These works clearly suggest effective utilization of DL approaches for ME based experiments can be successful, with scope for further exploration. Moreover, transfer learning, CNN, GAN, Capsule net are instances that clearly demonstrate success of DL approaches for MER. Consequently, we will be exploring DL techniques in Chapter 6.

2.4.4 Feature Classification

The task of assigning class labels to input data by designing predictive models using machine learning algorithms is generally known as classification. This stage involves classification of the input into various emotion types based on the extracted features. Designing such predictive models requires a set of data with known class labels to be used for training. To evaluate the performance of such models, one of the commonly used metrics is the classification accuracy. This metric is based on the class labels predicted by such models. Surveys showed [102-103] that the appropriate methods to achieve this for MER include support vector machine (SVM), k-nearest neighbour (kNN), random forest, extreme learning machine (ELM), sparse based representation, CNN based approaches, group sparse learning (GSL) and relaxed K-SVD. Multiple kernel learning (MKL) [70] has also been employed for ME analysis with better results than SVM for some instances. For improving the classification performance with divide and conquer based approaches, [70] also employed a random forest for ME analysis. The nearest neighbour approach is based on distance determined between unknown samples and known samples. The method was employed in one of the earliest works for ME where an accuracy of 65.83% was achieved using the SMIC database [104]. Compared with other classifiers, SVM has been most extensively used for classifying ME.

Throughout this thesis classification is implemented using SVM [105] therefore, working principle of this method is discussed briefly. The datasets are separated into two sets namely training and testing. Individual instances of data in the training set contain several attributes and class labels. Based on this information, if given test data attributes, the SVM will build a model capable of predicting class labels for each instance of such test data. In simple terms, SVM searches for an optimal hyperplane that best differentiates various classes. Figure 2.15 illustrates an instance where three possible hyperplanes are identified that can segregate the data belonging to two different classes. The selection of an optimal hyperplane that gives maximum segregation for data belonging to two different classes is illustrated in Figure 2.16.

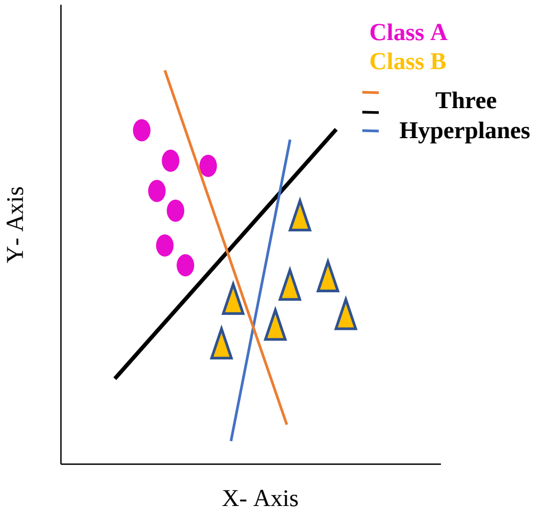


Figure 2.15. Illustration of three hyperplanes identified for data seggregation [105].

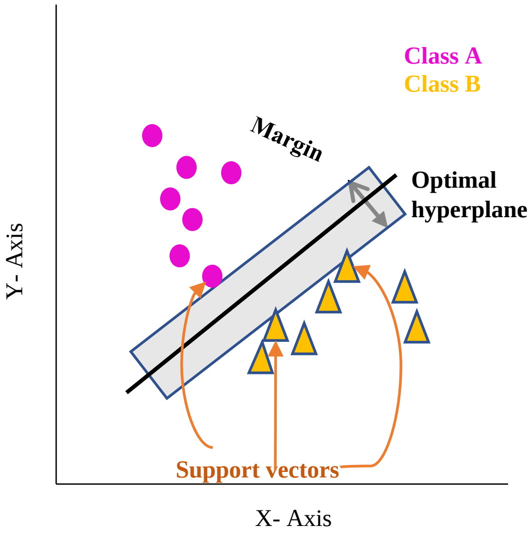


Figure 2.16. Illustration of optimal hyperplane that gives maximum data seggregation [105].

To identify the right hyperplane, it uses the kernel trick. The three most popular kernel functions used with SVM are linear, polynomial, and radial basis function (RBF) denoted by equation (2.11), (2.12) and (2.13) respectively [105]:

$$K\left(\mathrm{x}_{i\,,}\mathrm{x}_j\right) = \mathrm{x}_i^T\mathrm{x}_j \tag{2.11}$$

$$K\left(\mathrm{x}_{i\,,}\mathrm{x}_j\right) = \left(\gamma\, \mathrm{x}_i^T\mathrm{x}_j + r\right)^{\,d}, \gamma > 0 \tag{2.12}$$

$$K\left(\mathrm{x}_{i\,,}\mathrm{x}_j\right) = \exp\left(-\gamma\, ||\mathrm{x}_i\, - \mathrm{x}_j||^2\right), \gamma > 0 \tag{2.13}$$

where d, r and γ are the kernel parameters, $\left(\mathrm{x}_{i\,,}\mathrm{x}_j\right)$ represent training samples and T denotes the transpose operation. In equation (2.12), d refers to polynomial degree, r is coefficient, and γ in (2.12) and (2.13) is the gamma parameter that describes the scale of influence of each training sample.

2.5 Recognizing Micro Expression with Low Resolution Images

The advancements in MER techniques are accelerating at an exceptional rate in recent years. Envisaging a real environment, the recordings captured in our everyday life are prime sources for many studies, but these data often suffer from poor quality. Consequently, this has opened up a new research direction involving low resolution ME images. Identifying a particular class of ME among several classes is extremely challenging due to less distinct inter-class discriminative features. LR of such images further diminishes the discriminative power of micro facial features. Undoubtedly, this increases the recognition challenge by twofold. To address the issue of LR for facial ME, [41] proposed reconstructing HR images from LR images by employing a face hallucination algorithm on individual frames. At present datasets available for ME contain only HR images. For instance, the CASME II micro expression dataset contains HR images with resolution of approximately 280 x 340, whereas LR images are usually below 50 x 50 [41] resolution. Hence in [41], a LR micro expression image dataset was obtained by simulating three existing HR micro expression image datasets i.e., CASMEII, SMIC-HS and SMIC-subHS. The framework employed in their work to reconstruct super resolution (SR) ME images from LR images is depicted in Figure 2.17. Through experiments an improvement on overall classification accuracy was achieved using these datasets. However, low accuracy for individual classes was also observed alongside. From the results obtained, a drastic decline in the recognition accuracy was observed for expressions with exceptionally LR. Datasets from CASMEII and SMIC-HS yielded higher magnitude of misclassification than SMIC-subHS. It was observed that the reliability and validity of any FE analysis approach is directly affected by the resolution of the input image used hence acquiring decent resolution for the reconstructed facial ME images was crucial when employing SR techniques.

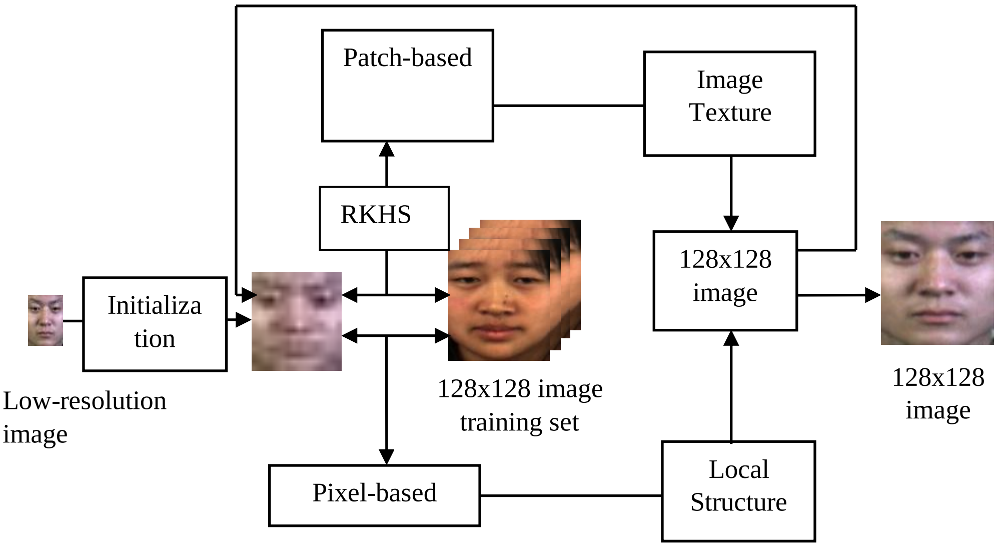
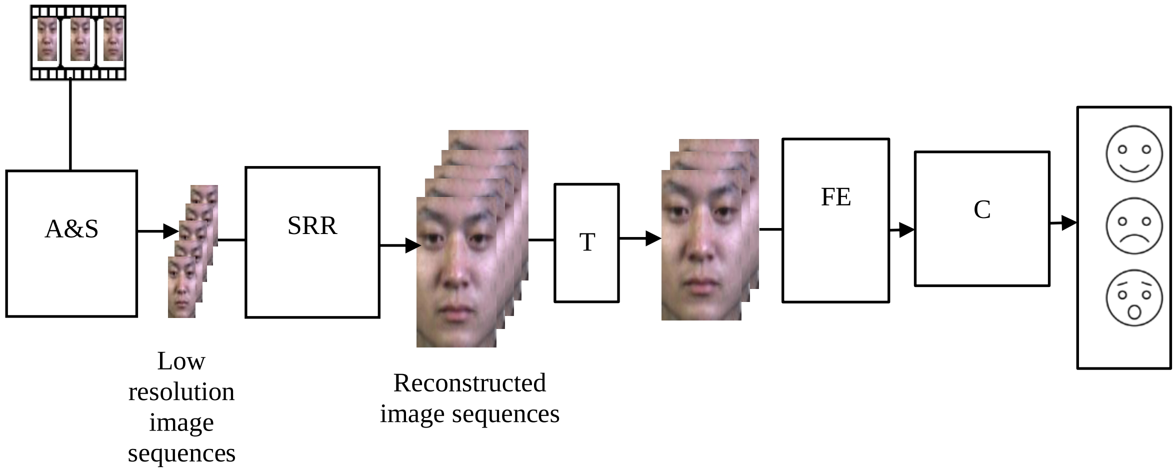


Figure 2.17. Reconstruction technique employing super-resolution for LR images [41].



Note :
A&S :Alignment & Segmentation SRR :Super resolution reconstruction
T : Temporal interpolation model (TIM) FE : Feature Extraction
C :Classification

Figure 2.18. Framework for microexpression recognition with LR images [41].

Work involving face SR for expression analysis has employed macro expressions, thus SR on ME is a unique concept introduced in [41]. Therefore, at the time of writing this thesis [41] is the only work that has employed the concept of SR for LR micro expression images. The general pipeline adopted in their work is presented in Figure 2.18. It must be noted that the solution provided in their work does not consider DL based methods. Taking this concept further, this thesis attempts to solve the LR problem in images for ME by employing several DL techniques which will be discussed in Chapter 6.

2.6 Application of Micro Expression Recognition System

The conception of ME can be traced back to 1969 when Ekman and Friesen [39] were examining an interview of a psychiatric patient. An extremely brief sad face was detected when the interview's video clip was played in slow motion. Thorough examinations affirmed that this lasted only for two frames which is approximately 1/12 of a second. This was followed by a quick dubious smile stretching for longer span. These observations revealed that the patient was trying to supress an underlying negative feeling, but some expressions leaked involuntarily. Such subtle expressions are termed as ME and are believed to reveal genuine emotions. Thus, application of ME in the field of psychology is evident. Similarly, Ekman also suggested that MER could be beneficial in understanding emotion of individuals with autism since they exhibit extremely fewer expressions on their face. Like facial macro expression [106-107], usefulness of ME can also be explored for assessing pain intensity for patients who are unable to express themselves. These instances illustrate tremendous potential of MER in identifying useful checkpoints that could assist in clinical diagnosis and investigation. Extending its application to mental health for estimating signs of depression, anxiety and suicidal tendencies can be very impactful and beneficial endeavour.

Another instance where MER can become useful is in certain negotiations for business, politics, public policies etc. Negotiations involve decision making and studies show that emotion is an important factor that influences such decision-making processes [48]. By carefully examining such ME, one can get an intuitive insight into the party/individual's state of mind during negotiations, which can be helpful in making informed decisions. Such decisive advantage can effectively save time by avoiding unfruitful conversations with people having little or no interest in affirming contracts.

A person is usually vulnerable during high stake situations due to which leakage of expressions are sometimes inevitable, therefore MER can be helpful for detecting a lie. This belief was strengthened through findings by Ekman [39], where he confirmed such leakage was a powerful tool in identifying deceit. In an experiment conducted by [108] involving a mock crime scenario, participants were asked to either be truthful or lie about stealing. Results obtained revealed that ME analysis was effective in differentiating between truthtellers and deceitful people. These findings clearly support the idea of utilizing MER for identifying tip-off during criminal investigations, law enforcement and similar fields[23,40,49]. Identifying security threats for border control, in airports etc are another set of instances where application of MER can be advantageous [40]. By examining an individual's ME, suspicious behaviour exhibited by them can be detected and can be dealt with suitably [40]. Thus, situations where screening individuals may be necessary for security reasons can certainly benefit from MER system.

Reading ME for obtaining cues regarding students' progress as part of classroom communication is another instance of ME application. Such application in academics field was explored by [109-110]. In [110] ME was obtained to determine the concentration level of students in a classroom. The information gained was then used for adjusting teaching methods to suit the student's state of mind. Both these works promote the successful application of ME for academic monitoring and meaningful learning.

Building an affective robot is another instance where MER finds its application. Such robots can read facial ME to improve communication between human and robot. It can be effective in helping people who have difficulty in expressing emotion like elderly, disabled, or autistic individuals. Many other fields exist where MER can be applied and is beyond the scope of this thesis to cover them all. Such a wide range of inter-disciplinary application of ME has drawn attention of many researchers and its further expansion is evident in coming days.

2.7 Summary

This chapter provided a review of various techniques that have been developed and applied to different components of MER systems. The basic model of a MER system along with its building blocks has been examined here. Various techniques and algorithms employed within these blocks for improving performance has also been presented. Through this extensive literature review, progress made so far in the field of automated MER can be comprehended. Amidst a wide range of approaches to ME analysis and recognition from various interdisciplinary groups, its resounding success is evident. With an increase in the interest in affect analysis with ME, substantial progress in the last decade has been identified. Exploring facial, ME has helped to uncover the current trends and challenges in this field. Study shows that accomplishments in ME analysis have successfully explored handcrafted as well as DL techniques, though approaches employing handcrafted methods are significantly higher in number. Based on the review it is fair to claim that the progress of DL based MER system was considerably influenced by designing novel databases through augmentation approaches. It has been observed that there exists significant works that have employed good quality ME images to achieve state-of-the-art recognition accuracy, whereas a small number of works exist that have explored low quality images for this task.

Designing methods for micro feature representation to effectively encode subtle movements is one research area within ME analysis which will be explored in this thesis. Also, we have explored both HR and LR micro expression images in experiments conducted for this thesis. In this thesis, work employing HR data is discussed in Chapter 4 and Chapter 5, and in Chapter 6 the work employing LR images for ME is presented.

Chapter 3

Image Super Resolution: Theories and Techniques

3.1 Resolution Implications for Micro Expression Analysis

In CV the need for good resolution images is vital for algorithms to achieve reliable and superior performance. Conventionally, analysis of ME has been performed using HR images which are ideal cases. These images are taken from datasets that are produced in ideal conditions with good lighting, no interference of illumination variations, full frontal view with no obstructions and resolutions of approximately above 150x150. However, in a real-world scenario, capturing expressions with HR may not always be possible particularly using low-cost surveillance cameras. External factors like ill pose, meagre lighting conditions, non-uniform illumination etc. can severely impact the quality of images captured using such low-cost devices. Very often faces captured using such cameras appear alongside several other objects hence are likely to take up only a limited space in the entire image. Thus, faces captured appear both very tiny as well as with poor resolution. The quality may also be affected if downloaded images, or images stored in restricted memory capacity etc. are employed. In general, exploring LR micro expression can be extremely beneficial particularly for crowd scenarios and poorly illuminated areas [41]. Since MER is a substantially widespread inter-disciplinary application area it is undeniable that this implicitly spawns situations where images to be analyzed maybe of poor quality. Due to further loss of discriminative features owing to reduced resolution, these images may not be of much use particularly for identifying certain minute facial details. Such insufficient resolution makes it extremely difficult for both humans and machines to utilize the available information.

To make these images useful, enhancing the textural information becomes essential and SR algorithms can be ideal to achieve this. A significant surge in the use of surveillance cameras, especially for monitoring public domains, has created a new challenge for recognizing ME collected under shallow lighting conditions. For these cameras, more emphasis is laid on capturing reliable recordings for longer period which is generally achieved by making a significant compromise on image/video resolution, thereby raising the need for algorithms that can deal with such resolution concerns [111]. SR is one such medium capable of addressing resolution challenges that are often engrained in images acquired using ordinary imaging devices. Reconstructed images obtained using SR algorithms often have improved pixel density subsequently offering more image details. Achieving good resolution using superior hardware is not always cost effective and therefore employing image processing algorithms seem more feasible [111]. The absence of discriminative facial details in ME along with faint muscle movement intensity that lasts for exceptionally short duration characterizes it to be laborious for recognition tasks. Extracting informative attributes from such LR images become effortful due to further loss in the availability of salient information which may have unfavorable influence on the performance of overall MER systems. As such, resolution of these ME images can be a pivotal factor during the recognition process. Differentiating among various classes of ME is already a challenge due to non-distinct features. Therefore, difficulties are further compounded when such expressions are captured with poor resolutions, as distinctive traits in such micro facial features reduces.

The very first work that studied the effect of resolution for ME was by [112]. The original images taken from the CASMEII database were downscaled to 75%,50% and 25%. Performance of three feature extraction techniques were tested on these downscaled images. At the lowest downscale level 3DHOG technique performed the best whereas at HR, the LBP-TOP method seemed to perform much better. At half the resolution histogram of optical flow orientation (HOOF) method gave the best performance in comparison to other two techniques. The work successfully realized the effects of resolution for ME but did not examine resolution enhancing techniques on such LR images. Moreover, the effect of image degradation other than LR was also not explored.

To have more relevance with real-life applications, [41] proposed using deteriorated ME images that were both blurred and down sampled. Three levels for LR were considered i.e., 16x16, 32x32 and 64x64. These LR micro expression images were then super-resolved using patch based and pixel-based face hallucination techniques on individual frames and was the first work to perform MER using deteriorated image quality. At present datasets available for ME contain only HR images. Hence in their work [41] LR micro expression image dataset was obtained by simulating three existing HR micro expression image databases i.e., CASME II, SMIC-HS and SMIC-subHS. Fast LBP-TOP was used for extracting the features which were then classified using SVM. The results indicated that employing significantly LR images at 16x16 level makes it extremely difficult to achieve decent recognition results. Their approach worked comparatively better on SMIC-subHS with less misclassification reported than other two databases. Employing SMIC-HS database images at size 16x16, a drastic improvement on the recognition results was reported particularly for positive labels. In contrast, substantially higher misclassification results were reported for CASMEII database. Another observation made for this database was that most of the data were misclassified into “others” category. When recognition accuracy obtained for SR images were compared with their corresponding LR images a significant improvement was noticed for all the three databases at all chosen resolution. By analyzing the overall reconstruction performance, observed through structural similarity index (SSIM), it was clear that the method produced best reconstruction results for SMIC-HS database followed by CASMEII and SMIC-subHS at 64x64 level. Same trend was observed for the other two levels also. However, observing peak signal to noise ratio (PSNR) suggested that reconstruction performance on SMIC-subHS was better than on CASMEII database at both 64x64 and 32x32 level. Though reconstruction values obtained for SMIC-subHS database was slightly less compared to other databases, yet it successfully produced best recognition results recorded at 74.65% which is much higher than that obtained for SMIC-HS and CASMEII database at 52.44% and 48.18% respectively. Lower volume of data samples along with fewer class categories in SMIC-subHS might have worked in its favor thereby producing better recognition results in comparison.

It must be noted here that to the best of the author’s knowledge, [112] is the first work to examine the effect of resolutions for MER. The work employed downscaled images but did not consider degrading images in their implementation. In [41], discussed earlier, the authors have addressed low quality issues for MER where quality of image is severely affected due to poor size and blurring. In their approach SR method was employed for enhancing such low-quality images without DL techniques. Therefore, from extensive literature study limited work addressing poor quality issues for MER was noticed. Inspired from these two works, this thesis takes their concept forward and explores the recognition process for ME employing low quality images.

To address the low-quality issues for facial ME images, this thesis proposes a novel approach that employs a SR technique using DL, GAN, and its variant. Therefore, this Chapter will discuss these SR algorithms in detail to be employed later in experiments for enhancing LR micro expression images.

It must also be mentioned here that at present public database available for ME contain only HR images therefore are unsuitable to be tested directly by SR algorithms. To simulate appropriate LR database suitable for SR methods, this thesis applies image degradation technique which shall be discussed in Section 3.6

3.2 Progression of Deep Learning Super Resolution Techniques

An image with dispersed and loosely aligned pixels consisting of comparatively fewer image details within them than standard resolution image is identified as a LR image. This obviously makes it appear pixelated, less precise, blurrier, and granular. On the other hand, image with more concentrated and compact pixel arrangement in addition to crisper and clearer appearance is classified as HR image. Implicitly, image details contained in it are much denser and condensed. Broadly, image with LR differ from HR image mainly in terms of pixel density per unit area and degree of coherence. A substantial lack of salient information (e.g., texture details, high frequency information etc.) in LR image makes the process of attribute extraction extremely challenging and laborious. The task of estimating a HR image by reconstructing an image from a LR input image of the same scene is generally known as image super resolution (ISR) and the reconstructed image is known as a SR image. Such methods are expected to overcome the influence of various degradation factors like blur, noise etc., acquired during image acquisition. Producing improved reconstructed scenes along with restoration of essential details is paramount for these methods. For instance, when super resolving LR facial images, recovering facial details is imperative. The challenge is not only to reconstruct the face, but also to maintain attribute consistency with the original HR images. Thus, restoring face details in the reconstructed image is vital for face SR algorithms, to facilitate FE analysis. SR image can be estimated using HR video or multiple images or single image. In this thesis SR image was produced from single image input using DL methods hence the scope of discussion largely focuses on DL techniques for single ISR. The concept of SR was first introduced in 1984 by Tsai and Huang [113] where an image was reconstructed by employing multiple frames and has now progressed with several advancements.

One of the early works that applied DL concepts for SR employed a fully convolutional neural network (FCNN) [114-115] with a very straight forward architecture. These types of networks do not have dense connections at their rear and the technique was named super resolution convolutional neural network (SRCNN). Due to the absence of pooling, the output images obtained were the same size as the input images. The network was composed of three convolutional layers and two rectified linear units (ReLU), where structurally ReLU succeeded every convolution layer with an exception in the final layer. The first layer in the SRCNN was used for extracting features from a given LR image input. The second performed non-linear mapping where the extracted features were mapped to its corresponding high dimensional representations. The final layer utilizes the prediction information borrowed from its neighborhood to generate the SR image as output. The architecture employed to formulate SRCNN is illustrated in Figure 3.1.

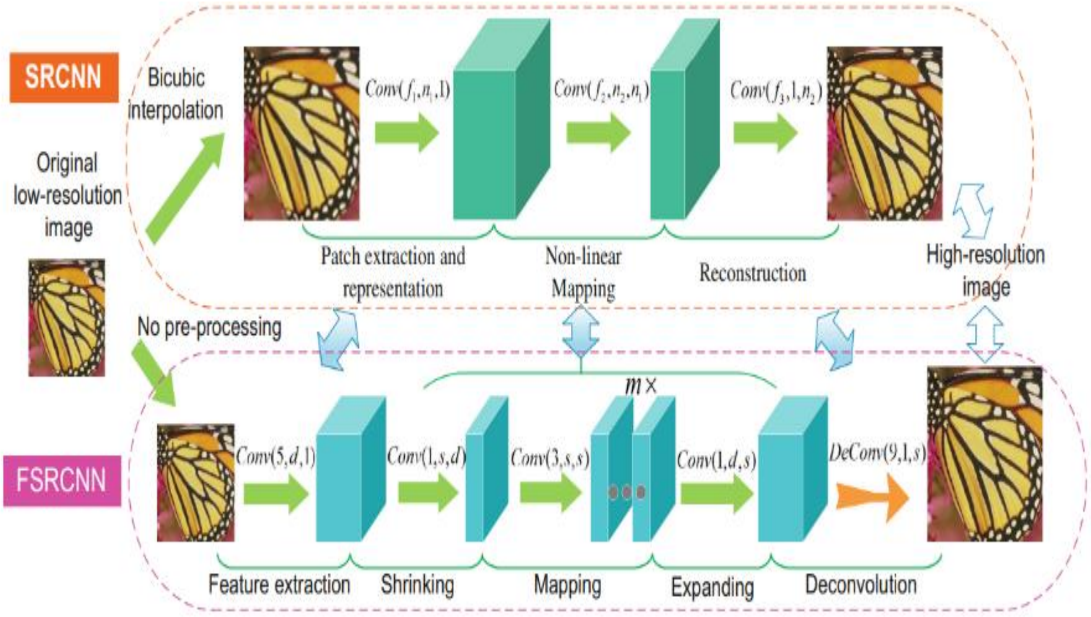


Figure 3.1. Illustration of network structure for super resolution convolution neural network(top) [116] and fast super resolution convolutional neural network(bottom) [116].

Later, an improvement on this method introduced by [116] performed 40 times faster with exceptionally higher restoration capability and was known as fast super resolution convolutional neural network (FSRCNN). It was composed of five components namely feature extraction, shrinking, mapping, expanding and deconvolution. This structure utilized for FSRCNN is illustrated in Figure 3.1. Another advantage of this method was that it was able to perform both training as well as testing with much higher acceleration for varying upscaling factors.

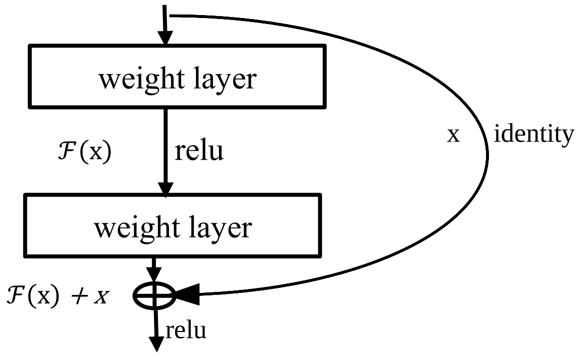


Figure 3.2. Illustration of residual learning [138].

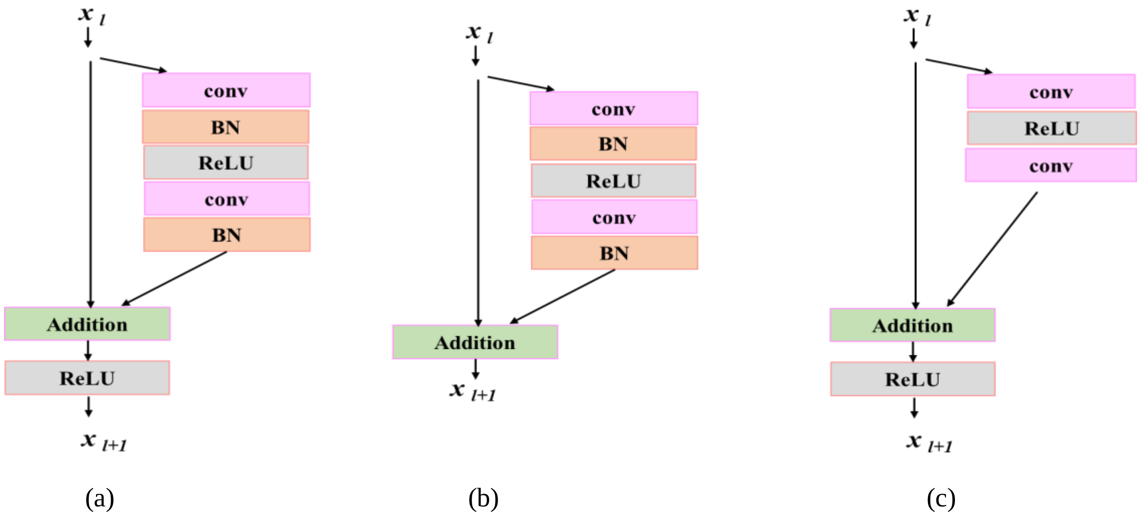


Figure 3.3. Illustration of residual block [118] used in (a) conventional ResNet, (b) Super resolution with ResNet and (c) for EDSR.

In contrast to these methods, very deep networks were designed to perform SR that utilized residual learning as illustrated in Figure 3.2. The basic idea of such approach is to learn the residue i.e., difference between given input and the ground truth. A deep residual learning (ResNet) based SR technique called SRResNet used residual blocks instead of convolution and demonstrated better performance than SRCNN [117]. By removing the batch normalization module from the conventional residual network (ResNet) (refer Figure 3.3 (a)) and SRResNet (refer Figure 3.3 (b)); [118] was able to optimize the model and named it as an enhanced deep super resolution network (EDSR) (refer Figure 3.3 (c)). Additionally, it was able to save memory usage by a significant margin of 40%. With further advancements, densely connected CNN architectures evolved for enhancing the network performance and became popular for solving the SR problem. The architecture named as SRDenseNet benefited due to an accelerated training process from the dense connections introduced between the convolutional layers. An illustration of such dense block utilized in SRDenseNet is presented in Figure 3.4 (a).

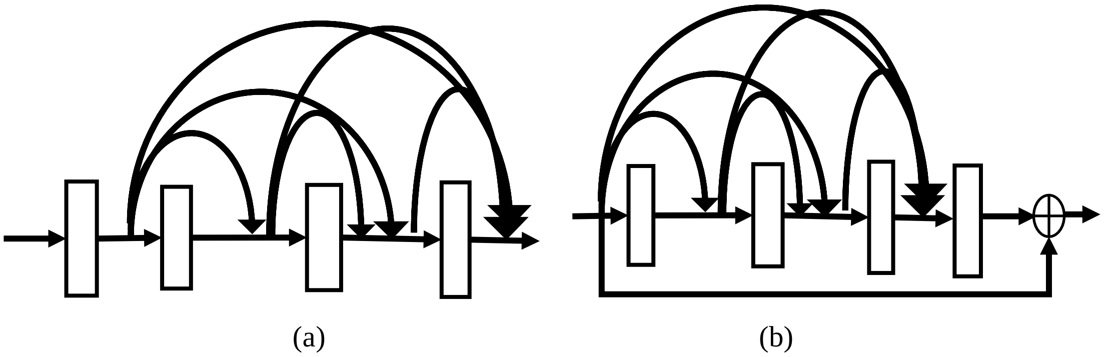


Figure 3.4. (a) Dense block used in SRDenseNet architecture (b) Residual dense block used in residual dense network [119].

A residual dense network (RDN) is another instance that uses this type of architecture with some modifications (refer to Figure 3.4(b)). For learning local patterns this technique considers utilizing hierarchical feature representations exhaustively. Using a very deep RDN architecture to solve image SR problem [119] achieved favorable results. With the evolution of generative models in various CV related fields, its usefulness was noted for SR algorithms too. GANs have a very powerful ability to learn and hence they become useful for SR tasks. The work in [120] used a GAN architecture to perform SR over a single image and called it a super resolution generative adversarial network (SRGAN). The image was super resolved with a factor of four and was the first work to have successfully upscaled images with this factor. Notably the images produced in their work were more realistic than those obtained from any other state-of-the art techniques used before. Using a novel approach by introducing a residual in residual dense block (RRDB), [121] was able to improvise on the visual as well as texture quality of its super-resolved output images. Several innovative deep convolutional neural networks (e.g., CNNs) are now available with variations that exploit RDN, residual dense blocks (RDB) and recursive learning architectures and have been successfully applied to SR problem. Borrowing the RDB and GAN based SR approaches from these works this thesis tests them on

ME images. Only those work closely related to the methods employed in this thesis are highlighted in this section. An elaborative discussion of the methods employed in this thesis are presented in the following sections.

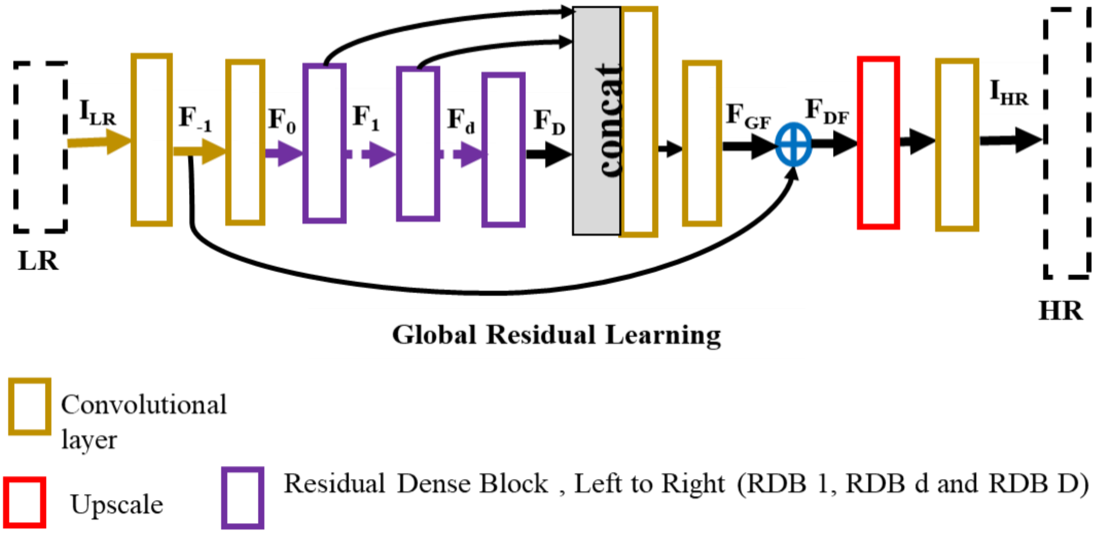


Figure 3.5. Residual Dense Network with contiguous memory [119].

3.3 Residual Dense Network

Most DL approaches utilized for SR before [119] suffered from certain flaws like increased computational complexity, low growth rate, loss of image details from its LR input, variation in scales etc. To overcome loss of image details, [119] introduced the RDN, capable of fully exploiting hierarchical features from the LR input. RDB was used as the building blocks for RDN structure. A detailed illustration of RDN utilized for performing ISR is given in Figure 3.5. Structurally, RDB consists of three primary components i.e., local feature fusion (LFF), local residual learning (LRL) and dense connected layers. Since the output of one RDB is directly connected to every layer of its succeeding RDB, the entire RDB structure supports contiguous memory amongst them. The RDN was introduced for exploiting the hierarchical features from all its convolutional layers which had not been accomplished by previous CNN based SR algorithms. Four components of RDN architecture include shallow feature extraction, dense feature fusion followed by up-sampling network as illustrated in Figure 3.5. In order to extract the required shallow features from the LR input it uses two convolutional layers. In Figure 3.5 the shallow features extracted by the first layer are denoted by F_{-1} and F_0 represents features extracted by the second layer. If the total number of RDB in the architecture is denoted by *D,* then the output of the d^{th} RDB is given by F_d. The hierarchical feature extraction process is then followed by dense feature fusion (DFF). This is realized by performing global feature fusion (GFF) in addition to global residual learning (GRL). The DFF is represented with the help of an equation given below [119]:

$$F_{DF} = H_{DFF}\,(F_{-1}, F_0, F_{1,}\ \ldots., F_D) \qquad (3.1)$$

In equation 3.1, F_{DF} denotes the output of DFF and H_{DFF} is the composite function of two operations namely, convolution and ReLU. Following the feature extraction process in both local and global space, the next step is up sampling and generating the HR image.

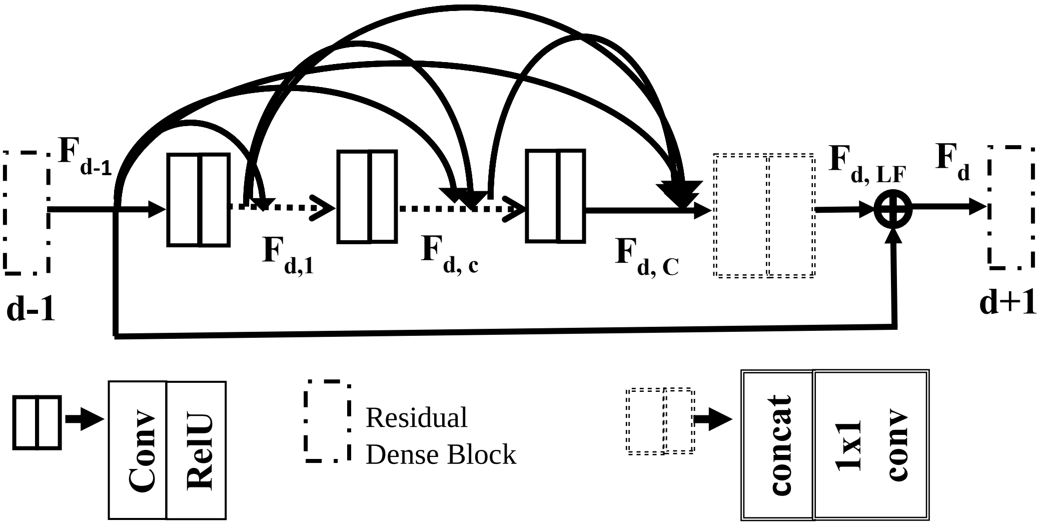


Figure 3.6. Architecture of residual dense block [119].

RDN is composed of RDB which is a densely connected CNN capable of extracting ample local features. A detailed illustration of the RDB architecture is presented in Figure 3.6. Also, due to a contiguous memory (CM) mechanism between several RDBs, it is capable of continuously learning from more improved features. To achieve this, every layer of the current RDB is fed with the state of the preceding RDB. If F_{d-1} denotes the input and F_d denotes the output for d^{th} RDB with G_0 feature maps for both input and output, then the output of its c^{th} convolutional layer is given by [119]:

$$F = \sigma\left(W_{d,c}\left[F_{d-1}, F_{d,1}, \dots, F_{d,c-1}\right]\right) \tag{3.2}$$

In equation (3.2), the activation function is the rectified linear unit (ReLU) denoted by σ, the weight for c^{th} convolutional layer is represented by $W_{d,c}$, and the feature maps produced by (d-1) RDB in their concatenated form is given by $\left[F_{d-1}, F_{d,1}, \dots, F_{d,c-1}\right]$. The convolutional layers in the d^{th} RDB are 1, ….,(c-1) which produce a feature map $G_0+(c-1)$ x G, where G represents feature map for $F_{d,c}$. Thus, a direct connection is established between successive layers and the output of each layer and RDB. This is followed by LFF where in order to downsize the extracted features the d^{th} RDB is directly fed with $(d\text{-}1)^{th}$ RDB feature maps in concatenated form. Also, the output information is regulated using a 1x1 convolutional layer. Since the architecture is a combination of LRL and dense connections, this architecture is often known as RDB. The advantages of utilizing LRL are it enhances the flow of information and boosts the network representation capability.

To extract global features F_{GF} the network further uses DFF composed of GFF and GRL. Here all the feature maps produced by all the RDB within the architecture are concatenated. This is then fused with the two convolutional layers of 1x1 and 3x3. As the final step GRL is then employed before performing the upscaling denoted as F_{DF} in Figure 3.5.

3.4 Generative Adversarial Network

Ideally, GAN [122] based SR algorithms consist of three vital components namely generator, discriminator, and a loss function. The generator function estimates a HR image for the corresponding LR image whereas, the discriminator estimates if the generated image is realistic enough. Both these models try to tweak their parameter settings based on the outcome obtained by both the models to produce improvised results. For instance, if the discriminator fails to give correct prediction, then it uses the error information to avoid making similar mistakes in next successive rounds. However, if the discriminator is successful in making the correct prediction, then it means the generator model failed hence this time the generator tries to update its parameters and improvise its fake image generating abilities. If frequency of correct predictions is high then it demonstrates the better capability of discriminator, while higher the error count for the discriminator better the generator’s competency. By repeatedly performing these procedures it forces the generated data to get as close as possible to the actual data. Additionally, the loss function essentially helps in optimizing the overall GAN framework by quantifying the similarity between the real data and the generated data. A basic framework to illustrate GAN architecture with its generative and discriminative models is presented in Figure 3.7.

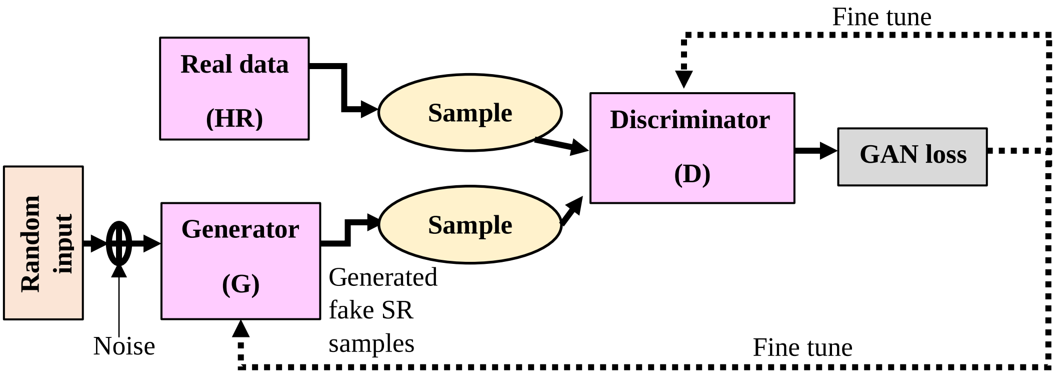


Figure 3.7. Architecture of generative adversarial network [123].

During implementation both generative and discriminative models are trained simultaneously. Initially, the discriminator network is trained on real data for a period of time generally termed as epoch using a forward propagation. Following this, it is again trained on fake data generated by the generative model. Here the discriminator is expected to make a prediction if the image is fake. During this training phase of the discriminator, the generator model remains in idle state. Similarly, in the successive phase the generator is trained keeping the discriminator in idle state. Predictions made by the discriminator model in the previous phase is utilized to train the generator model in this phase which helps in tuning its results. This is also repeated for number of times measured in terms of epoch. Therefore, in order to train the discriminator as well as the generator it backpropagates the values computed as GAN loss. SR techniques that have employed GAN and are closely aligned with the work presented in this thesis are presented in the next subsections.

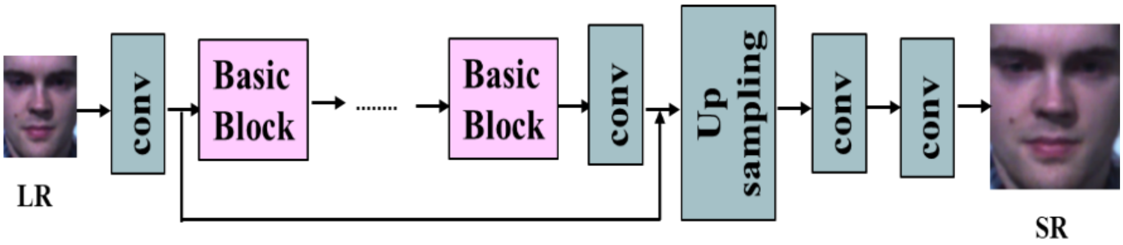


Figure 3.8. Basic architecture of super resolution network [120-121].

3.4.1. Artefact Cancelling Generative Adversarial Network

Broadening the GAN application, it has been utilised for ISR too in [120] and the method was known as SRGAN (see Figure 3.8). The network employed in the generator for SRGAN was a feed-forward CNN with two layers built using residual blocks (see Figure 3.9); with weights and biases obtained through the loss function optimization. The

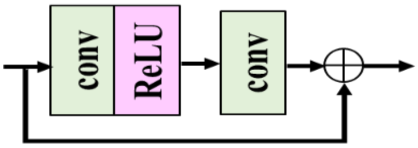


Figure 3.9. Residual block [119].

layers were built using 64 feature maps, batch normalization (BN) layers and a 3x3 kernel and activation function named ParametricReLU. On the other hand, for building the network for discriminator LeakyReLU was employed. Further the network contained 8 convolutional layers of 3x3 kernels. Similar to the VGG network the kernel was incremented by employing a factor of two to obtain a resultant feature map of 512. To compute the probability estimation, the feature map was succeeded by two dense layers along with an activation function namely, sigmoid. The architecture of the generator and discriminator network employed in SRGAN is illustrated in Figure 3.10 and Figure 3.11 respectively. For estimating the loss at perceptual level, the method employed a weighted sum of two components namely adversarial loss and VGG based content loss. The approach using VGG feature loss as well as adversarial loss attempts to eliminate the noise. This method is referred to as noise-cancel throughout this thesis.

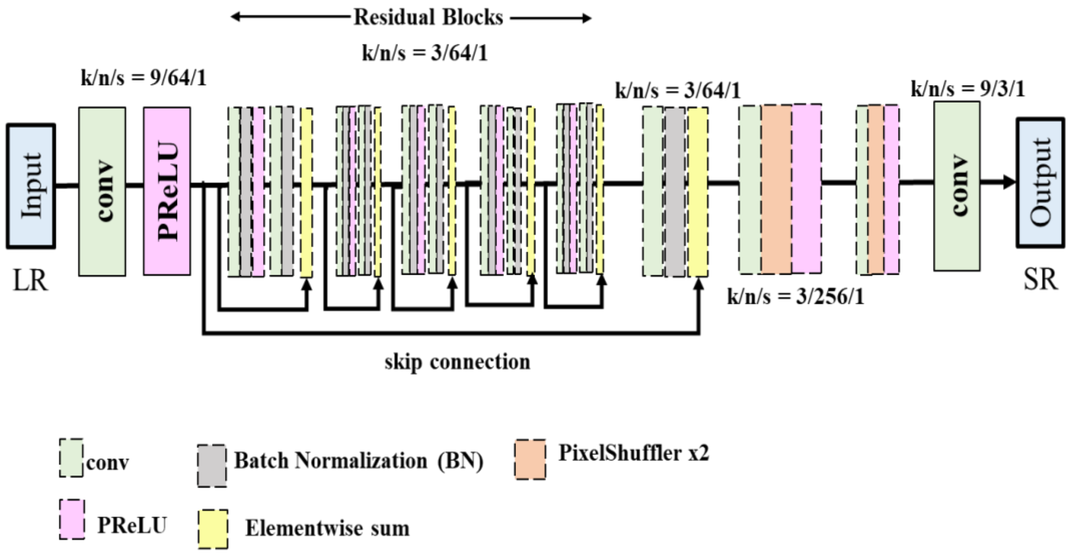


Figure 3.10. Architecture of generator network with specifications for kernel size(k), number of feature maps and stride(s) for each layer of convolution [120].

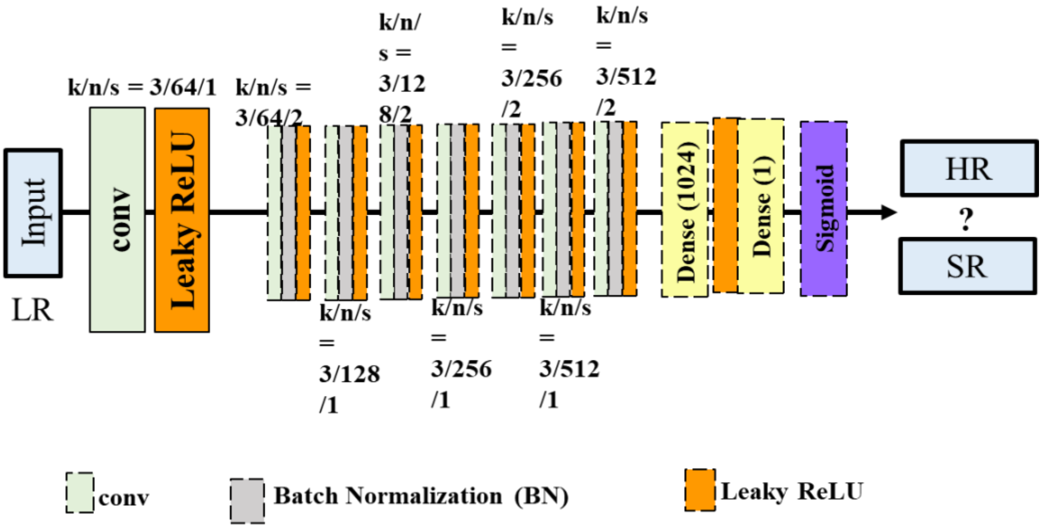


Figure 3.11. Architecture of discriminator network with specifications for kernel size(k), number of feature maps and stride(s) for each layer of convolution [120].

3.4.2. Enhanced Super Resolution Generative Adversarial Network

By introducing 23 residual-in-residual dense blocks (RRDB) into the generator network structure along with improvisation on adversarial and perceptual loss, [121] successfully built the enhanced super resolution generative adversarial network (ESRGAN). The new architecture was basically an upgrade of the original SRGAN, with removal of BN and introduction of residual scaling (see Figure 3.12). Removal of BN benefited the network with scaled down complexity of computation and removal of artifacts from the generated images. Here, several RRDBs are employed, where each of these RRDBs are built using several RDBs. Further, each RDB consists of several convolutional layers (see Figure 3.12) and every convolutional layer that exists within the RDB consists of feature maps same as that discussed in Section 3.3. and utilizes residual scaling denoted by β.

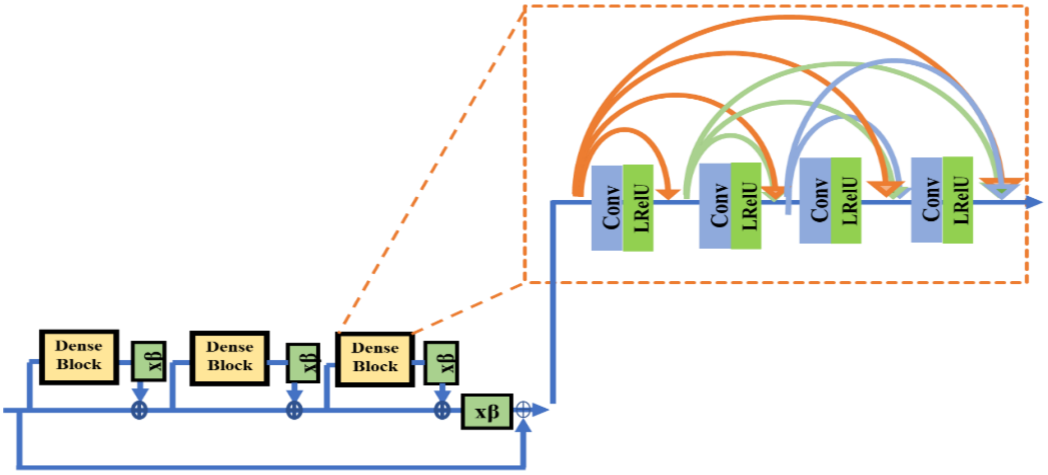


Figure 3.12. Residual in residual dense block (RRDB) [121].

The architecture of the discriminator employed is known as relativistic average discriminator (RaD). To gather salient information like textural attributes and sharp edge characteristics and continuously learn from them; both real and generated data are provided during adversarial training. In contrast to SRGAN in this architecture the perceptual loss was minimised since the VGG features were utilized before employing activation function. Through experiement it was observed that the ESRGAN approach was able to produce much sharper images and with higher details than SRGAN. The model generates SR images from LR image by scaling it with a factor of four. LR images used during experiments were obtained using bicubic kernel function. For network interpolation, it undergoes two training first based on PSNR and second based on GAN then derives the final interpolation model using parameters from both PSNR and GAN models.

3.4.3. Further Improving Enhanced Super Resolution Generative Adversarial Network (nESRGAN+)

To bring further improvements on images generated by ESRGAN at the perceptual level, the work in [124] introduced new blocks instead of the usual RRDB. In this new structure an additional level of residual learning was introduced inside every dense block and was known as residual-in-residual dense residual block (RRDRB). The method was developed to bridge the gap between the images generated by the ESRGAN method and their corresponding ground truth. An overview of this further improved design is illustrated in Figure 3.13 and is referred to as nESRGAN+. This new RRDRB was built with a more superior network structure than the usual RRDB with further denser network, with an extra layer of residual learning augmented into the structure compared to ESRGAN architecture. The residuals are added at an interval of every two layers. Moreover, Gaussian noise is also injected after each residual in this architecture.

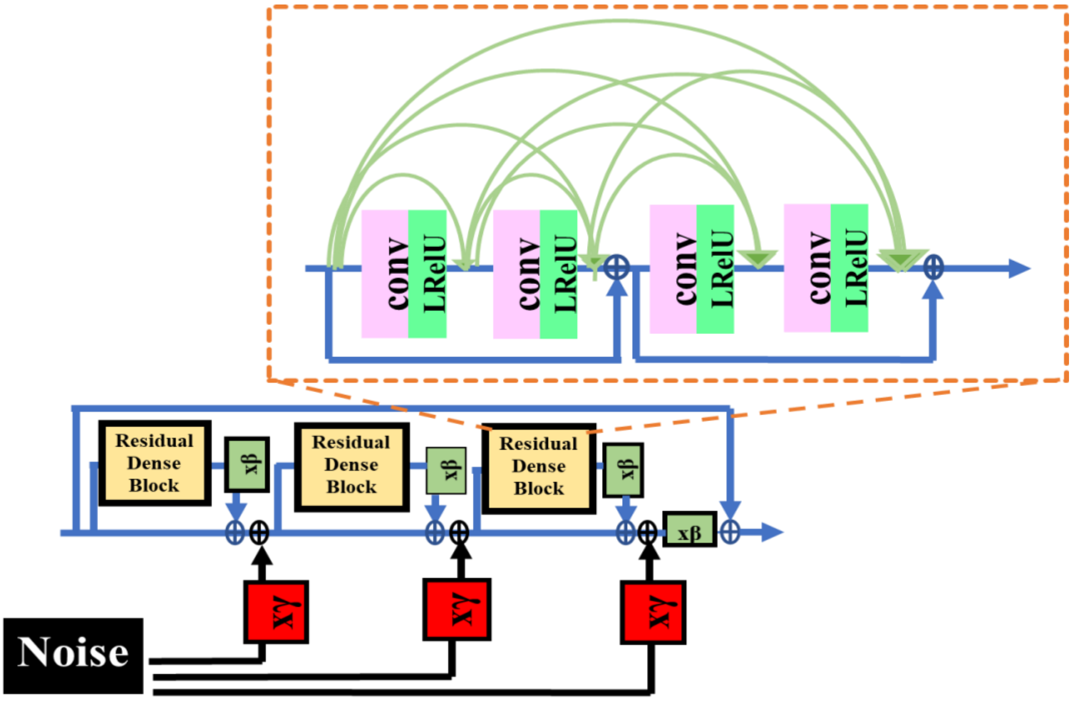


Figure 3.13. nESREGAN+ architecture employed for super resolution [124].

Experimentally it was found that the images generated using RRDRB had substantially better visual quality than its previous variant. The method was tested on LR images with a scaling factor set to four. Sampling down the original HR images through the bicubic kernel, these LR images were acquired. For implementation, the generator was built using 23 blocks and trained accordingly. The method was successful in restoring most of the image textures that existed in the original HR image. Through qualitative measurements like PSNR and perceptual index (PI) it was established that this method produced comparatively superior super-resolved images than both SRGAN and ESRGAN methods.

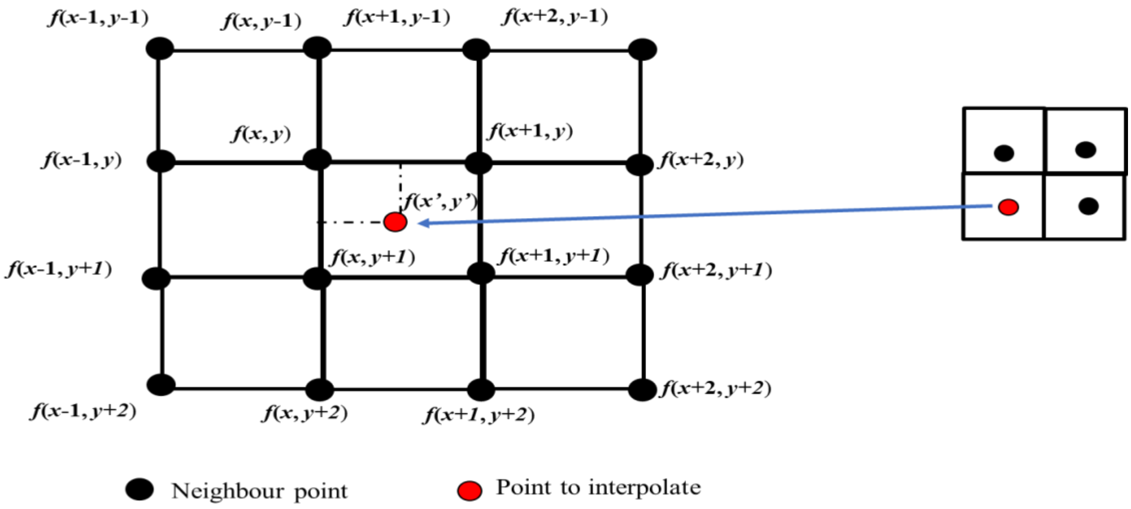


Figure 3.14. Illustration of 4x4 neighbourhood for computing 16 coefficients [125].

3.5 Bicubic Interpolation

The basic working principle of interpolation techniques is to estimate the value at some unknown positions by utilizing the information from a set of known data points. For such methods as the quantity of known data increases, as will the accuracy estimation for the pixel under consideration. Bicubic interpolation takes a simple approach for estimation by utilizing information from a neighborhood of size 4x4 (refer to Figure 3.14) [125]. The weighted average of the 16 nearest neighbors is computed where the weight for each known point is based on its distance from the target interpolation point. By applying a third order polynomial function this method ensures that within four corner points the required surface can be fitted. It utilizes the value of intensity at these four points in addition to the derivatives along three directions i.e., diagonal, vertical and horizontal. The interpolated area is represented using equation 3.3 [125]:

$$f_i(x, y) = \sum_{j=0}^{3} \sum_{j=0}^{3} a_{ij} x^i y^j \tag{3.3}$$

Here, $f_i(x, y)$ denotes the interpolated area for the point (x, y) and a_{ij} denotes the coefficients. Sixteen coefficients are computed in total, among them four are computed from the intensity values at four corners. Further, from the diagonal derivates, four other coefficients are computed. Lastly, taking the horizontal and vertical directions and utilizing their spatial derivatives information; eight coefficients are computed. Since the estimation is based on a greater number of known pixels, the method is regarded as one of the standard methods and often used for making fair comparisons with similar methods.

3.6 Image Degradation

The principal factor that determines the quality of any image is spatial resolution and is represented by the number of pixels per unit area for a given image. Due to growing demand to have good resolution images for digital image processing and its allied fields, data are usually captured from cameras possessing satisfactory resolution. Such HR images often enhance the results significantly. Datasets that are currently available for ME contain images captured using cameras with good resolutions and under a controlled environment. As such they are HR images and are not useful for SR tasks. Therefore, new sets of LR databases containing low quality images suitable for working with SR algorithm need to be built.

To construct deteriorated images, down sampling and Gaussian blurring are applied on the existing HR micro expression images for all three databases to obtain three corresponding sets of simulated databases. This process of creating images with loss of quality from its HR images is known as image degradation and can be expressed using equation (3.4) [41]:

$$L = SYZ + x \tag{3.4}$$

The symbol *S* denotes down sampling, *Y* denotes blurring, *Z* denotes high resolution, *x* is other additive noise and L represents the LR image obtained.

3.7 Summary

This chapter introduced the theory of resolution for ME and why they are important during the recognition process. Early work examining implications of resolution for MER in computer vision involved methods without DL models. To improve the research by introducing DL methods into ME analysis, this chapter moves on to discuss the concepts of SR and progression of DL techniques in SR domain. It also clearly depicts the steady growth of GAN models but, has a long way to go before being well-established for facial ME analysis. Facial expression analysis using SR tends to focus more on macro expression. Therefore, applying these methods with focus on facial ME seems relevant and achievable. All theories and methods described in this chapter provide the foundations on which the contribution Chapter 6 will be based upon.

Chapter 4

Local Phase Quantization for Micro Facial Feature Extraction

4.1 Introduction

Fundamentally, any digital image is composed of a finite number of pixels, expressed mathematically by a two dimensional function $f(x,y)$, where x and y represent the two spatial coordinates and its finite intensity at any image position is denoted by f. Objects in such images can be described by observing characteristics and patterns within them. Algorithms for determining and describing such patterns are known as feature extraction algorithms and the steps involved form the feature extraction process. It is believed that information contained in an individual's face are more discriminative along the horizontal direction than those vertically aligned, which is ideal for appearance-based techniques [126]. Due to its horizontal orientation, such facial features seem to remain undisturbed by any changes in illumination [126]. For experimenting, in this section a feature extraction technique is to be applied to ME facial images taken from spontaneous ME datasets CASMEII. It should be noted that these micro facial images have been captured in an environment with appropriate lighting hence, encoding facial micro textures using appearance-based methods seems realizable. Another advantage of using methods that fall within the appearance-based category is that providing emotion label information is sufficient for it to undergo a training process. This section of thesis explores the handcrafted extraction method which employs an appearance-based approach named local phase quantization (LPQ).

The LPQ extraction method was first introduced in [127-129] for extracting textures from blurred images. The use of the LPQ method was further seen for FE analysis in later years [88, 130-132]. In a comprehensive study by [133] it was found that most of the work for FE recognition reported exceptionally high recognition accuracy by using LPQ and its variants. The facial images considered in these experiments [88,130-132] for performing expression analysis contained macro expressions. Survey revealed that comparatively more work that employ LPQ method for macro expression [88,130-132] exists than ME [89][134]. The work in [134] employs local phase quantization on three orthogonal planes (LPQ-TOP) for AU detection for ME and is different from work in this thesis which focuses on recognition. In the work by [89] this method is employed for ME, but it focuses on designing cross database micro expression recognition (CDMER) rather than a straightforward MER system. In their work [89], instances of ME used during training and testing phases belonged to two different datasets. Hence images used to train a model were different from images used to test that model. An average recognition accuracy of 63.79% (decorrelation 0.1) and 64.05% (decorrelation 0) was achieved using LPQ-TOP on varying combinations of training and testing data taken from HS, NIR and VIS variations of the SMIC dataset. Similarly, using training images from CASMEII and testing on images from HS, NIR and VIS variations of SMIC and vice versa produced an average accuracy of 38.51% (decorrelation 0.1) and 41.83% (decorrelation 0). Inspired by the massive success of the LPQ based method for macro expression analysis and the CDMER framework, this thesis takes forward the usage of LPQ-TOP and test its suitability to perform as a micro facial feature extraction technique where both training and testing images belong to the same dataset (i.e., non-cross database environment). The performance of the extraction method on the chosen database is measured by analyzing the classification results obtained by employing features extracted by LPQ-TOP method. Thus, work described in this thesis does not consider designing a CDMER and focus remains on building a MER framework by employing the LPQ-TOP method.

In this chapter the preliminary investigations performed using the CASMEII database to test the suitability of employing LPQ-TOP and TIM for designing an automated MER system are presented and forms the first contribution for this thesis. Results obtained here also form the basis on which its application is extended in further Chapters.

The remainder of this chapter is organized as follows. Section 4.2 introduces the LPQ based approach particularly as a feature extraction technique. Preliminary results obtained by experimenting with this method using the CASMEII dataset and an initial analysis is presented in Section 4.3, followed by a summary of the work in Section 4.4.

The contribution made in this chapter has been published in Sharma, P., Coleman, S., Yogarajah, P. & Laurence, T. (2019). Micro expression classification accuracy assessment. IMVIP 2019: Irish Machine Vision & Image Processing, Technological University Dublin, Dublin, Ireland, August 28-30. doi:10.21427/kbny-0a41.

4.2 Local Phase Quantisation Method

The LPQ method is based on computing a discrete Fourier transform (DFT). The quantized phase of DFT is mathematically obtained for a given image patch, generally also referred to as a neighbourhood, to describe its underlying texture, known as phase quantization. Three primary building blocks of this method are phase information, blur insensitivity and histogram. Experiments in [127-129] show that if statistically independent samples are taken to perform quantization, then maximum information can be preserved, and de-correlation was introduced to achieve this. It is a well-established fact that the LPQ method draws its working principle from quantization of Fourier transforms [127]. Equations (4.1) to (4.5) describe the process when the phase computed for a DFT is invariant to blur. Similarly, equations (4.15) and (4.16) describe the LPQ operator. Given an original image $f(\mathrm{x})$, its corresponding observed image g(x) with spatial blurring, that is invariant spatially, can be represented using convolution, expressed by [127-129]:

$$\mathrm{g}(\mathbf{x}) = (f \ * h) \ (\mathbf{x}) \tag{4.1}$$

where $h(\mathbf{x})$ symbol represents a point spread function (PSF), * notation is used to denote two-dimensional convolution and $\mathbf{x}$ stands for a vector of coordinates, which is expressed by $[x,y]^T$. This expression further can be represented in the Fourier domain as [127-129]:

$$G(\mathbf{u}) = F(\mathbf{u}).\ H(\mathbf{u}) \tag{4.2}$$

Here, the symbol $G(\mathbf{u})$ represents the DFT of the blurred image, $F(\mathbf{u})$ signifies the DFT of the original image and $H(\mathbf{u})$ serves as the DFT of the PSF. Furthermore, $\mathbf{u}$ is considered to be a vector of coordinates, expressed as $[u,v]^T$. Since the focus of this method is to obtain phase information, the equation can be expressed in terms of a sum [127-129]:

$$\angle G(\mathbf{u}) = \angle F(\mathbf{u}) + \angle H(\mathbf{u}) \qquad (4.3)$$

In equation (4.3), the three symbols $\angle G(\mathbf{u})$, $\angle F(\mathbf{u})$ and $\angle H(\mathbf{u})$ represent the corresponding phase angle for $G(\mathbf{u})$, $F(\mathbf{u})$ and $H(\mathbf{u})$. Presuming $h(\mathbf{x})$ is centrally symmetric, i.e., $h(\mathbf{x})=h(-\mathbf{x})$, the Fourier transform is always expected to be a real value, consequently this causes the phase to be reduced to a two-valued function represented as[127-129]:

$$\angle H(\mathbf{u}) = \begin{cases} O & if \quad H(\mathbf{u}) \geq 0 \\ \pi & if \quad H(\mathbf{u}) < 0 \end{cases} \qquad (4.4)$$

Ultimately the equation then becomes:

$$\angle G(\mathbf{u}) = \angle F(\mathbf{u}) \text{ for all } \angle H(\mathbf{u}) \geq 0 \qquad (4.5)$$

Three approaches were proposed by [128] for computing the local phase information namely, short term Fourier transform (STFT), Gabor filters and least square filters. Following the best results obtained with STFT to compute the phase information in [128], in this thesis a STFT based approach is selected.

4.2.1. Short Term Fourier Transform

A short-term Fourier transform (STFT) for an image $f(\mathrm{x})$ can be computed by taking an image patch of size m x m denoted as $f_{\mathrm{x}}(\mathbf{y})$. These image patches $f_{\mathrm{x}}(\mathbf{y})$ are mathematically defined by basis function using equation (4.6) [128]:

$$\phi_{\mathbf{u}}^{\mathcal{F}}(\mathrm{y}) = e^{-j2\pi\mathbf{u}^{T}\mathbf{y}} \qquad (4.6)$$

The STFT can now be computed using 2D convolution * given by:

$$f(\mathrm{x}) * \phi_{\mathbf{u}}^{\mathcal{F}}(\mathrm{y}) \qquad (4.7)$$

Here, STFT is computed for four low frequencies, **u** = {$\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$}. Additionally, for rows and columns, computations can be performed using one dimension because the basis functions are easily separable [128].

Elaboratively, to examine the phase, LQP basically considers a local M x M neighbourhood represented by the notation $\mathcal{N}_x$ at each pixel position, **x,** for a given image *f*(**x**). For computing a STFT the following equation is used [127-129]:

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y} \in \mathrm{N}_{\mathbf{x}}} f(\mathbf{x} - \mathbf{y}) \;\; e^{-\mathrm{j}2\pi \mathbf{u}^T \mathbf{y}} \;=\mathbf{w}_{\mathrm{u}}^{T} \mathbf{f}_{\mathrm{x}} \tag{4.8}$$

In equation (4.8) $\mathbf{w}_{\mathrm{u}}$ denotes basis vector computed at **u** frequency. For the neighbourhood $\mathcal{N}_x$, its pixel information is stored in a vector denoted by $\mathbf{f}_x$. Using 1-D convolution consecutively for rows and columns, the STFT is evaluated for all positions in an input image. It computes local Fourier coefficients at each pixel location for four frequency points: $\mathbf{u}_1=[a\ ,\ 0]^T$, $\mathbf{u}_2=[0\ ,\ a]^T$, $\mathbf{u}_3=[a\ ,\ a]^T$, $\mathbf{u}_4=[a\ ,\ -a]^T$ (see Figure 4.4). Here, "*a"* represents a scalar frequency that satisfies $H(\mathbf{u})\geq 0$.

4.2.2. Decorrelation

As noted by [128], the scalar quantization process discussed in Section 4.2.1 can be achieved in a theoretical scenario where the coefficients undergoing quantization are not co-dependent statistically. However, while experimenting, situations may arise where such coefficients are correlated and employing scalar quantization is not recommended. Therefore, to have a better approach for such cases utilizing vector quantization seems more suitable. Alternatively, decorrelating the data prior to performing quantization can also be relevant [128]. At this stage the real and imaginary components of the frequency components need to be separated. This ultimately results in a vector for each pixel position that can be expressed as in equation (4.9), where *Re* represents the real and *Im* represents the imaginary parts respectively [127-129].

$$\mathbf{F}_{\mathrm{x}} = ([Re\{\, F(\mathbf{u}_1,\mathbf{x})\},Im\{F(\mathbf{u_1},\mathbf{x})\}],\ldots.$$
$$\ldots, [Re\ \{\, F(\mathbf{u_4},\mathbf{x})\},Im\{F(\mathbf{u_4},\mathbf{x})\}])^{\mathrm{T}} \qquad (4.9)$$

The real components are then concatenated into a vector and the step is also repeated for imaginary components to generate its vector. Combining equation (4.8) and (4.9), the final vector representation is expressed as:

$$\mathbf{F}_{\mathrm{x}} = \mathbf{W}\mathbf{f}_{\mathrm{x}} \qquad (4.10)$$

For an image patch *f*(**x**), σ^2 denotes variance between pixels that are adjacent to one another. Therefore, the resultant matrix representing covariance for all the data samples *M*, belonging to the neighborhood $\mathcal{N}_x$ is given by (4.11) [127-129]:

$$\mathbf{C} = \begin{bmatrix} 1 & \sigma_{1,2} & \cdots & \sigma_{1,M} \\ \sigma_{2,1} & 1 & \cdots & \sigma_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M,1} & \sigma_{M,2} & \cdots & 1 \end{bmatrix} \qquad (4.11)$$

Assuming linear dependence for $\mathbf{F}_{\mathrm{x}}$, the covariance matrix is measured as [128]:

$$\mathbf{D} = \mathbf{\Phi}\ \mathbf{C}\mathbf{\Phi}^{T} \qquad (4.12)$$

In order to decorrelate these vectors whitening transformation is employed given by equation (4.13) and equation (4.14) is the singular value decomposition employed to compute orthonormal matrix **V** [127-129]. Whitening transformation follows a linear transformation model which essentially converts a given vector and covariance matrix into a set of new vector having an identity matrix as its covariance matrix.

$$\mathbf{G}_{\mathrm{x}} = \mathbf{V}^{T}\mathbf{F}_{\mathrm{x}} \tag{4.13}$$

$$\mathbf{D} = \mathbf{U}\,\mathbf{\Sigma}\mathbf{V}^{T} \tag{4.14}$$

Resultant vectors are obtained by computing $\mathbf{G}_{\mathrm{x}}$ for every location for the given image, which then undergoes a quantization process.

4.2.3. Quantisation

The uncorrelated samples obtained from the steps discussed above are then quantised. To do so, binary scalar quantizer shown in equation (4.15) is used and quantizes the signs of real and imaginary parts of the coefficients obtained in the previous step. Here, f_j represents the j^{th} component for a given vector $F(\mathbf{x})$ [127-129]:

$$q_j = \begin{cases} 1, & if \quad f_j \ \geq 0 \\ 0, & if \quad f_j \ < 0 \end{cases} \tag{4.15}$$

The resultant eight-bit binary coefficients are represented in the form of integers using the binary coding technique using equation (4.16), where the integer values lie within the range of 0-255 given by [127-129]:

$$b = \sum\nolimits_{j=1}^{8} q_j 2^{j-1} \tag{4.16}$$

The symbol q_j represents the quantized vector obtained from equation (4.15) and (4.16) for the j^{th} component. Such integer values are composed for all the image positions and stacked into a histogram. This histogram is essentially a 256-dimensional vector containing features extracted from an input image (see Figure 4.1).

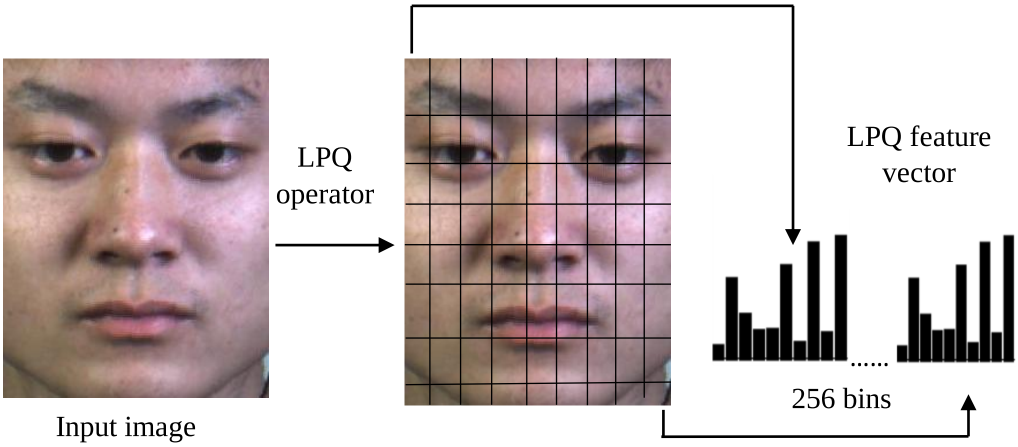


Figure 4.1. Extracting feature from each block and concatenating them into single feature vector [88].

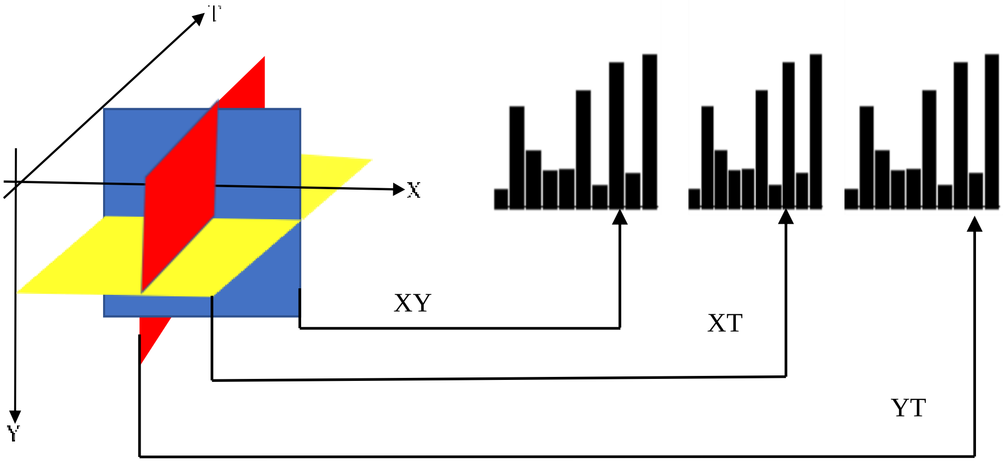


Figure 4.2. Concatenated histogram obtained from three orthogonal planes (XY, XT & YT) [88].

4.2.4. LPQ-TOP Method

To extract features, as a first step the whole facial image is divided into non overlapping blocks, then a feature vector is computed for each of these blocks using the LPQ approach. A graphical illustration of this step in a simplest form is presented in Figure 4.1. The LPQ method is ideal when working with static images however, for implementation in this thesis, including the temporal sequence is essential to examine ME dynamics hence, the extension of LPQ to include the time domain, denoted as LPQ-TOP is employed. The spatial domain data are provided by the XY plane whereas information relating to the temporal domain is provided by the XT and YT planes (see Figure 4.2).

The descriptors extracted from each of these planes are represented in the form of a histogram where each of these histograms are concatenated to form a single feature vector. A simple illustration of this procedure for extraction and concatenation of feature vectors is presented in Figure 4.3. Since features are extracted independently along all three orthogonal planes, i.e., XY, XT and YT, it produces 256 x 3 bins per volume, measured in space-time. The overall procedure followed for obtaining features using LPQ, along with the selection of the neighbourhood described in this section is graphically demonstrated in a step-by-step manner in Figure 4.4 and Figure 4.5. It effectively also demonstrates the use of the equations outlined in this section as and when applicable. The extracted facial micro features in this stage form the foundation for recognition at a later stage. In initial experiments conducted in this thesis this technique was employed to extract facial micro features from image sequences taken from the CASMEII dataset.

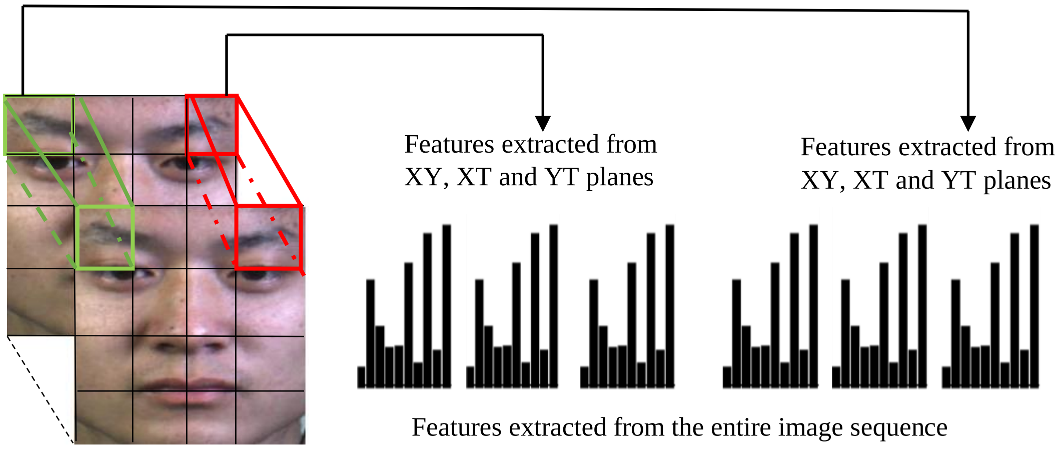


Figure 4.3. Features extracted from each block representing a sequence, concatenated to form feature vector [88].

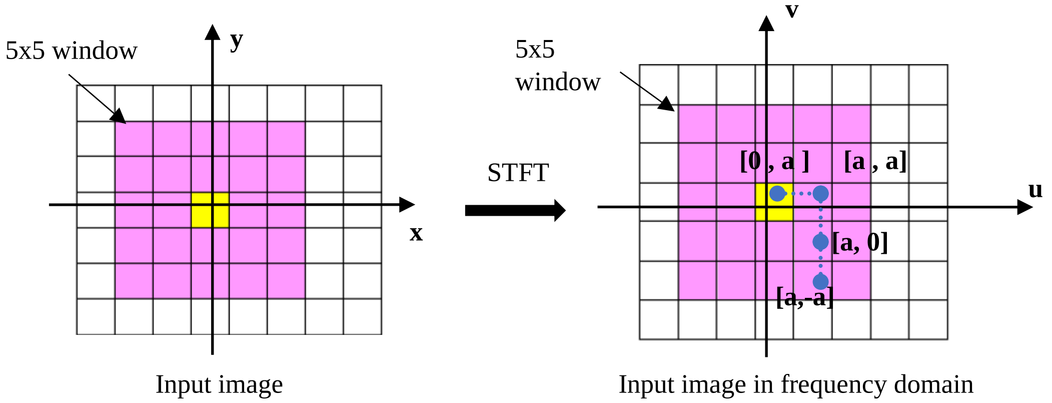


Figure 4.4. LPQ Fourier frequencies and M x N neighbourhood [131].

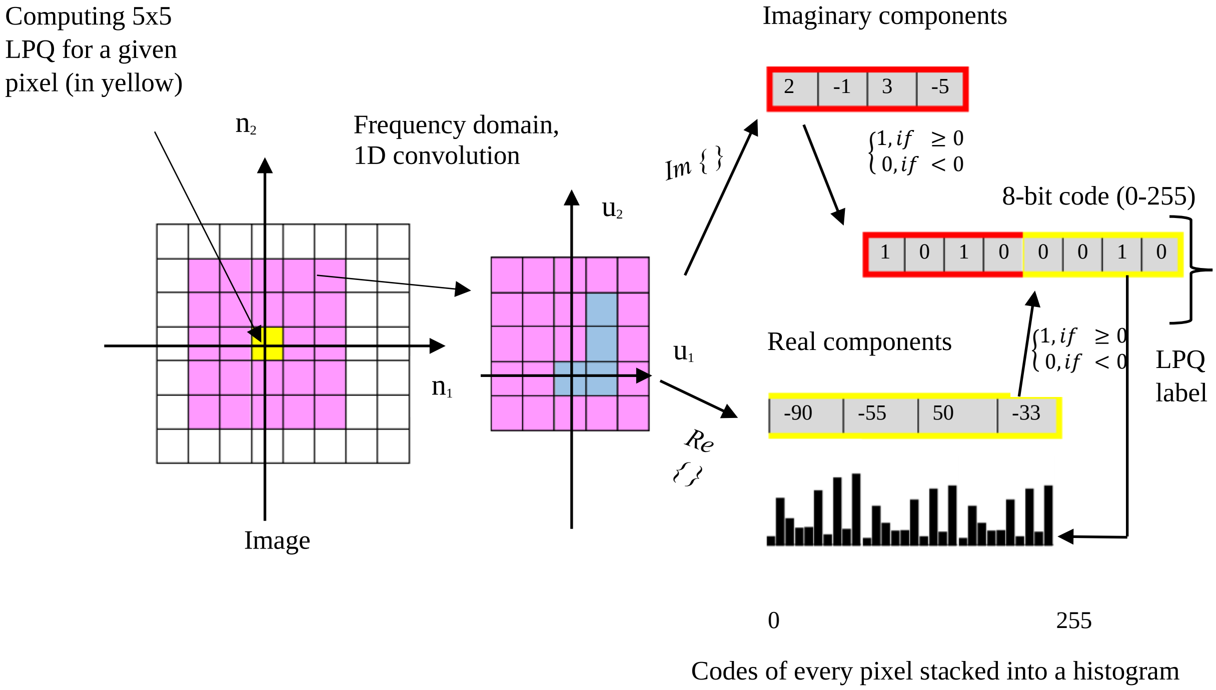


Figure 4.5. Step-by-step layout for computing LPQ over facial image [131].

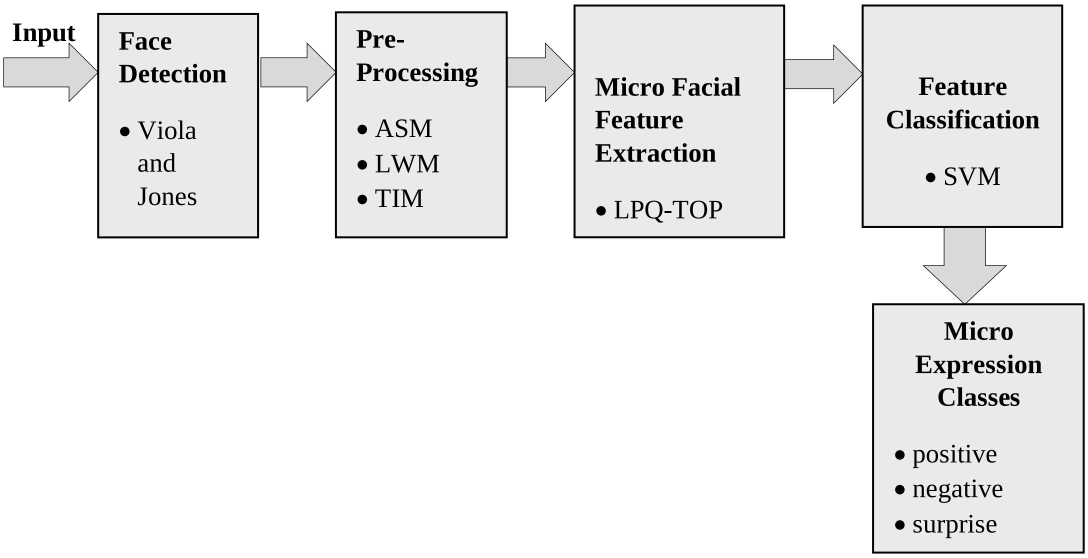


Figure 4.6. Bare bone structure for micro expression recognition system.

4.3 Experiments and Results

The general pipeline used for performing MER consists of face detection, pre-processing, feature extraction and feature classification as demonstrated in Figure 4.6. In this section the details of the experiments performed using this pipeline are discussed. The experimental setup and parameter specifications used along with the results obtained and its discussion is also presented in this section.

4.3.1. Face Detection and Pre-Processing

The experiments have been conducted using the CASMEII spontaneous ME dataset containing faces captured with a resolution of 280 x 340 pixels [47]. The raw dataset contains facial data, unaffected by flickering or illumination issues, which can be readily used. For detecting a face from the input, Viola, and Jones (V & J) [71] detector as described in Section 2.4.1 is used.

Following the work in [40], for face alignment this thesis uses ASM [72] and LWM [76]. Total 68 landmarks are identified using ASM which are then used for normalizing images further. The ME videos in the dataset have non-uniform frames, therefore taking inspiration from [40,89], this thesis applies TIM [77] to interpolate frames and maintain frame uniformity across the input data. The working principle of TIM has already been discussed in section 2.4.1.

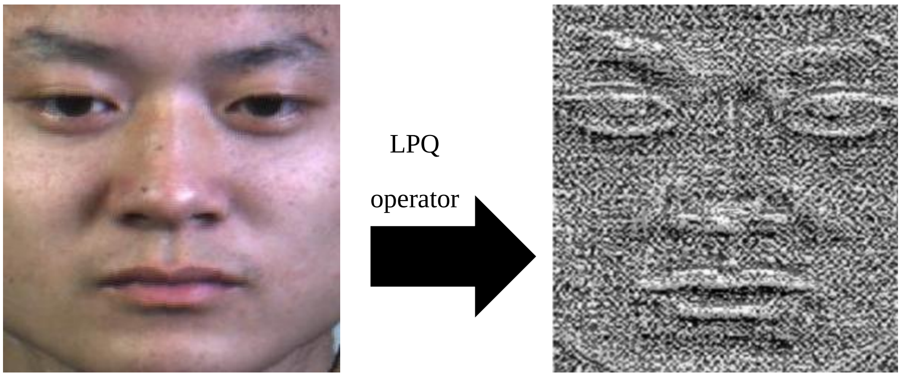
Originally the data in this dataset were labelled into five classes i.e., happiness, disgust, surprise, repression, and others. In experiments for this thesis three classes namely positive, negative and surprise were considered. The motive behind choosing three classes of expressions is that the key focus of this research is to be successful as a MER system particularly targeting autistic faces in future. This type of faces tends to manifest very minimal expressions on their face therefore the concept of classifying ME into three classes as positive, negative and surprise seemed more realizable. Also, in the five classes, fear and sad expressions were not used for baseline evaluation however in this thesis these expressions are also utilized. During implementation in this thesis, happy expressions have been given a positive label and surprise expressions are left unchanged. Expressions like fear, sad and disgust are categorized under a negative label. Samples originally labelled as others in this dataset are a mixture of various expressions due to which input expressions have a high chance of being misclassified as others during classification. This type of interference within the classification algorithm can diminish overall performance. To avoid this and to preserve inter class discriminative characteristics of input samples, expressions labelled others have not been considered for this initial experiment. Additionally, expression labelled as repressions also remain unused for this thesis in current experiment to align the expression labels as close as possible to autism individuals.

4.3.2. Micro Facial Feature Extraction and Classification

Experimentally, features are extracted using the LPQ-TOP method described in Section 4.2. During implementation, a neighbourhood size of 5x5 was used. The feature vector obtained thereafter is passed to the classification algorithm. For conducting experiments, SVM [105] is used to perform classification of these data into three classes. The classification accuracy is computed as follows:

$$A = \frac{N_C}{N} \times 100 \tag{4.17}$$

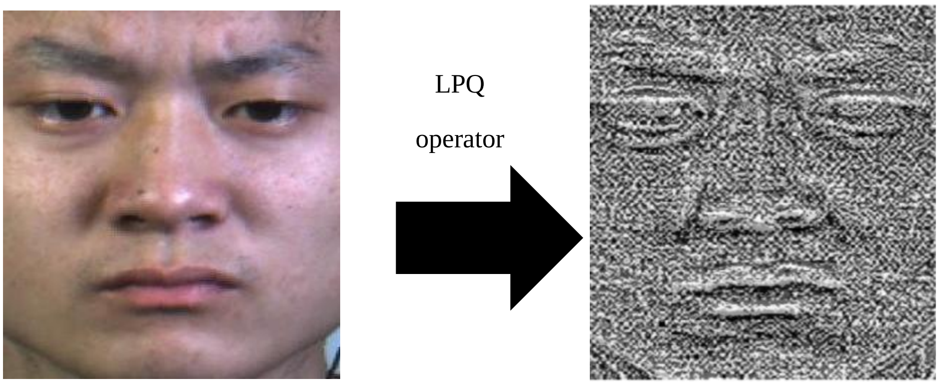
In equation (4.17), ME images correctly classified are denoted by N_C , N denotes the total number of ME images used and A is the rate of accuracy obtained. A total of 130 CASMEII images have been used with 33 positive, 25 surprise and 72 negative labels. Experimentally, [89] found that setting decorrelation to zero produced the best results, hence this section uses the same parameter for the implementation of LPQ-TOP with the XY, XT and YT planes set to [5,5,5].



Happy expression (Input image)
(a)

LPQ representation
(b)

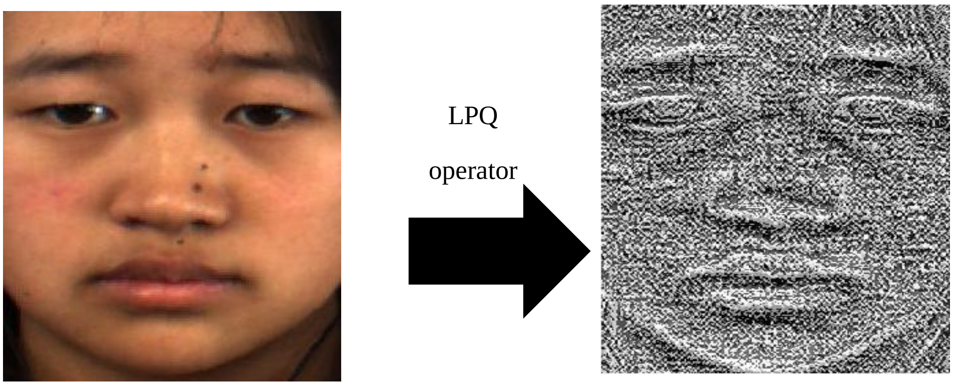
Figure 4.7. Instance of LPQ representation on XY plane, derived for positive label.



(a)
Disgust expression (Input image)

(b)
LPQ representation

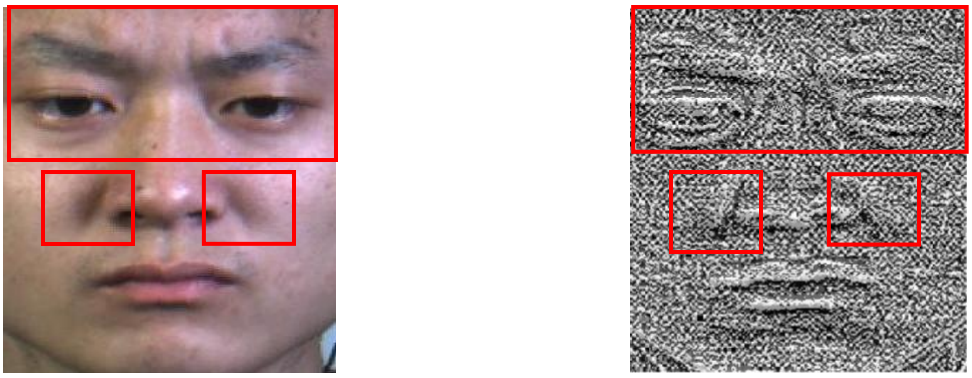
Figure 4.8. Instance of LPQ representation on XY plane derived for negative label.



Surprise expression (Input image)
(a)

LPQ representation
(b)

Figure 4.9. Instance of LPQ representation on XY plane derived for surprise label.



Input image
(a)

Texture patterns captured by LPQ operator
for negative label (disgust expression)
(b)

Figure 4.10. Highlighting the texture patterns captured by LPQ descriptor.

4.3.3. Results and Discussion

An instance of the LPQ representation obtained for a static image of a positive label (happy expression) in experiments performed for this thesis is demonstrated in Figure 4.7(b). It can be observed that the original expression (see Figure 4.7.(a)) is very subtle, yet LPQ was successful in extracting its relevant features thereby achieving an effective description of the expression (see Figure 4.7.(b)). A similar observation can be spotted for a negative label (disgust expression) from its original and LPQ representations demonstrated in Figure 4.8. For the surprise label, Figure 4.9 demonstrates a successful LPQ representation derived from its subtle original image. From these three figures (Figure 4.7, Figure 4.8, and Figure 4.9) it can be observed that LPQ descriptors were successfully implemented. It was effective in characterizing facial attributes required for identifying the ME that appear on the face. Its strong discriminative ability can also be visualized from these graphical results which demonstrates effective depiction of spatial texture patterns. For instance, the texture patterns for an expression captured by the LPQ operator, highlighted in red in Figure 4.10, is clearly visible in the processed image and can easily be perceived as a disgust expression. Hence the relevance of the LPQ approach for ME analysis has been established experimentally.

Table 4.1. Micro expression recognition results on CASMEII dataset with varying combinations of orthogonal planes.

	LPQ	**LPQ- XTYT**	**LPQ -XT**	**LPQ-YT**	**LPQ-TOP**
Accuracy (%)	53.57	61.11	60.06	60.89	61.16

For extracting features, the LPQ method was tested with varying combination of planes, the results obtained for each combination is presented in Table 4.1. From these results it is evident that by including all three planes, it yielded best recognition performance in comparison to other combinations of planes. The second-best performance was achieved by employing XTYT planes which was lower by a small value of 0.05% compared with the previous combination. The next best performance was yielded by using YT plane followed by XT plane. The lowest performance was obtained when time domain was excluded with an accuracy which is 7.59% lower than when all planes were employed. This clearly shows the significance of incorporating time domain for MER.

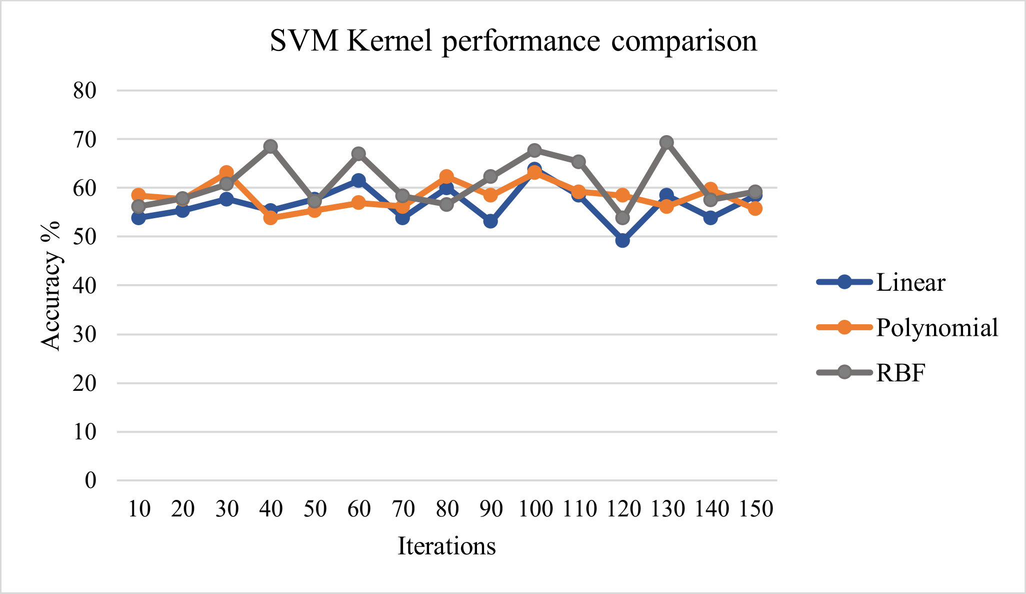


Figure 4.11. Performance comparison of SVM kernel on features extracted from CASMEII using LPQ-TOP.

The available datasets were divided into training and testing sets in a 70:30 ratio. SVM was employed for training and testing the data with three kernel functions, namely linear, RBF and polynomial. Parameters for each of these kernels were searched using grid search, followed by k-fold cross validation. A SVM with multi-class classification using a one-vs-all strategy was utilized. The penalty coefficient C was set between the search range of 0.1 to 1000 to obtain the best penalty value. Similarly, for the γ parameter, the search range was set between 0 to 1 with interval of 0.1. The coefficient r was set to zero, and degree was set between 1 to

10. Performance of each of these kernels was recorded during the experiments and calculated repeatedly over several iterations (150); the results are illustrated in Figure 4.11. From this figure it is evident that the RBF kernel produced the best performance over several iterations. Though performance of the polynomial kernel exceeded that of RBF in some instances, it could not outperform the RBF in overall. Performance of the linear kernel was lower than the RBF in the majority of the iterations with a similar trend observed when compared with the polynomial kernel. The lowest performance recorded for the linear kernel was 49.23% and

53.84% for both polynomial and RBF. Peak performance recorded for the linear kernel was 63.8%, whereas for the polynomial and RBF it was 63.07% and 69.23% respectively. Thus, it can be seen that the peak performance of the linear kernel was slightly higher than that of the polynomial kernel. However, when computing the overall performance of these kernels, the RBF surpassed other kernels with 61.16%, evidenced by the average recognition rate which was 56.7% for the linear kernel, whereas using the polynomial it was 58.3%. Therefore, SVM with the RBF kernel was able to produce the best performance for LPQ-TOP on CASMEII database. Performance metrics like precision, recall and F1 score obtained from the experiments is presented in Table 4.2.

Precision is a measure that reveals the proportion of samples correctly predicted as true positives compared with total positive instances predicted by the model. The metric can be computed by applying the equation given below [143].

$$Precision = \frac{TP}{(TP + FP)} \ \ ; \ 0 < Precision < 1 \tag{4.18}$$

The recall measure helps to identify the number of positive instances missed by the model during its prediction stage. This estimation is computed using mathematical equation given below [143].

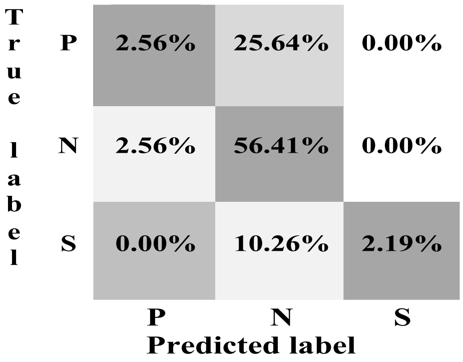
$$Recall = \frac{TP}{(TP + FN)};\ 0 < Recall < 1 \tag{4.19}$$

The F1 score is computed using harmonic mean of these both precision and recall values, mathematically expressed using an equation given below [143].

$$\frac{2}{\left(\frac{1}{Precision} + \frac{1}{Recall}\right)} \tag{4.20}$$

Table 4.2. Performance metrics obtained on CASMEII.

Performance Metric	Value
Precision	0.63
Recall	0.62
Accuracy%	61.16
F1 Score	0.53



True label	P	N	S
P	2.56%	25.64%	0.00%
N	2.56%	56.41%	0.00%
S	0.00%	10.26%	2.19%
	Predicted label		

Figure 4.12. Confusion matrix obtained for CASMEII data.

From the performance metrics obtained using the CASMEII database, presented in Table 4.2, it can be seen that using the proposed pipeline precision of 0.63 was obtained which depicts the model had a low false positive rate, the recall value was 0.62 and F1 score was 0.53. Observing the confusion matrix in Figure 4.12 it can be clearly seen that negative samples had higher classification accuracy, compared to positive and surprise samples which may be due to the data distribution that is more bias towards negative class labels. Moreover, among the misclassified samples, the majority of them were wrongly classified as negative samples.

Table 4.3. Accuracy % comparison for CASMEII.

Feature Extraction Method	Accuracy %	Class Label used (ours)		Expressions used (ours)
LPQ-TOP	**61.16 (ours)**	3	positive	Happy
LBP-TOP	55.87 [40]		negative	Sad, Fear, Disgust,
HOG-TOP	57.49 [40]		surprise	Surprise
HIGO -TOP	57.09 [40]			

Exceptionally less misclassified data were wrongly classified as surprise. The accuracy values reported in Table 4.3 give an insight on the significance of this method for feature extraction using the CASMEII dataset. The value obtained highlights that LPQ-TOP was able to perform well on this dataset by achieving an accuracy rate of 61.16% for three class emotion classification. In this initial experiment, LPQ-TOP based recognition approach has achieved 5.29% higher accuracy than LBP-TOP reported by [40]. Similarly, this accuracy is 3.67% higher than HOG-TOP and 4.07% higher than HIGO-TOP [40]. Significantly less work could be found using the CASME II dataset for classifying ME into three classes of emotion hence comparison of accuracy delivered by the proposed approach for three class emotion has been established wherever applicable. Meanwhile, less work that have employed LPQ-TOP using CASMEII for MER could be found due to which appropriate comparison could not be established. However, the performance of LPQ-TOP using CASMEII is compared with those in [40] for SMIC using LBP-TOP, HOG-TOP, and HIGO-TOP.

Table 4.4. Accuracy % Comparison for CASMEII & SMIC.

Feature Extraction Method	CASMEII	SMIC-HS	SMIC-VIS	SMIC-NIR	SMIC-subHS
LPQ-TOP	**61.16 (ours)**	-	-	-	-
LBP-TOP	55.87 [40]	57.93 [40]	70.42 [40]	64.79 [40]	77.46 [40]
HOG-TOP	57.49 [40]	57.93 [40]	71.83 [40]	63.38 [40]	80.28 [40]
HIGO-TOP	57.09 [40]	65.24 [40]	76.06 [40]	59.15 [40]	80.28 [40]

This comparison of the preliminary result with those obtained by [40] using variations of the SMIC dataset is presented in Table 4.4. According to the values listed in the Table 4.4 one can see that the performance of LPQ-TOP when compared with results obtained in [40] using SMIC datasets are higher in some cases. For instance, from experiments conducted in this thesis the accuracy is 3.23% higher than LBP-TOP and HOG-TOP when compared with the performance recorded across SMIC-HS. This accuracy is also closer to those obtained using LBP-TOP and HOG-TOP and 2.01% higher than HIGO-TOP on SMIC-NIR dataset. This comprehensive analysis of the preliminary results helps to visualize successful use of LPQ-TOP in a MER task. It should be noted that performance recorded in [40] for HOG-TOP and HIGO-TOP is much higher when used on the SMIC-VIS and SMICsub-HS datasets. One reason behind this exceptional performance of both the methods on SMICsub-HS could be due to its smaller sample size as it is collected from only eight participants containing more evenly distributed data. Nevertheless, in experiments conducted for this thesis the features extracted using LPQ-TOP classified using SVM have produced the highest classification accuracy compared with other extraction methods on CASMEII. It must be mentioned here that employing similar approach in a cross-database environment the highest accuracy obtained on CASMEII in [89] is 48.46% which is quite lower compared to the results obtained in this thesis in a non-cross database environment.

Examining these initial results, it is sufficient to claim that performance of LPQ-TOP technique is as competitive as any other traditional feature extraction methods employed so far for micro facial images. Likewise, its usefulness is established amidst these early experiments and intermediate results obtained. Hence, this technique seems to have some potential for additional investigation, as such exploring it further alongside other methods will be considered with more elaborative experiments in next chapters.

4.4 Summary

To summarize, the current research work was devoted in assessing the significance of the LPQ-TOP method for facial micro feature extraction. It also attempts to determine the method’s effect when combined with supervised classification. In this chapter the focus remains on the applicability of the LPQ-TOP method for a MER system. Experiments are conducted using the CASMEII dataset to establish this. A comprehensive analysis based on results obtained through initial experiments is provided to highlight its importance for expression analysis. Absence of similar work for MER made the entire process of tuning the LPQ algorithm time consuming and more challenging. Through careful examination it is found that the method was successful in describing micro facial features thereby empowering the classification process. Results demonstrate that the performance of this method is quite appreciable when compared with other binary and gradient based approaches. Despite the fact that the LPQ method is computationally slightly higher than the local binary method, the positive outcome indicates it can be exploited further and has scope for improvement. Deeper investigation may be needed to maximise its usage. Since work employing CASMEII has not considered three class classification, it was difficult to make direct comparison with other works that use it. At the time of writing this thesis and to the best of author’s knowledge, the only work that has employed the LPQ based method for recognition of ME without an AU, is specifically for designing CDMER. Therefore, this chapter proposes to extend its use for building MER pipeline for a non-cross database scenario along with TIM. By further employing phase quantization approach for the next few experiments, it aims to draw clearer insight on this method particularly for ME domain. To achieve this, the thesis further exploits the LPQ-TOP and TIM into the recognition pipeline in Chapter 5 and provides a comprehensive analysis of the performance obtained on several databases.

Chapter 5

Amplifying Spontaneous Facial Micro Expression to Achieve Recognition Boost

5.1 Introduction

Exceptionally short duration and faint intensity of muscle progression in ME naturally limits the recognition capability of algorithms. This has been a perennial challenge which needs to be meticulously dealt with. Chapter 2 revealed that previously more work employing image sequences for analysing facial ME were prevalent than video sequences. To exploit the full potential of such expressions, designing video-based MER systems have emerged rapidly in recent years. Such approaches employing videos can provide the groundwork for building real time ME analysis. Finding ME accurately in a video sequence is still a challenge since it is noticeable in few frames due to its brief time span. The research direction of this section is towards approaches that aid enhancement of existing facial micro features to compensate the elusive nature of these minute expressions. Thus, this section surveys existing video magnification-based MER systems extensively.

The concept of magnifying subtle changes in videos that are difficult to be perceived by naked eye was first introduced in [79]. The magnification was applied for observing blood flow as well as motion with small scale, and the method was named Eulerian video magnification (EVM). The method was tested in some dynamic environment where changes are almost invisible to naked eyes like the heartbeat of a new-born, guitar string vibrations, blood flow on the wrist etc., and was successful in revealing such indistinct motion which usually goes undetected by human eyes. The same technique was introduced for ME by [49] to amplify the muscle movement in videos. Three filters i.e., Butterworth, ideal and second order infinite impulse response (IIR) were examined to identify the best choice. Due to a narrow range of bandpass, the ideal filter performed poorly on ME videos. Using a second order IIR filter with the magnification factor set to 20 and spatial frequency set to 16, the best performance was obtained. The Butterworth filter performed better than the ideal filter but could not outperform the IIR filter during experimentation by [49]. Using the IIR filter, the magnification approach was tested on the CASMEII dataset and achieved recognition accuracy of 75.3%[49]. Using HIGO for extracting EVM magnified features on the CASMEII dataset, [40] achieved an impressive recognition performance of 78.14%. Similar to the work discussed earlier, in experiments by [40] too, an IIR filter with a wider bandpass was employed. The method was tested with varying magnification factors to identify the best parameter combinations on various datasets including three variations of the SMIC dataset. On SMIC-HS the method achieved accuracy of 75%, similarly on SMIC-VIS it was 83.10% and 71.83% on SMIC-NIR. Utilizing a hybrid approach in [135], where motion magnified with EVM was extracted using spatio temporal texture map (STTM), almost 5% more accuracy than [49] was achieved. Combining TIM and magnification as a single component in [78], the framework recorded accuracy of 70.85% on CASMEII. To build a layout for compound MER system, [136] magnified basic ME using EVM which were then used for producing compound ME images. Due to muscle articulation achieved using magnification, creating synthesized images for compound expressions became much easier. To deal with the noise present in the magnified ME clips, [136] employed the Emotion Avatar Image (EAI) technique.

Evidently the video magnification approach seems to complement facial features, therefore extending previous work described in Chapter 4, this section examines the influence of magnification on phase quantization method. Keeping the previous framework (refer to Figure 4.6) intact this section introduces EVM taken from [79] to achieve ME amplification. Also, for a thorough investigation, the proposed approach is tested on seven different spontaneous ME datasets and hence demonstrate the robustness of the approach.

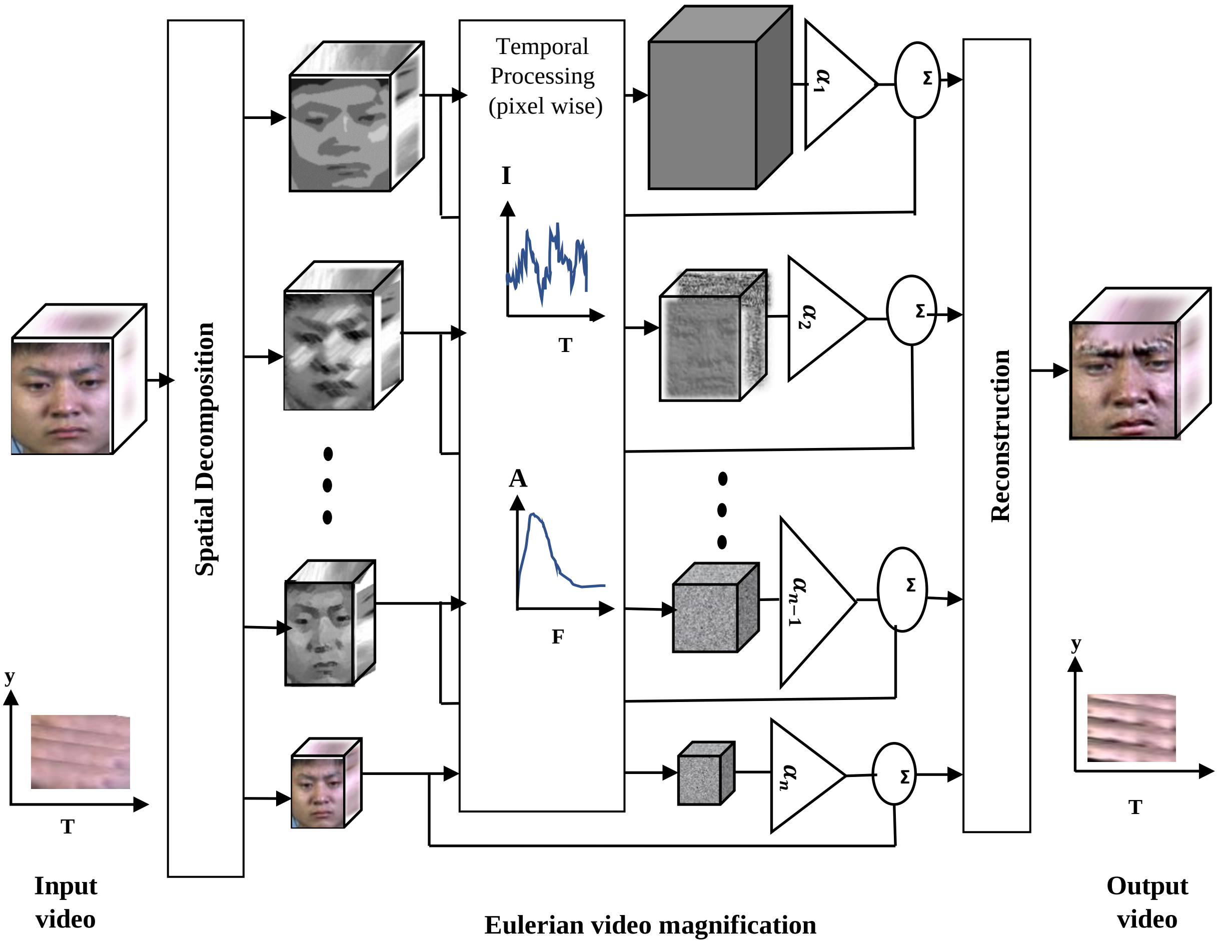
The remainder of this chapter is organized as follows. In Section 5.2 the video magnification technique employed for experiments is discussed. A brief description of the proposed approach is outlined in Section 5.3. Experimental results obtained along with its comprehensive analysis is presented in Section 5.4 and finally, a summary of the work is presented in Section 5.5.

A portion of experiments and results presented in this chapter has been published in Sharma,P., Coleman, S., Yogarajah, P. , Taggart, L. and Samarasinghe P., (2021) "Magnifying Spontaneous Facial Micro Expressions for Improved Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 7930-7936, doi: 10.1109/ICPR48806.2021.9412585.

5.2 Eulerian Video Magnification

Given an input video, the EVM technique [79] operates on both the spatial and temporal domain. In simple terms, the video frames first undergo decomposition in the spatial domain, next the filtering process is applied in the temporal domain. To perform magnification, it takes any pixel location and examines color values in time series. If any variation is observed in the temporal series for a chosen frequency band, then the amplification process is initiated. Using a linear approximation approach that assumes steadiness in brightness, the EVM approach is effective in magnifying motion without explicitly tracking it. Since magnification is achieved without relying on motion estimation, this reduces the overall computational complexity [79]. Also, another major advantage of this approach over a Lagrangian based method is that amplification of both spatial based motion as well as temporal changes can be achieved through this single framework. The method boasts of maintaining a rationale between spatial and temporal video attributes, achieved due to the uniform filtering process applied over its pixels. Due to a straightforward implementation of the magnification algorithm, the method is easy to use and allows users to directly regulate various parameters appropriately.

The overall process can be described in three phases, spatial decomposition, temporal processing for amplification, and reconstruction as illustrated in Figure 5.1. During spatial decomposition the sequences in an input video are broken down in terms of several spatial frequency bands. This is followed by the next phase where temporal processing is carried out on each of these extracted spatial bands. Here, a band pass filter is applied in order to derive the appropriate frequency band. By multiplying these extracted signals with a magnification factor denoted as α, the final amplified video is reconstructed. The steps followed to perform the magnification process using EVM framework are detailed in Section 5.2.1.



Note:
I : Intensity A : Amplitude F : Frequency (Hz) T : Time $\boldsymbol{\alpha}$: Magnification factor

Figure 5.1. Illustration of Eulerian video magnification framework [79].

5.2.1 Eulerian Motion Magnification

In order to realize motion magnification, Eulerian motion magnification (EMM) utilizes a series expansion approach based on Taylor series of the 1st order [79]. There exists an association between motion magnification and temporal processing which can be described using motion translation in a one-dimensional signal. Given an amplification factor α the signal to be synthesized is expressed as [79]:

$$\hat{I}\,(x,t) = f\big(x + (1+\alpha)\delta(t)\big) \tag{5.1}$$

In equation (5.1), for a given pixel position *x* at time *t*, its intensity is represented as $I(x,t)$. As an effect of translation these intensities are expressed using displacement $\delta(t)$; where, $I(x,t) \ = \ f$ (x | $\delta(t)$) and $I(x,0) = \ f(x)$. Here, α represents an amplification factor applied to unify these signals. Considering time, 't' and $f\ (x + \ \delta(t))$ for a given image at position *x*, applying first order Taylor series expansion the equation is given as [79]:

$$I(x,t) \approx f(x) + \delta(t)\frac{\partial f(x)}{\partial x} \tag{5.2}$$

For a motion signal $\delta(t)$, assuming it does not lie beyond the frequency range of the temporal bandpass filter, then this results in an expression [79]:

$$B(x,t) = \delta(t)\frac{\partial f(x)}{\partial x} \tag{5.3}$$

In equation (5.3), the effect of applying a broadband temporal bandpass filter to *I*(*x*, *t*) is denoted by *B*(*x*, *t*). The next step is to compute a new signal $\tilde{I}(x,t)$, by adding the amplified bandpass signal factored by α, back to *I* (*x*, *t*) expressed as [79]:

$$\tilde{I}(x,t) = I\ (x,t) + \alpha\ B(x,t) \tag{5.4}$$

Substituting the values of $I\ (x,t)$ and $B(x,t)$ in equation (5.4) with values taken from equation (5.2) and (5.3), and simplifying these expressions the equation becomes:

$$\tilde{I}(x,t) \approx f(x) + (1+\alpha)\ \delta(t)\frac{\partial f(x)}{\partial x} \tag{5.5}$$

Extending the influence of Taylor's series expansion for amplification with larger deviation denoted by $(1+\alpha)\ \delta(t)$, the expression transforms into [79]:

$$\tilde{I}(x,t) \approx f\big(x + (1+\alpha)\delta(t)\big) \tag{5.6}$$

For situations where $\delta(t)$ lies outside the frequency range of passband, application of broadband temporal band pass filter to I (x, t) will produce a resultant bandpass signal given by [79]:

$$B(x,t) = \sum_k \gamma k \delta_k(t) \frac{\partial f(x)}{\partial x} \tag{5.7}$$

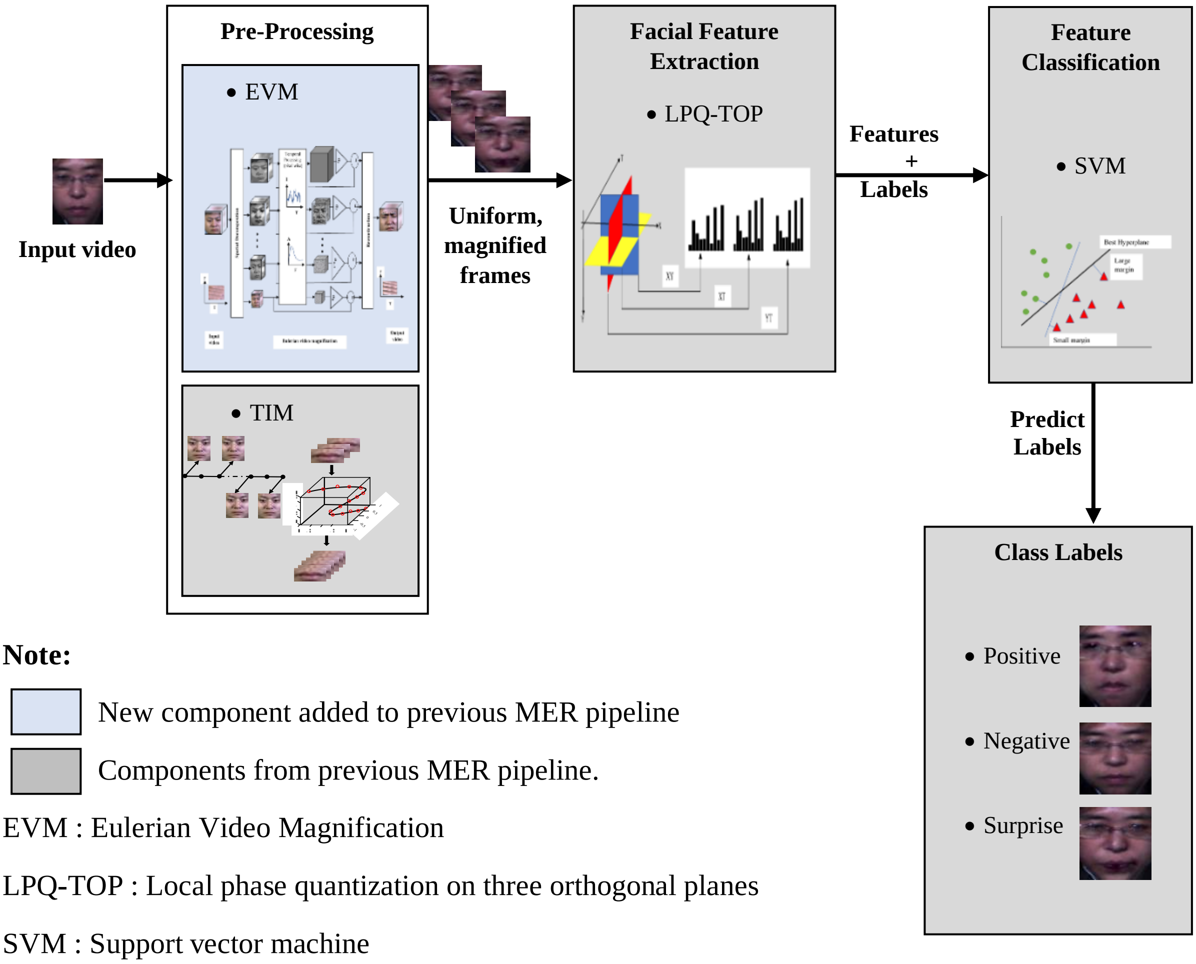
Variations in temporal spectral components of $\delta(t)$ at a given index k is denoted by $\delta_k(t)$ in equation (5.7) attenuated by γk (known as temporal filtering factor). Due to the equivalence between the frequency dependent amplification factor and the temporal frequency reliant attenuation i.e., $\alpha k = \gamma k \alpha$, the outcome is a motion magnified video expressed as [79]:

$$\tilde{I}(x,t) \approx f(x + \sum_k (1 + \alpha_k)\delta_k(t)) \tag{5.8}$$

The amplification achieved with a factor $(1 + \alpha)$ using Taylor series expansion is theoretically true considering the assumptions described earlier, however in reality it suffers from certain limitations when working with fine grained frames depicting feeble motion. To decide on the best amplification factor, α that can produce motion magnification with highest precision for a given wavelength, bounds are employed to regulate the entire process, using equation (5.9).

$$(1 + \alpha)\delta(t) < \frac{\lambda}{8} \tag{5.9}$$

Here, λ represents the spatial wavelength given by 2π / ω for a frequency ω, α is the magnification factor, δ(t) represents the observed motion.



Note:

- ▭ New component added to previous MER pipeline
- ▨ Components from previous MER pipeline.

EVM : Eulerian Video Magnification

LPQ-TOP : Local phase quantization on three orthogonal planes

SVM : Support vector machine

TIM : Temporal interpolation model

Figure 5.2. Microexpression recognition pipeline with EVM,TIM, and LPQ-TOP.

5.3 Proposed Approach

The demand to design a competent automatic MER system in CV is increasing noticeably owing to its cross-discipline applications. However, due to its faint and rapid characteristics recognizing such expressions is still extremely challenging. To deal with this issue, this section introduces video magnification into the existing MER framework that was tested earlier using the CASMEII dataset. The basic framework employed in this thesis for the

MER problem has already been discussed in Chapter 4, along with an appropriate illustration (see Figure 4.6). Utilizing the components from the basic framework, in this section it is extended further to address three class MER problem.

This proposed framework along with the new component is illustrated in Figure 5.2. It consists of EVM in the pre-processing stage, which is the new addition, to be utilized with previously employed components. In the previous work (Chapter 4), the data was processed with TIM, then the LPQ approach was chosen and applied along three orthogonal planes, denoted as LPQ-TOP, for extracting facial micro features. Both these methods are revisited in this section to demonstrate its proficiency when combined with EVM, which had not been addressed in the experiments discussed in Chapter 4. This section continues to exploit the MER pipeline from the previous experiment, introducing new additions to it and presenting extensive research conducted along with a thorough analysis to examine effects of magnification on LPQ-TOP method. Ultimately, the extracted features are used as input to a SVM for training, testing and classification. Thus, usage of TIM [77], LPQ-TOP [127-129] and the SVM [105] method for feature extraction and classification remain unchanged in this experiment too. The class labels considered for experiments in this chapter remain as negative, positive and surprise which is same as those used in Chapter 4. The effectiveness of the proposed approach is determined based on several tests and performance evaluation obtained over a variety of datasets namely SAMM, SMIC (HS, NIR & VIS), CASME, CASMEII and CAS(ME)^2. To ensure the work in this thesis remains relevant with fewer explicit facial expressions and reduced expressiveness in faces with autism, relabeling of samples seemed appropriate and consequently three labels i.e., positive, negative and surprise were chosen. Moreover, to have more uniformity in class labels across all datasets relabeling was endorsed as some datasets consist of more than three labels. Throughout the experiments conducted in this section, happy expressions are relabeled as positive, whereas all expressions originally labelled as surprise remain unchanged. Expressions originally labeled as disgust, sadness and fear are relabeled as negative for all datasets excluding SMIC, since it is already labeled into these particular three classes. Additionally, the original SAMM and CAS(ME)^2 dataset contain an anger label which is also categorized under negative in this implementation. The CASME dataset consists of tense label, which is classified as negative in this work. Some data originally labelled as repressions, helpless, pain, contempt, and others have not been considered in this work since these labels are present in selected datasets only. Moreover, these expressions do not seem to be a good fit considering expression limitations in autism. The procedure for extracting features using LPQ-TOP outlined in Section 4.2 has been followed here.

To comprehend the effect of magnification and build a working model of the proposed concept for solving the three class MER problem, the entire experiment is conducted in two phases. In the first phase the features are extracted with LPQ-TOP from data belonging to all seven databases individually, without introducing magnification to obtain seven corresponding sets of feature vectors. Classification is performed individually on each of these vectors to categorize the input into appropriate labels. In the second phase of the experiment, magnification of the ME videos for all seven datasets is implemented individually using the EVM technique described in Section 5.2. Following this, the corresponding magnified data are processed with TIM. Features are then extracted using LPQ-TOP from these magnified frames to obtain another corresponding set of seven feature vectors which are then forwarded for classification. To provide a fair analysis of LPQ-TOP performance, this section compares it’s results with other popular feature extraction methods generally employed for MER. Details of experiments conducted using the proposed approach along with its corresponding results obtained are provided in Section 5.4. It must be mentioned here that the literature review in Chapter3 revealed that comparatively lesser work employing SAMM, CAS(ME)^2, and CASME database could be found. Moreover, there was an absence of similar work employing magnification on these databases due to which appropriate comparison could not be established. The work used for comparison employing CASMEII and SMIC database is somewhat relevant to the work described in this thesis therefore used for comparison, however magnification factor and extraction methods are different to that used in this thesis.

The work in this Chapter has two contributions in the field of MER. First, it presents a comprehensive examination and analysis to ascertain the usefulness of employing TIM and EVM with LPQ-TOP to boost recognition performance. Second, it offers a thorough investigation of LPQ-TOP using several datasets contributing to its cognizance as a potential micro facial feature extraction technique substantiated by results that are comparable with some of the existing methods which are popularly employed in this domain.

5.4 Experiments and Results

The details of the experiments conducted using the proposed approach along with the specifications of parameters used throughout this set of experiments are presented in this section. The EVM approach introduced by [79] follows four steps to produce a magnified video. First is performing spatial decomposition, second is applying a bandpass filter. Third, it multiplies the extracted bandpass signal with the amplification factor. Fourth is adding this amplified signal back to the original video. To decompose video into various spatial frequency band, a Laplacian pyramid is constructed. Then for extracting the desired frequency band, it can use either an ideal filter, Butterworth or second order IIR. Empirically [79] found that the ideal bandpass filter was more suitable to amplify color whereas the second order bandpass IIR filter seemed suitable for magnifying both color and motion. To examine the effect of all three filters on ME, in this thesis, a pilot test was performed on a video taken from the CASMEII dataset. Results obtained thereafter were compared with instances of image sequences extracted from the original ME video without performing magnification. Figure 5.3. illustrates the original image sequence for the disgust expression to be used for a fair visual comparison with various results obtained in this set of experiments. Another need for conducting the pilot test was to identify a suitable magnification factor for amplifying ME videos.

Thus, the overall work was divided into three experiments. First was the pilot experiment, second was extracting features from non-magnified image sequences for seven individual datasets and performing classification. In the third set of experiments the videos from seven datasets were magnified, then feature extraction was performed followed by the classification process.

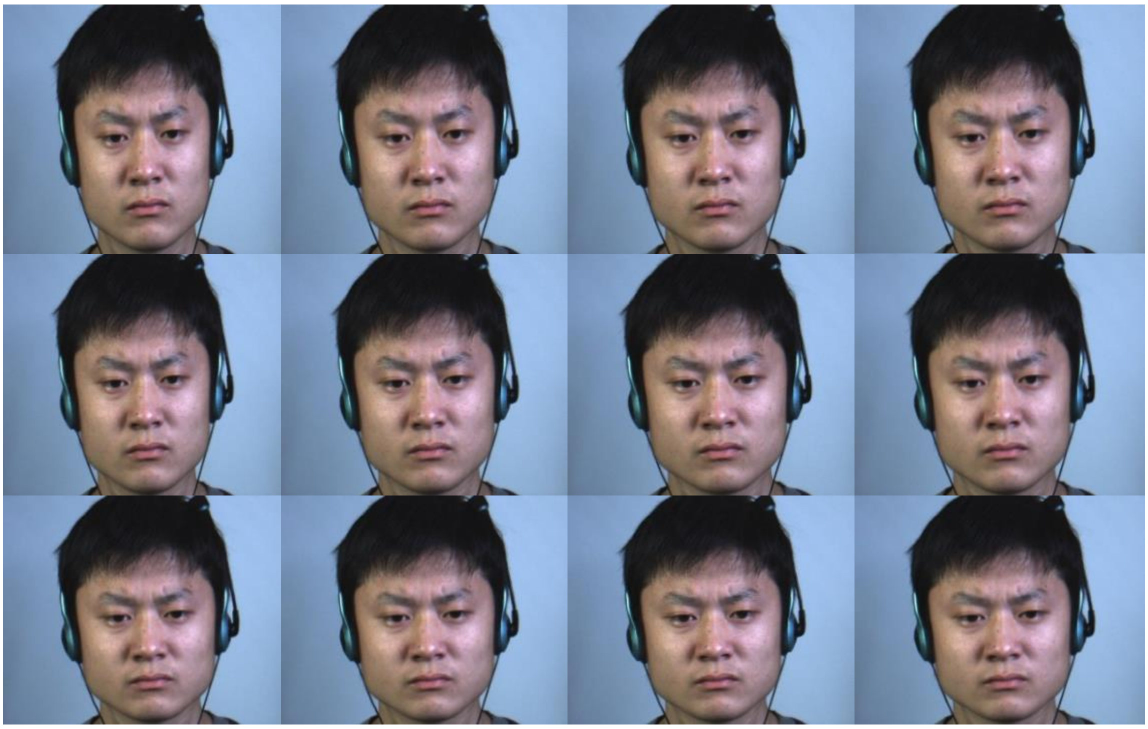
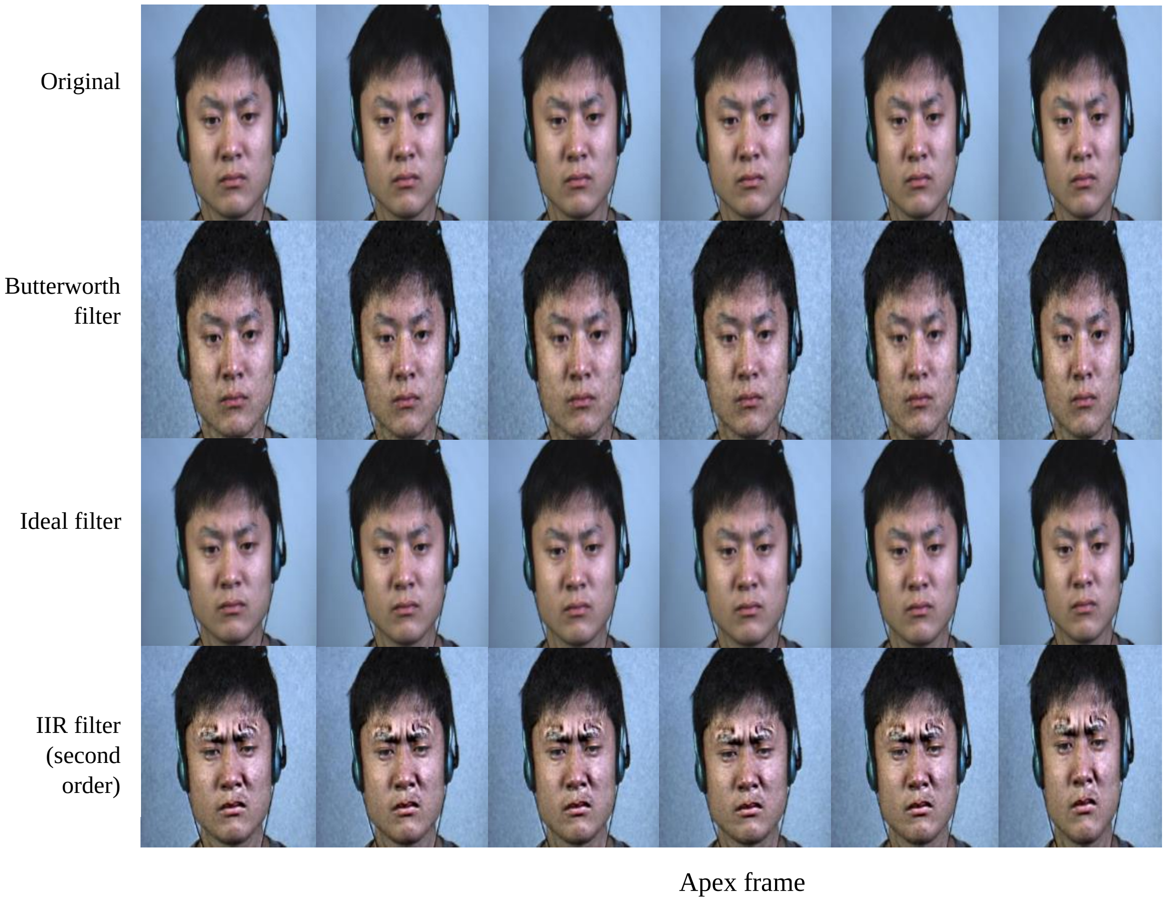


Figure 5.3. Non-magnified raw image sequence for disgust micro expression.



Note: Every fourth frame in each row represents the apex frame.

Figure 5.4. Illustration of magnified image sequences for disgust micro expression obtained using different bandpass filters.

5.4.1 Pilot experiment

Through the pilot experiment the target was to make an appropriate selection of two entities i.e., bandpass filter and amplification factor, to be applied in the rest of the experiments. In order to verify the effect of different filters used in the magnification process, three sets of experiments were conducted with the magnification factor set to 26, chosen randomly. In the first experiment a Butterworth filter was applied to obtain the amplified ME. Image sequences extracted from the amplified video obtained is presented as layer 2 in Figure 5.4. Similarly, the ideal filter was applied to obtain second magnified video represented by layer 3 in Figure 5.4. Finally, by applying the IIR filter in the third experiment it obtained magnified frames as represented by layer 4 in Figure 5.4. Comparing the original frames with the results obtained one can visually find no difference between magnified and non-magnified frames using the ideal filter. The effect of magnification can hardly be noticed in the magnified frames (see layer 2 in Figure 5.4). Also, the presence of noise due to magnification is not very prominent in this case. The magnified frames obtained using Butterworth filter suffer from noise compared with the ideal filter results. Visibly this can be realized by observing the frames presented in layer 2 in Figure 5.4. Here, the presence of magnified noise is overshadowing the magnified expressions obtained. Observing the magnified frames obtained using the IIR filter one can easily perceive the exaggeration occurring above the eyebrow regions. Clearly, the muscle movement is much more visible when compared with its non-magnified frames. Keeping the magnification factor uniform throughout this pilot test, the results presented here have been obtained. By carefully examining these results and following the work by [49] and [79], selecting the IIR filter to perform magnification for experiments in this thesis seemed reasonable.

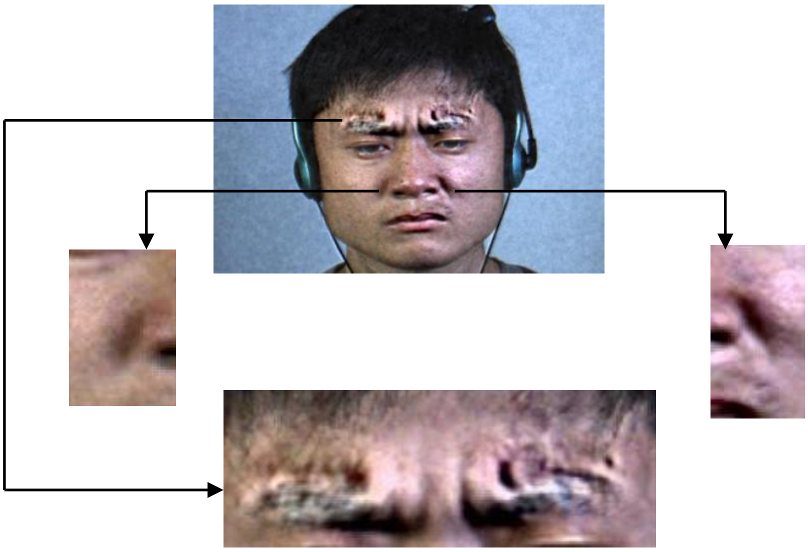
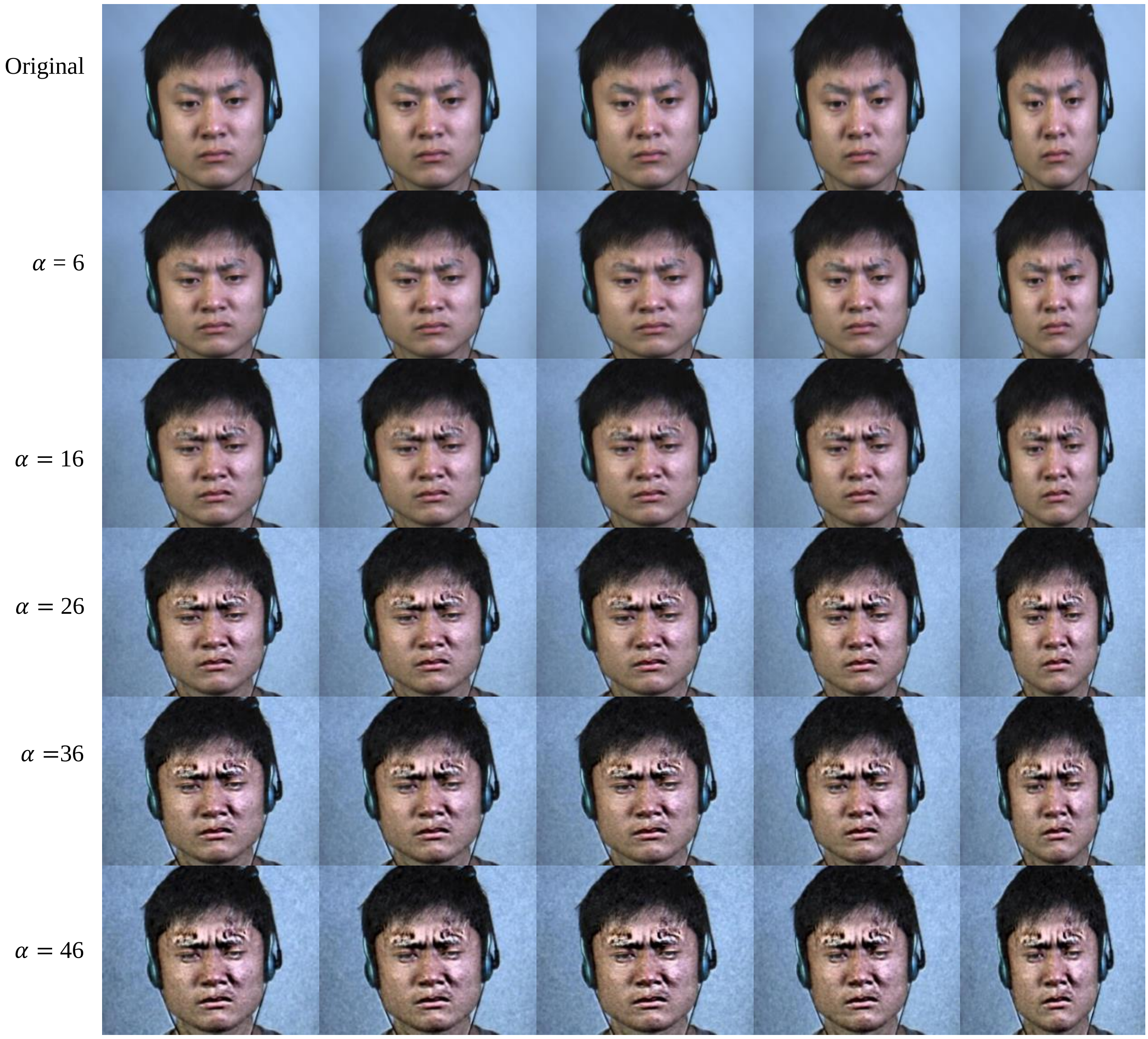


Figure 5.5. Highlighting areas with appearance of muscle motion exaggeration after applying EVM.

An instance of ME magnified using EVM, highlighting areas with appearance of exaggerated muscle is presented in Figure 5.5. Evidently from Figure 5.5 one can visualize that magnification has a profound effect on accentuating minute expressions.



Note: Every third frame in each layer represents the apex frame.

Figure 5.6. Demonstrating magnified image sequences obtained for disgust micro expression at different settings of magnification factor α.

In order to have a guideline for selecting the amplification factor to be applied in the rest of the experiment, another pilot test was conducted by regulating magnification factor, α in equation (5.9). Magnification method, applying the IIR filter was tested on the disgust video ME for five different α values, set at 6, 16, 26, 36 and 46. These magnified image sequences extracted from individual results, obtained using five different α values are presented in Figure 5.6. The exaggeration of muscle movement obtained by setting α= 6 is not very discriminative indicating that the effect of magnification is less recognizable. Extremely feeble magnification effect can be perceived here when observed minutely but this degree of amplification is not enough for conducting experiments in this thesis. Continuing with the pilot testing phase and regulating amplification factor the experiments move on to higher α values. As the α value increases to 16 the effect of magnification can be seen on the facial muscles, especially on areas around the nose and above the eyebrows (see row 3 in Figure 5.6). This effect is visibly more perceivable when compared with the results obtained with α set to 6. A higher exaggeration is more prominent as the magnification factor reaches 26. At this setting motion magnification along with slight color variations can be observed. Beyond this amplification factor (i.e., α= 36 & =46) the visibility of color variations due to amplification is very high and interferes with the appearance of motion magnification to a large extent. Comparatively, the images obtained in these two cases seem noisier than previous cases. Since the appearance of muscle motion is very prominent when α is set at 26 and the interference due to color changes seem negotiable; this setting seemed suitable for experiments to be performed in this thesis. Moreover, comparing the results between α=16 and α=26, colour changes in both cases is almost similar. Also, exaggeration of motion is visibly quite superior for α=26 by comparison. Besides, [49] and [137] were successful in achieving good recognition using α=20 and α=26 respectively, thus for this thesis experiments with a magnification factor set above 20 was chosen. Since the obstruction due to color changes for α= 36 and α=46 is very high and dealing with these can be very time consuming, these settings were not considered in this work. Also, due to very low visibility of magnification on facial muscles for α=6, this setting too did not seem appropriate for conducting experiments. Thus, throughout the experiments hereafter the magnification factor was set at 26 and an IIR filter was employed for realizing motion magnification using EVM technique outlined in Section 5.2.

5.4.2 Experiments employing non-magnified data

Following the work in [40] and [89], the TIM parameters were chosen in this set of experiments for processing the raw data without introducing a magnification technique. As highlighted in Section 5.3, since this section focuses on categorizing facial ME into three classes, the chosen labels are positive, negative and surprise. Also, as previously discussed relabelling has been obtained as described in Section 5.3. A total count of 133 ME images have been considered from the CASME dataset. Here the distribution of data is biased towards a negative label. For CASMEII a total of 122 instances were used with distribution of 32 positive, 65 negative and 25 surprise labels. This data distribution is better than those in the CASME dataset. Very few ME data was available in CAS(ME)^2 with a total count of 51 and with a distribution of below twenty data for positive and surprise labels.

With data distribution skewed more towards negative labels, 121 ME data were used from the SAMM dataset. Comparatively, this dataset contained less noisy data than others. Extremely few data labelled as surprise in this dataset could be employed in this work, due to low availability. Among all the datasets considered, SMIC-VIS and SMIC-NIR contained the most evenly distributed data. Both the datasets had a distribution of 28 positive, 23 negative and 20 surprise labels. The highest count of data employed in our work was from SMIC-HS, with a total count of 164. Distribution of data in this dataset was satisfactory with 51 positive, 71 negative and 42 surprise labels.

Thus, one can see that majority of the available ME datasets have unbalanced class division. The total datasets used along with the specifications of their distribution into three class labels are presented in Table 5.1. This data distribution remains common for both the experiments described in Section 5.4.2 and Section 5.4.3.

Table 5.1. Dataset used and their class distribution.

Dataset	**Total Samples Used**	**Positive Label**	**Negative Label**	**Surprise Label**
CASME	133	9	104	20
CASMEII	122	32	65	25
CAS(ME)^2	51	12	29	10
SAMM	121	26	80	15
SMIC-VIS	71	28	23	20
SMIC-NIR	71	28	23	20
SMIC-HS	164	51	71	42

Table 5.2. LPQ-TOP performance on seven datasets (without magnification).

Dataset	**Accuracy %**	
	Decorrelation(0.1)	**Decorrelation(0)**
CASME	83.42	**83.45**
CASMEII	61.09	**61.16**
CAS(ME)^2	**63.6**	63.2
SAMM	70.0	**70.4**
SMIC-VIS	**65.6**	64.91
SMIC-NIR	**63.3**	62.9
SMIC-HS	61.11	**62.8**

The raw data have been up sampled with TIM before performing feature extraction. While employing the LPQ-TOP method the neighbourhood size is set to 5. To perform cross validation for training, along with multi-class classification using one-vs-all strategy, experiments were performed using 10-fold. Experiments for classification were carried out using SVM with three different kernels namely polynomial, RBF and linear. Performance of LPQ-TOP is assessed on all seven datasets using classification accuracy. The accuracies are presented in Table 5.2. The performance of the chosen extraction method i.e., LPQ-TOP is impressive with the highest accuracy of 83.45% recorded on the CASME dataset.

The lowest performance using the LPQ-TOP approach was obtained with the CASMEII dataset with accuracy of 61.16%. Similar recognition accuracy is observed on SMIC and CAS(ME)^2 datasets, with 65.6%, 63.3% and 62.8% accuracy obtained on VIS, NIR and HS variations of SMIC datasets respectively. For CAS(ME)^2 the value obtained was closer to SMIC-NIR at 63.6% accuracy. Among all these datasets, SAMM contains the cleanest data along with exceptional facial resolution. Here the approach reached an accuracy of 70.4% which is reasonably good performance, considering no magnification was applied yet.

5.4.3 Experiments employing magnification process

Experimentally this section investigates if intensifying muscle movements on the face during the initial stages by explicitly utilizing EVM followed by TIM can assist LPQ-TOP in producing recognition performance improvement on facial ME using various datasets. Further it also examines if magnification with EVM can help realise uniform increase in recognition rate across all the chosen datasets. By carefully observing the recognition rate obtained for all cases, an analysis is performed to give useful inferences. Keeping the parameters consistent with Section 5.4.2 during extraction process, this second phase of experiment was conducted.

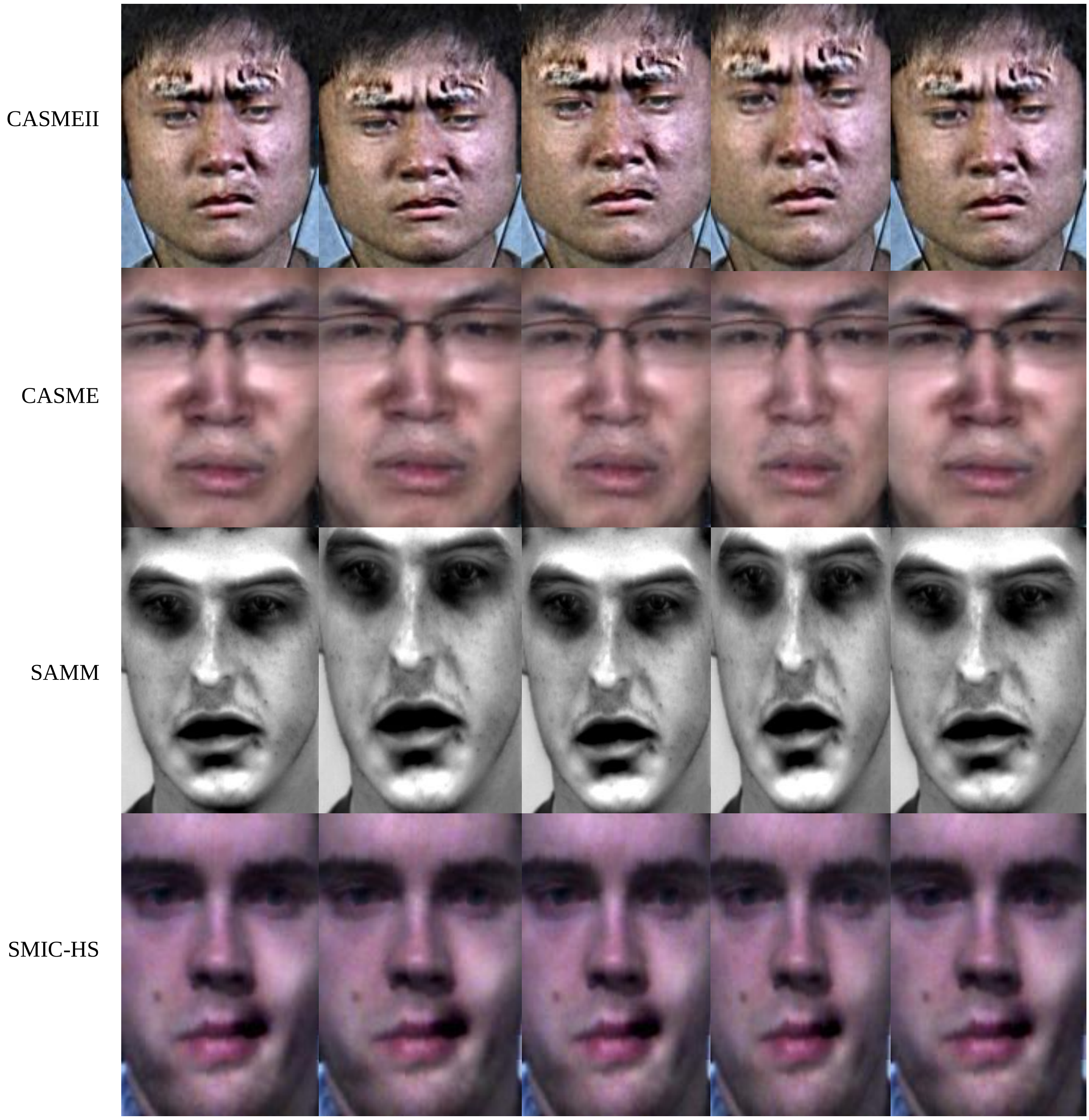


Figure 5.7. Magnified micro expression image sequences obtained for various datasets with the amplification factor α= 26 using IIR filter.

As discussed in Section 5.3, the proposed framework employed magnification in combination with TIM and the LPQ-TOP feature extraction method for ME analysis. The sample distribution used here is the same as presented in Table 5.1. These samples were first magnified by applying EVM, implemented using IIR filter and the magnification factor set to 26. The effect of applying EVM with the mentioned parameters and filter on various dataset can be visualized from image sequences presented in Figure 5.7. Due to the magnification process the almost invisible muscle movement on every dataset is amplified and has become very prominent. For instance, in layer 2 representing the CASME dataset, the amplified muscle near the mouth as well as the nose region can be perceived clearly. Similar observations around the mouth region can be seen for layer 4 representing the SMIC-HS dataset. This observation is consistent for the SAMM image sequence too, representing substantial muscle movement. The features were then extracted from these magnified data from all seven datasets in different experiments. Following the similar process outlined in Section 5.4.2, SVM was applied to train and test with these features. Performance obtained across all datasets by applying the proposed approach is presented in Table 5.3.

Table 5.3 Accuracy % obtained using LPQ-TOP and its comparison across various datasets.

Dataset	**Accuracy % (LPQ-TOP)**		
	No Magnification	***With Magnification***	***% Increase***
CASME	83.45	88.20	4.75
CASMEII	61.16	74.50	13.34
CAS(ME)^2	63.60	68.50	4.90
SAMM	70.40	72.07	1.67
SMIC-VIS	65.60	73.80	8.20
SMIC-NIR	63.30	70.42	7.12
SMIC-HS	62.80	65.80	3.00

Analogous to classification results for non-magnified datasets, in this case also, the highest accuracy of 88.2% is obtained for the CASME dataset by employing magnification. After magnification, the lowest performance using the proposed approach was obtained for SMIC-HS dataset resulting in an accuracy of 65.8%. An impressive performance boost using the

CASMEII dataset after introducing magnification has been noticed with accuracy recorded at 74.5%. On CAS(ME)^2 dataset a reasonable effect of magnification can be witnessed with accuracy of 68.5%. With a very competitive accuracy at 73.8% and 70.42%, performance boost for SMIC-VIS and SMIC-NIR is also notable. Almost an insignificant effect of magnification is observed on the SAMM dataset with an accuracy of 72.07%. The results obtained demonstrate that EVM can enhance the extraction capability of LPQ-TOP to achieve improved recognition performance for the majority of datasets. Observing the recognition boost achieved, significant influence of magnification in improving the performance of LPQ-TOP is undeniable and this influence is observable among all chosen datasets.

Further examining these recognition values revealed that, even though a rise in recognition performance after introducing magnification is obvious for all datasets, no uniform pattern could be observed in these increased recognition rates. For instance, a significant boost in recognition accuracy of 13.34% was recorded on the CASMEII dataset after employing magnification in contrast to a small boost of 1.67% achieved on the SAMM dataset even though the number of samples used is approximately same for both these datasets. Moreover, the data are also recorded with same fps for both, yet a significant variation in the influence of magnification is observed on these two datasets. Diverging from this pattern, between SMIC-VIS and SMIC-NIR the rate of increase varies by only 1% (approx.) , indicating a more uniform rate of increase in the recognition accuracy among these datasets. It is important to note here that these two SMIC datasets contain data recorded with the same rate of fps and the volume of data is the same for both. These characteristics observed among SMIC datasets are similar to characteristics shared between the SAMM and CASME dataset, yet a uniform magnification influence could not be observed among the latter two datasets. The SAMM and CASMEII databases have non-uniform distribution of data whereas both SMIC datasets contain more balanced data which may have resulted in this variation, with results appearing more favourable using the SMIC datasets. Certainly, the classification technique also seems to have benefitted due to uniform data distribution evident from the rate of increase in the recognition accuracy obtained which are more uniform for SMIC database.

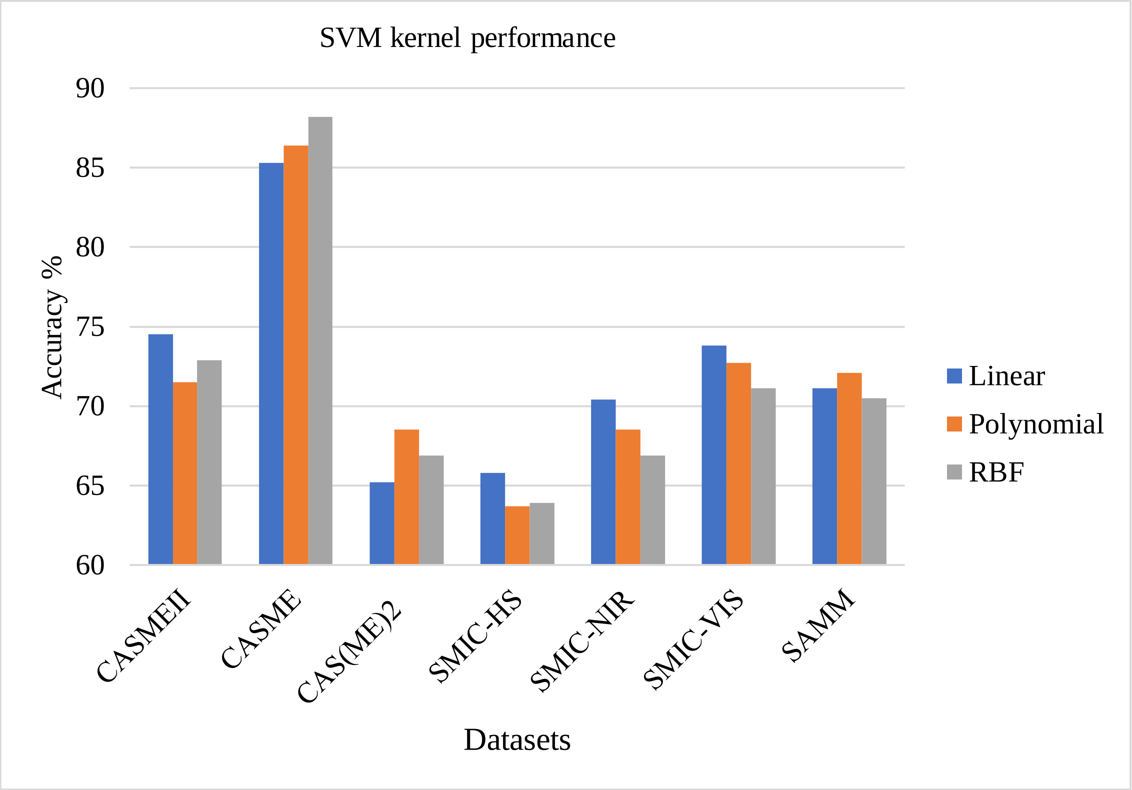


Figure 5.8. SVM kernel performance comparison on various dataset.

Overall, considering all the datasets employed in this experiment, the average increase in recognition accuracy was computed as 6.14% (approx.) which indicates successful implementation of the proposed pipeline for MER problems addressing three class labels. The performance comparison of three SVM kernels on seven datasets used in this experiment is presented in Figure 5.8. Among the three kernels tested, experimentally it was observed that the linear kernel produced best results for the SMIC dataset. It produced the best accuracy and outperformed the other two kernels by quite a large margin on all three variations of the SMIC dataset used. This kernel also performed well on CASMEII dataset as demonstrated by the results. However, on the CASME dataset the RBF achieved highest accuracy and outperformed the linear kernel. Also, performance of the polynomial kernel was slightly better than linear kernel when tested on the CASME dataset. Comparatively by employing the polynomial kernel the SVM recorded improved accuracy on both SAMM and CAS(ME)^2 datasets. Particularly in the case of CAS(ME)^2 , the accuracy obtained using this kernel can be seen to be higher than linear and RBF. Performance of all the three kernels was almost similar using SAMM dataset but considering the overall performance, the polynomial kernel resulted in better performance with a very small margin.

Table 5.4. Accuracy % comparison between LPQ-TOP and other methods.

Dataset	**Our Work (Accuracy %)**	**Other Authors**	
	EVM +TIM+LPQ-TOP	**Accuracy %**	**Method [40]**
CASME	88.2		-
CASMEII	74.5	78.14	HIGO + mag
		63.97	HOG + mag
		60.73	LBP-TOP + mag
CAS(ME)^2	68.5		-
SAMM	72.07		-
SMIC-VIS	73.8	81.69	HIGO + mag
		77.46	HOG + mag
		78.87	LBP-TOP + mag
SMIC-NIR	70.42	67.61	HIGO + mag
		64.79	HOG + mag
		67.61	LBP-TOP + mag
SMIC-HS	65.8	68.29	HIGO + mag
		61.59	HOG + mag
		60.37	LBP-TOP + mag

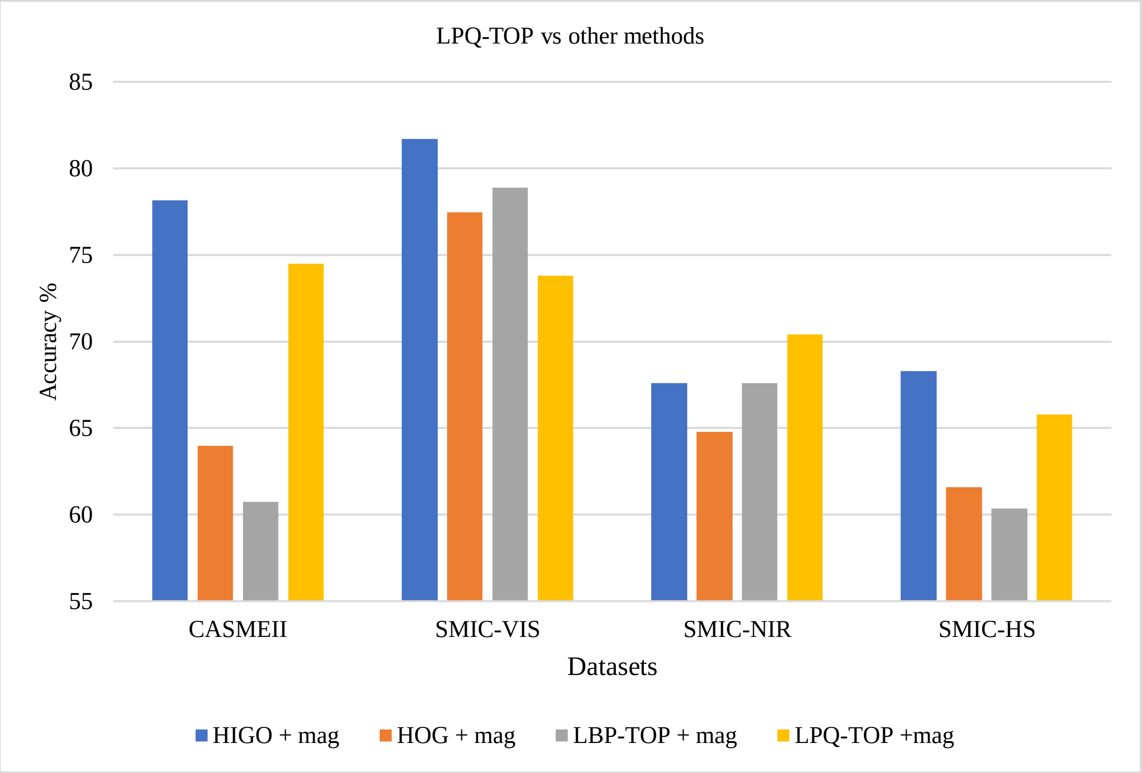


Figure 5.9. Performance comparison of LPQ-TOP and other methods.

5.4.4 Proposed approach vs. other methods

To determine the overall performance of the proposed approach, comparison with other similar approaches is conducted. The accuracies obtained after introducing magnification on all seven datasets in comparison to other methods are presented in Table 5.5. For a fair comparison of LPQ-TOP performance with magnification compared with other methods, an analysis is presented in Figure 5.9. Carefully observing the results, one can notice that the proposed approach is able to achieve a performance boost which is comparable with some of the existing approaches on the majority of datasets. For instance, using the CASMEII and SMIC-HS datasets, the proposed approach is able to produce accuracy which is 10.53% and 4.21% higher than HOG with magnification (HOG+mag). Likewise, using both these datasets the proposed approach once again produced 13.77% and 5.43% higher accuracy than LBP-TOP with magnification (LBP-TOP + mag). From these observations it was understood that the chosen approach resulted in performance which was much better than HOG + mag and LBP-TOP+ mag. However, for the same two dataset the performance of the proposed approach seemed slightly less competent than the third method i.e. (HIGO + mag). The performance obtained using the proposed approach on SMIC-NIR dataset recorded 5.63%, 2.81% and 2.81% higher accuracy than all the three existing methods i.e., HOG, HIGO, and LBP-TOP respectively, when implemented along with magnification.

Using the SMIC-VIS dataset, the proposed approach obtained an accuracy of 73.8% yet is still lower than other methods in comparison. Consequently, the proposed approach does not seem to meet the expected results for this dataset. Due to lack of similar work on CASME, CAS(ME)^2 and SAMM datasets appropriate comparisons could not be made. However, experimentally the proposed approach was able to gain a significantly improved recognition performance using the CASME dataset, recorded as 88.2%. This is by far the highest accuracy obtained using the proposed approach across all datasets. For the SAMM and CAS(ME)^2 datasets using the same approach the accuracy recorded was 72.07% and 68.5%. Therefore, the results obtained utilizing the proposed approach look promising and are closely comparable specially with LBP-TOP and HOG-TOP.

In the experiments conducted some of the samples have been excluded, this may have given an added advantage to our method thus resulting in high recognition accuracy. From these results one can conclude that the performance of LPQ-TOP with magnification is comparable with both binary and gradient based methods. Examining these results more closely it is obvious that magnification has a substantial influence on the CASMEII dataset followed by SMIC-VIS and SMIC-NIR. On the other hand, the performance of the LPQ-TOP method was consistently better on the CASME dataset. From this thorough examination it can be inferred that overall performance of the proposed approach was competent enough to accomplish a good MER performance, therefore can be compared and contrasted with similar existing techniques. Experimental results show that the proposed approach can be viewed as an alternative approach for solving MER problems. Findings from this work have shown that the feature extraction method LPQ-TOP has undeniably benefited from magnification which has led to boost in recognition accuracy for all datasets.

Experimentally, a higher magnification factor seems to work positively when combined with TIM and LPQ-TOP method. Moreover, fewer class labels used in this experiment to maintain higher relevance with ME exhibited by ASD individuals may have worked in favour of the proposed approach.

5.5 Conclusion

This work provides an extensive performance analysis of an approach that combines magnification, interpolation, and phase quantization. The proposed methodology is tested on seven different spontaneous ME datasets to ascertain the performance boost achieved by the LPQ-TOP approach when employed together with magnification and interpolation. The results highlight the usefulness of the approach for dealing with MER problems. However, it should also be noted that even though performance boost has been achieved for every instance, yet it lacks consistency in the rate of increase in the recognition accuracy across the datasets. An average boost of 6.15% has been achieved using the proposed approach which is promising. This clearly substantiates the advantage of employing magnification to significantly aid extraction techniques in efficiently distilling relevant facial micro features thereby boosting overall recognition performance to a large extent. A closer investigation revealed a clear performance bias towards the CASMEII dataset after magnification in comparison to other datasets sharing similar data count and frames per second.

Furthermore, this work also presents an extensive comparison between some popular existing feature extraction methods and quantization methods, where tests were conducted both with and without EVM on several spontaneous ME datasets. Evidently the results obtained successfully establish the competency of the LPQ-TOP technique particularly as a facial micro feature extraction technique with abundant scope for further exploration. Indeed, through this work a novel pipeline aimed to solve the three class MER problem, particularly by employing video, is realized successfully. Moreover, the results obtained here can be utilized for making comparisons while designing new methodologies in the future. Additionally, experiments performed provide the groundwork for MER frameworks that can be extended for autism screening, detection, and diagnosis as future application. The results obtained from the proposed pipeline at this stage looks promising and the methodology can be explored further.

Imbalanced classification is one of the limitations of the proposed approach due to an unequal distribution of available data samples with the exception of the SMIC datasets. In order to perform competent ME analysis, having an adequate number of training samples is crucial and this is often a challenge due to unavailability of adequate volume of spontaneous ME data samples and datasets. Therefore, unavailability of sufficient ME data is the second limitation of the current work. Nevertheless, the results obtained by exploiting LPQ-TOP method are positive and shall be explored further in Chapter 6.

Chapter 6

Image Super Resolution for Micro Expression Analysis

6.1 Introduction

ISR based ME analysis is proposed to address the image quality issues often generated while capturing recordings as highlighted in the earlier chapters. To regain the discriminative features lost due to quality issues, DL based SR methods are investigated in this section for transforming low resolution ME images into SR images. Additionally, this chapter presents an exhaustive performance analysis of various SR algorithms employed for ME image reconstruction along with a comparative performance analysis of the overall pipeline on three simulated ME databases. Taking forward the SR concept for ME introduced by [41], this thesis proposes a pipeline that utilizes GAN and its variant along with other DL approaches to achieve ISR on low quality ME images. To best of the author's knowledge, at present both DL and GAN approaches have not been utilized specifically on low resolution ME images, this work is a first attempt to realize it. The proposed pipeline aims to combine the best features from both handcrafted methods as well as DL techniques. Low resolution ME images obtained by simulating data from SMIC (HS&VIS) [64] and CASMEII [47], databases are used to evaluate the performance of proposed approach.

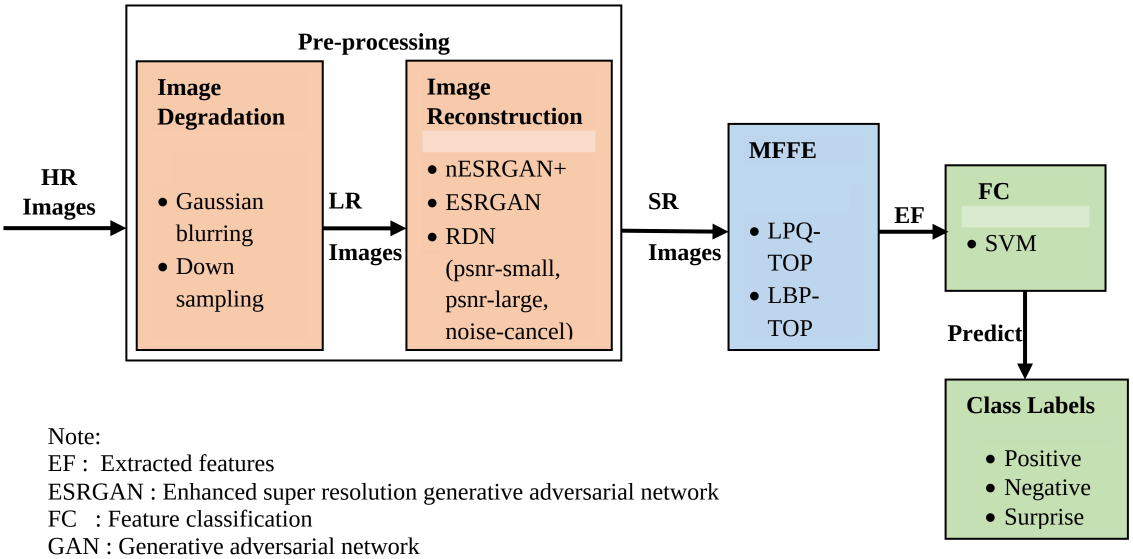
The remainder of this chapter is organized as follows. In Section 6.2 the overall pipeline proposed for reconstructing ME from LR images and its recognition process is described. Experimental results obtained along with its comprehensive analysis is presented in Section 6.3. This is followed by a summary of the work presented in Section 6.4.

A portion of experiments and results presented in this chapter has been published in

- Sharma, P., Coleman, S., Yogarajah, P., Taggart, L. and Samarasinghe, P. (2022), Evaluation of Generative Adversarial Network Generated Super Resolution Images for Micro Expression Recognition. In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - ICPRAM, ISBN 978-989-758-549-4, pages 560-569. DOI: 10.5220/0010820100003122.
- Sharma, P., Coleman, S., Yogarajah, P. et al. Comparative analysis of super-resolution reconstructed images for micro-expression recognition. Adv. in Comp. Int. 2, 24 (2022). https://doi.org/10.1007/s43674-022-00035-x.

6.2 The Proposed Micro Expression Reconstruction & Recognition Pipeline

To deal with the issue of LR in images consisting of ME, this thesis proposes a framework with a DL and GAN based SR image reconstruction module along with three other standard modules namely, image degradation, micro facial feature extraction and feature classification as shown in Figure 6.1. The initial two modules, image degradation and image reconstruction, are also known as pre-processing modules. They are employed to prepare the datasets for the recognition task. This is followed by a micro facial feature extraction module where the essential micro features are extracted from the input facial images. As a final step, the classification module manages the task of assigning appropriate class labels based on the extracted features.



Note:
EF : Extracted features
ESRGAN : Enhanced super resolution generative adversarial network
FC : Feature classification
GAN : Generative adversarial network
HR: High resolution
LBP-TOP : Local binary pattern on three orthogonal planes
LPQ-TOP : Local phase quantization on three orthogonal planes
LR : Low resolution
MFFE : Micro facial feature extraction
RDN : Residual dense network
SR : Super resolution
SVM : Support vector machine

Figure 6.1. Outline of proposed approach employing image reconstruction before micro expression recognition.

For experimenting, the HR images from the ME database are initially input into the image degradation model to generate the corresponding LR images. The degradation model applied to achieve this has already been discussed in Section 3.6. For every existing HR database, the proposed pipeline obtains two sets of LR databases, i.e., 64x64 and 32x32, by applying a down scale factor of two and four.

Following this, for every set of LR64 databases four different SR methods psnr-small [119,139,], psnr-large [119,139], noise-cancel [120,139] and ESRGAN [121, 139] are applied in different set of experiments. Here, the two methods, psnr-small and psnr-large perform SR using the RDN architecture discussed in Section 3.3 whereas ESRGAN and the noise-cancel technique utilize a GAN architecture to generate SR images. The generator for noise-cancel utilizes the RDN structure from Section 3.3 (refer to Figure 3.6) and the discriminator follows the architecture discussed in Section 3.4.1 (refer to Figure 3.11). The ESRGAN generates SR images by utilizing the structure described in Section 3.4.2. Likewise, for every set of LR32 databases ESRGAN and nESRGAN+ super resolution methods are applied. The architecture employed by nESRGAN+ model to generate SR images is as discussed in Section 3.4.3.

In addition to these techniques, SR experiments using bicubic interpolation [125] are also performed. Therefore, experimentally the pipeline generates SR images using the bicubic interpolation technique discussed in Section 3.5. By comparing the results obtained from DL and GAN based SR methods with the bicubic method, this thesis validates the reconstruction performance on ME images and makes useful inferences.

Every SR database obtained at both 64x64 and 32x32 resolution is used as input to the feature extraction method at different instances. Throughout this chapter, for every set of HR, LR and SR databases both LBP-TOP and LPQ-TOP feature extraction techniques have been employed individually. The corresponding feature vectors obtained thereafter are then given as input into the classification module to perform appropriate expression classification. Throughout the experiments, classification has been achieved using SVM as discussed in Section 2.4.4. Parameter specification of both the feature extraction method and SVM classification are presented in Section 6.2.1. Specifications for the three databases used in this Chapter are presented in Table 6.1 followed by Table 6.2, which presents a summary of resolutions, notations and different SR methods employed throughout the experiments. Actual parameters and other experimental settings used for all the methods shall be discussed in Section 6.3.

Table 6.1 Spontaneous micro expression dataset used.

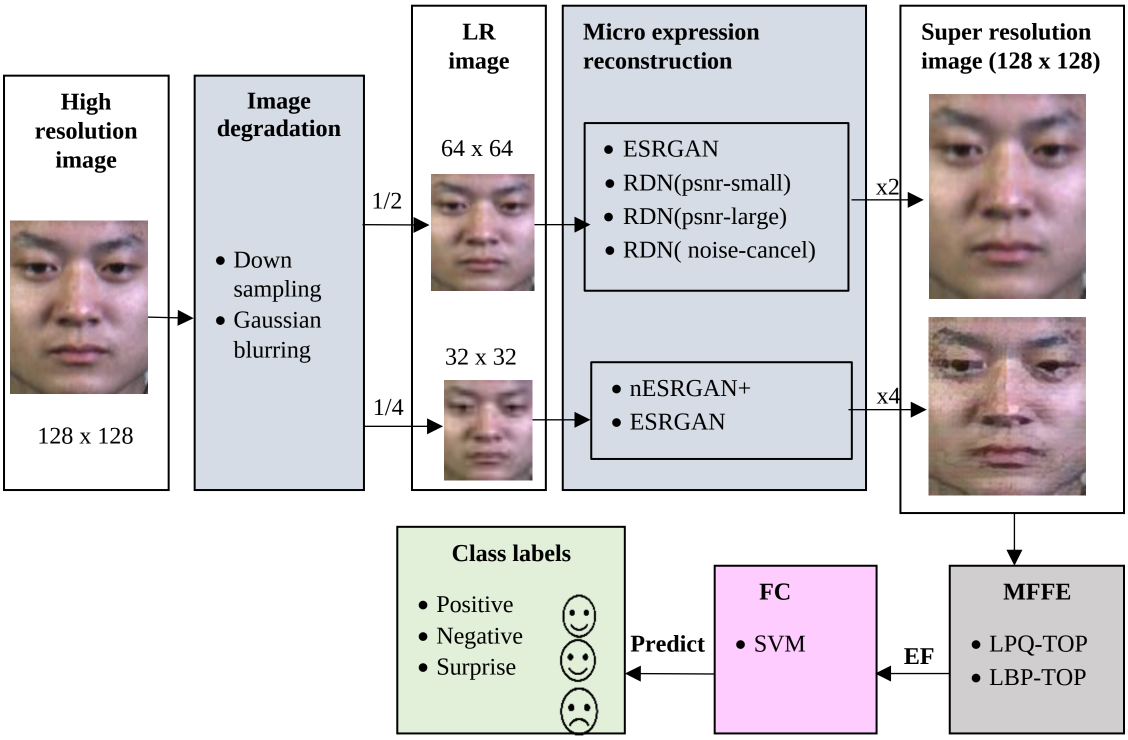
Dataset	**Subjects**	**Data Count**	**Classes**	**Class Label & Distribution**	**Facial Resolution**	**Speed (fps)**
CASME II	26	246	5	Happy- 32, Disgust -63, Surprise - 25,Repression - 27, Others - 99	280 x 340	200
SMIC-VIS	8	71	3	Positive -28, Negative-23, Surprise -20	130 x 160	25
SMIC-HS	16	164	3	Positive -51, Negative-70, Surprise -43	190 x 300	100

Table 6.2 Summary of notations, resolution and methods used.

Low Resolution		**SR Method Used**	**Scale Factor**	**Super Resolution**	
Notation Used	**Input Resolution**			**Final Resolution**	**Notation Used**
LR64	64 X 64	psnr-small	2	128x128	SR64
LR64	64 X 64	psnr-large	2	128x128	SR64
LR64	64 X 64	noise-cancel	2	128x128	SR64
LR64	64 X 64	ESRGAN	2	128x128	SR64
LR64	64 X 64	bicubic interpolation	2	128x128	SR64
LR32	32 X 32	ESRGAN	4	128x128	SR32
LR32	32 X 32	nESRGAN+	4	128x128	SR32
LR32	32 X 32	bicubic interpolation	4	128x128	SR32

6.2.1 Feature Extraction and Classification Parameter

For extracting features using the LBP-TOP method, the radii in two spatial directions i.e., for X and Y was set to 1. Similarly, the radii for the axis in the time domain, i.e., T, was set to 4. Further the neighbouring points for all the three planes i.e., XT, YT and XY was set to 8. Parameter settings for LPQ-TOP feature extraction include a neighbourhood with the size set to 5 and decorrelation set to 0.1. Experiments were conducted using multi-class classification with SVM to classify data from CASMEII as happy, sad, disgust, repression, and others whereas data from SMIC-HS and SMIC-VIS were classified into positive, negative and surprise.



Note:
EF: Extracted features, ESRGAN: Enhanced super resolution generative adversarial network, FC: Feature classification, LBP-TOP: Local binary pattern on three orthogonal planes, LPQ-TOP: Local phase quantization on three orthogonal planes, LR: Low resolution, MFFE: Micro facial feature extraction, nESRGAN+: Further improving enhanced super resolution generative adversarial network, psnr: Peak signal to noise ratio, RDN: Residual dense network, SVM: Support vector machine

Figure 6.2. Detailed illustration of pipeline to reconstruct micro expression images from low quality data and its recognition process.

6.3 Experiments, Results, and Analysis

A detailed illustration of the overall pipeline with specific techniques employed at different stages to achieve MER utilizing LR images is presented in Figure 6.2. In this section the experimental setup, parameters settings and results obtained using the methods and modules outlined in Figure 6.2 are discussed.

6.3.1 Image Degradation

From the specifications presented in Table 6.1; it can be clearly seen that facial resolution for all three databases vary. Therefore, to maintain uniformity across all datasets, all HR images are initially set to 128x128 following the work in [41] and will be referred as HR128 in this thesis. Down sampling these HR128 by factors 2 & 4 its corresponding sets of LR64 and LR32 are obtained for all the three databases individually. Instances of HR128 and their corresponding LR images at these two levels for all three databases are presented in Figure 6.3.

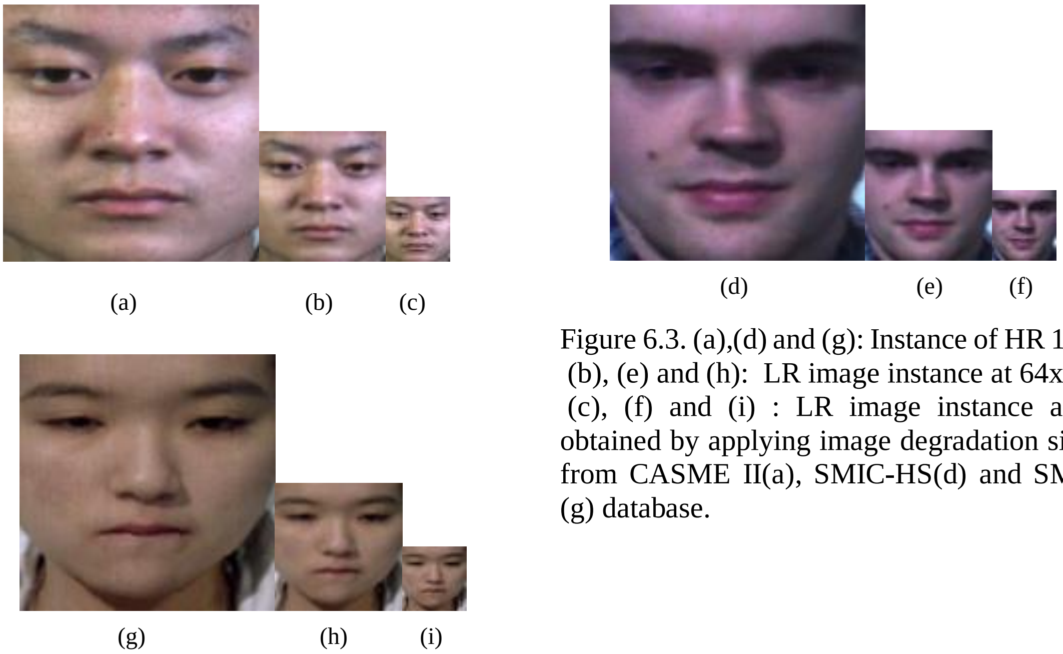


Figure 6.3. (a),(d) and (g): Instance of HR 128x128; (b), (e) and (h): LR image instance at 64x64 and (c), (f) and (i) : LR image instance at 32x32 obtained by applying image degradation simulated from CASME II(a), SMIC-HS(d) and SMIC-VIS (g) database.

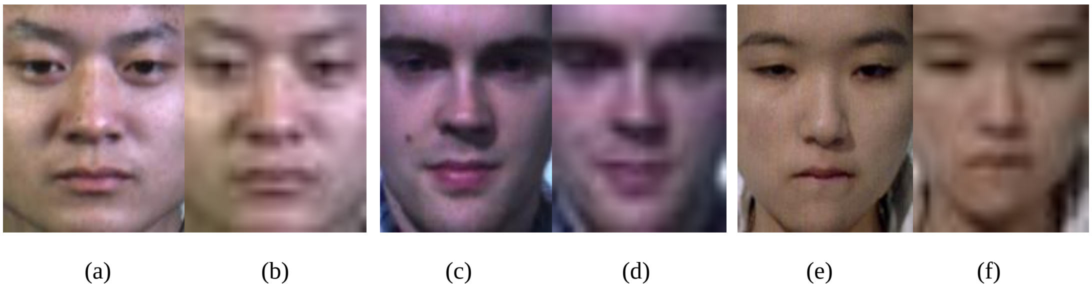


Figure 6.4. (a) (c) (e) Before applying degradation; (b) (d) (f) After applying degradation on CASMEII, SMIC-HS and SMIC-VIS respectively.

Images produced using such degradation models appear blurrier and of reduced quality. These effects along with the extent of loss of image details in the degraded images are clearly visible in the images shown in Figure 6.4. The LR images obtained by applying degradation on the corresponding HR images are depicted in Figure 6.4 (b), (d) and (f) respectively. Subtle expressions in HR images illustrated by Figure 6.4 (a), (c) and (e) are more obvious compared with the expressions on the degraded images presented in Figure 6.4 (b), (d) and (f). Evidently, one can notice loss of image details in the low-quality images generated by the degradation model. Thus, it can be said that both HR and LR images differ in terms of quality, as well as resolution. These newly created databases are now suitable to be used with SR algorithms.

6.3.2 Image Reconstruction

The experiments for image reconstruction were performed in two phases based on resolution of the LR image i.e., 64x64 and 32x32. In the first phase the LR64 instances of three databases were employed individually for the SR task. For each of these databases, five different sets of SR experiments were conducted by individually employing psnr-small, psnr-large, noise-cancel, ESRGAN and bicubic techniques with scale factor set to 2. The images obtained from all these methods after SR at this scale factor is referred to as SR64. Therefore, corresponding sets of SR database were obtained for every LR64 database.

Similarly, in second phase LR instances of three databases at 32x32 resolution were employed individually for SR task. For each of these databases, three different set of SR experiments were conducted by employing ESRGAN, nESRGAN+ and bicubic techniques with scale factor set to four. For all these SR methods the final resolution obtained was 128x128, to be referred as SR32.

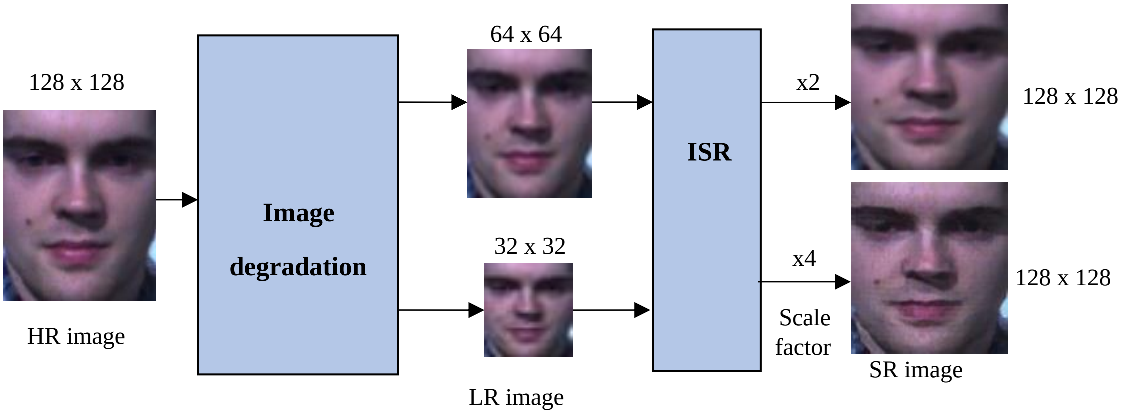


Figure 6.5. An abstract view to demonstrate resolution levels for image degradation and image super resolution model.

A basic sketch of this procedure followed for obtaining SR images from their corresponding LR images using different scale factor is illustrated in Figure 6.5. With reference to the basic model built using the RDN architecture given in Chapter 3 (see Figure 3.5 and Figure 3.6), parameter *D* refers to the number of RDB, C refers to the number of convolutional layers that are stacked inside a RDB, G refers to the number of feature maps of every convolutional layer that exists in RDBs, G_0 refers to the output filters i.e., the number of feature maps for convolutions that are outside of RDBs and of every RDB output. The values for these parameters employed in the psnr-large model is C=6, *D*=20, G=64, G_0=64 and scale factor x2, where the RDN network employed was trained on large image patches with large PSNR values. For psnr-small model parameter values used were C=3, *D*=10, G=64, G_0=64 and scale factor x2, where the RDN network employed was trained on small image patches with smaller PSNR values.

In the noise-cancel model parameters were set to C=6, *D*=20,G=64, G_0=64, scale factor x2, 3x3 kernel and activation function ParametricReLU for the RDN structure. The model was built by training the generator network on both VGG feature loss as well as adversarial loss. Different sessions of training were performed taking different sets of data. The discriminator employed in this method is the one illustrated in Figure 3.11. Different from the generator network, here it uses LeakyReLU activation function with its α parameter set to 0.2. Additionally, it also utilizes VGG loss which is based on ReLU and the network is optimized with Adam optimizer. Thus, using the psnr-small, psnr-large, and noise-cancel models, corresponding sets of SR64 datasets were obtained for every LR64 database.

ESRGAN is the fourth approach tested in this work built using ten RRDB, with three RDB in each of these RRDB. Furthermore, each of these RDB is built using four convolutional layers and inside each RDB there are 32 convolution output filters. Additionally, the architecture is fitted with 32 output filters for every RDB. With the learning rate set at 0.004, 100 decay frequency and decay factor at 0.5, the training parameters were set. The network was optimized using Adam optimizer and LReLU was used as activation function. The weight of the loss function is set to 1 for the generator and 0.003 for the discriminator during training. The GAN is optimized using Adam with β1 set to 0.9 and β2 set to 0.999 during this training phase. The discriminator is implemented with a kernel size of 3 and α set to 0.2 in LeakyReLU. The size of all convolutional layers is kept as 3x3 throughout the experiment. However, for local and global feature fusion, the size is set to 1x1. The model built was capable of upscaling images and supported scale factors two and four. Therefore, for every LR64 database its corresponding SR64 database containing SR images scaled by factor 2 were obtained. Similarly, for every LR32 database its corresponding SR32 database containing SR images scaled by factor 4 was obtained. The network structure employed here has been already discussed briefly in Section 3.4.2. For implementation of these four models discussed here the settings have been borrowed from [139].

The nESRGAN+ uses 3x3 convolutional kernels, 10 residual blocks and x4 scale factor and its architecture are described in Section 3.4.3. The loss function set at 0.005, decay factor at 0.01, learning rate set to 1x10^{-4} was considered. The Adam optimizer with parameters β1 and β2 set to 0.9 and 0.999 respectively was used. The model built was trained to upscale images by a scale factor set to 4. Most of the parameter settings of ESRGAN were kept intact while implementing nESRGAN+. These architecture and parameter settings for this model have been adapted from [124]. With this model for every LR32 database employing upscale factor four its corresponding SR32 database consisting of SR images was obtained, with resolution at 128x128.

To summarize, three LR64 databases, simulated from CASMEII, SMIC-HS and SMIC-VIS, were used in the SR experiments at scale factor two. For each simulated instance of LR64 databases, all four SR algorithms (i.e., psnr-small, psnr-large, noise-cancel and ESRGAN) were employed in four different sets of experiments to obtain four corresponding sets of super-resolved images. Similarly, another three sets of LR32 databases simulated from CASMEII, SMIC-HS and SMIC-VIS were used to perform the SR experiments at scale factor 4. For this case ESRGAN and nESRGAN+ models were employed to obtain corresponding SR images.

A summary of resolutions, different SR methods and database employed to perform various sets of experiment has already been presented in Table 6.2. The visual perception while assessing SR and HR image quality through human eyes may not always seem consistent hence, image quality needs to be assessed using quality metrics for these reconstructed images. These quality metrics are useful in determining performance of the SR algorithms.

6.3.3 Image Quality Assessment

Two widely used methods to assess SR image quality are peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) [121, 140], so these methods are chosen to assess quality of the super-resolved images obtained in the experiment. Values obtained for PSNR (measured in decibels, dB) and SSIM reflect the quality and rate of distortion of the reconstructed images compared with the original HR images. In simple terms they estimate structural correlation between the original and input image. SSIM is based on those structures that are typically visible in an image. The maximum value for SSIM is one which means closer the SSIM values are to one better is the reconstructed image quality. In case of PSNR, as its value increases, so too does the quality of reconstructed images. PSNR can be estimated by comparing the reconstructed image with an ideal image as follows [140].

$$PSNR = 10\,log\,10\left(\frac{max^{\,2}}{MSE}\right) \tag{6.1}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left(I(i,j) - I'(i,j)\right)^2 \tag{6.2}$$

In equation (6.1), max refers to maximum possible pixel intensity for a given input image and MSE refers to mean squared error. In equation (6.2) the number of rows is given by m, the number of columns is given by n, I is the HR original image and I' is the degraded image, i is the row index and j is the column index. SSIM is estimated using the following [140]:

$$SSIM(f,g) = l(f,g)c(f,g)s(f,g) \tag{6.3}$$

Here $l(f,g)$ estimates mean luminance closeness between two images f and g and is known as the luminance comparison function. Similarly, $c(f,g)$ estimates contrast closeness between two images and is known as the contrast comparison function. The correlation coefficient is estimated using the structure comparison function $s(f,g)$ between two images i.e., f and g. For SSIM, positive values can range between 0 to 1, where 0 means no correlation between two images and 1 means high correlation between two images. These two measures i.e., PSNR and SSIM computed for various SR algorithms for all three databases along with their analysis will be discussed in Section 6.3.4.

Table 6.3. Peak signal to noise ratio (PSNR (dB)).

Resolution	**SR Method**	**Dataset**		
		CASME II	**SMIC-HS**	**SMIC-VIS**
64 X 64	psnr-small	**34.7**	**37.41**	**36.67**
64 X 64	psnr-large	**34.59**	36.58	**36.12**
64 X 64	noise-cancel	31.06	30.38	31.15
64 X 64	ESRGAN	33.71	35.79	35.36
64 X 64	bicubic	34.7	36.45	35.57
32 X 32	ESRGAN	**27.65**	**29.5**	**28.95**
32 X 32	nESRGAN+	14.83	23.08	15.73
32 X 32	bicubic	30.83	32.3	31.64

Table 6.4. Structural similarity index (SSIM).

Resolution	**SR Method**	**Dataset**		
		CASME II	**SMIC-HS**	**SMIC-VIS**
64 X 64	psnr-small	**0.954**	**0.9827**	**0.9701**
64 X 64	psnr-large	**0.953**	0.9789	0.9672
64 X 64	noise-cancel	0.9261	0.9412	0.925
64 X 64	ESRGAN	0.9503	**0.9826**	0.9667
64 X 64	bicubic	0.9555	0.9771	0.9616
32 X 32	ESRGAN	**0.7811**	**0.8502**	**0.8189**
32 X 32	nESRGAN+	0.6559	0.7601	0.7527
32 X 32	bicubic	0.9032	0.9365	0.9111

6.3.4 Image Reconstruction Result Analysis

This section discusses the performance of all five SR algorithms based on the PSNR and SSIM metrics. Every set of super-resolved image sequence obtained were then compared with their corresponding HR images to obtain PSNR and SSIM values by utilizing equations (6.1), (6.2) and (6.3). The average PSNR and SSIM values computed for all reconstructed SR instances at SR64 and SR32 for all three databases are listed in Table 6.3 and Table 6.4 respectively.

Observing these image metrics, it is seen that psnr-small model was able to generate higher quality super resolved images across all databases at 64x64 image resolution. Specifically, the best reconstruction performance was obtained using the SMIC-HS database with this model achieving PSNR/SSIM values of 37.41dB/0.9827. On CASMEII database the performance of psnr-large model was very close to psnr-small model with PSNR/SSIM values behind that of psnr-small by a very small value of 0.11dB/0.001 respectively. For SMIC-HS database the images produced by psnr-small and ESRGAN model were structurally very close with a difference of 0.0001 SSIM value but psnr-small method produced 1.62dB higher PSNR value than ESRGAN. However, observing PSNR value alone it is seen that reconstruction performance of psnr-small and psnr large on SMIC-HS database is almost equal with a difference of only 0.83dB. Similar observation regarding PSNR metric can be made between psnr-large and ESRGAN model where later is lacking by a nominal value i.e., 0.79dB. Examining PSNR/SSIM values, a competitive performance between psnr-small and psnr-large models can be observed on SMIC-VIS database, where psnr-small model is ahead by 0.55dB/0.0029. When observing SSIM value alone for this database, structural performance of psnr-large and ESRGAN is almost same with later lacking by a value as small as 0.0005.

At 64x64 for bicubic and other reconstruction methods, inspecting PSNR values it is noticed that the bicubic performance is exactly the same as that of psnr-small method on CASMEII. All other reconstruction methods have produced image with lesser quality than bicubic method on this database. Using the SMIC-HS and SMIC-VIS databases the reconstruction performance of both psnr-small and psnr-large are superior to the bicubic method. Although reconstruction performance of the ESRGAN is below the bicubic but is still closer to bicubic in comparison to noise-cancel method. Examining the SSIM values in Table 6.4, one can notice that reconstruction performance of all the methods is inferior to bicubic method on CASMEII. Best value obtained on this database is 0.954, which is 0.001 below bicubic method, though the value is low by exceptionally small margin. On SMIC-HS and SMIC-VIS database the reconstruction performance of three methods i.e., psnr-small, psnr-large and ESRGAN are superior to bicubic method, thus these methods seem to perform well on both these databases. Clearly performance of the reconstruction methods on CASMEII is not satisfactory in comparison to other two databases.

Moving on to values obtained for 32x32 images it can be clearly seen that ESRGAN is able to outperform nESRGAN+ model across all databases. The best reconstruction performance given by ESRGAN at this resolution is for SMIC-HS with 29.5dB/0.8502 PSNR/SSIM metric values. The reconstruction performance of nESRGAN+ is far behind with its best performance metrics at 23.08dB/0.7601 for the same database.

Overall, it can be said that images reconstructed using psnr-small, psnr-large and ESRGAN model was almost similar where the psnr-small model was ahead by a narrow margin. Also, all these three models produced comparatively superior results compared to noise-cancel model for 64x64 images. Similarly, at 32x32 the ESRGAN produced far better results than nESRGAN+ model but overall performance was still lower than those obtained at higher resolution i.e., 64x64. While comparing results obtained using ESRGAN at both 64x64 and 32x32 levels, it can be observed that the model is able to produce comparatively better result by employing higher resolution images. For instance, highest performance given by ESRGAN is 35.79dB/0.9826 on SMIC-HS database for 64x64 images, however using the same model a dip in performance is noticed when lower resolution images (i.e., 32x32) are employed with PSNR/SSIM metrics value 29.5dB/0.8502. This clearly strengthens the common belief that the resolution employed at input directly affects the reconstruction performance of SR algorithms and same can be observed for ME images as well. Comparing reconstruction performance using PSNR/SSIM values for all these methods with bicubic it can be noticed that at 32x32 their performance is less superior than bicubic technique. Thus, the bicubic method seems to perform better at this level.

Images generated by employing various SR methods on three databases is presented in Figure 6.6 and Figure 6.7. From these results one can observe the visual quality of images generated by various SR algorithm at different scales. By comparing the visual quality of reconstructed image instances in both the figures, the significance of resolution and image quality employed during input phase seem obvious. For instance, visually the images obtained from LR64 after reconstruction presented in Figure 6.6 is much clearer and less noisy when compared to those images presented in Figure 6.7 that were obtained from their corresponding LR32 images. Thus, as expected, images obtained from lower resolution i.e., LR32 are quite poor compared to those obtained from a higher resolution i.e., LR64.

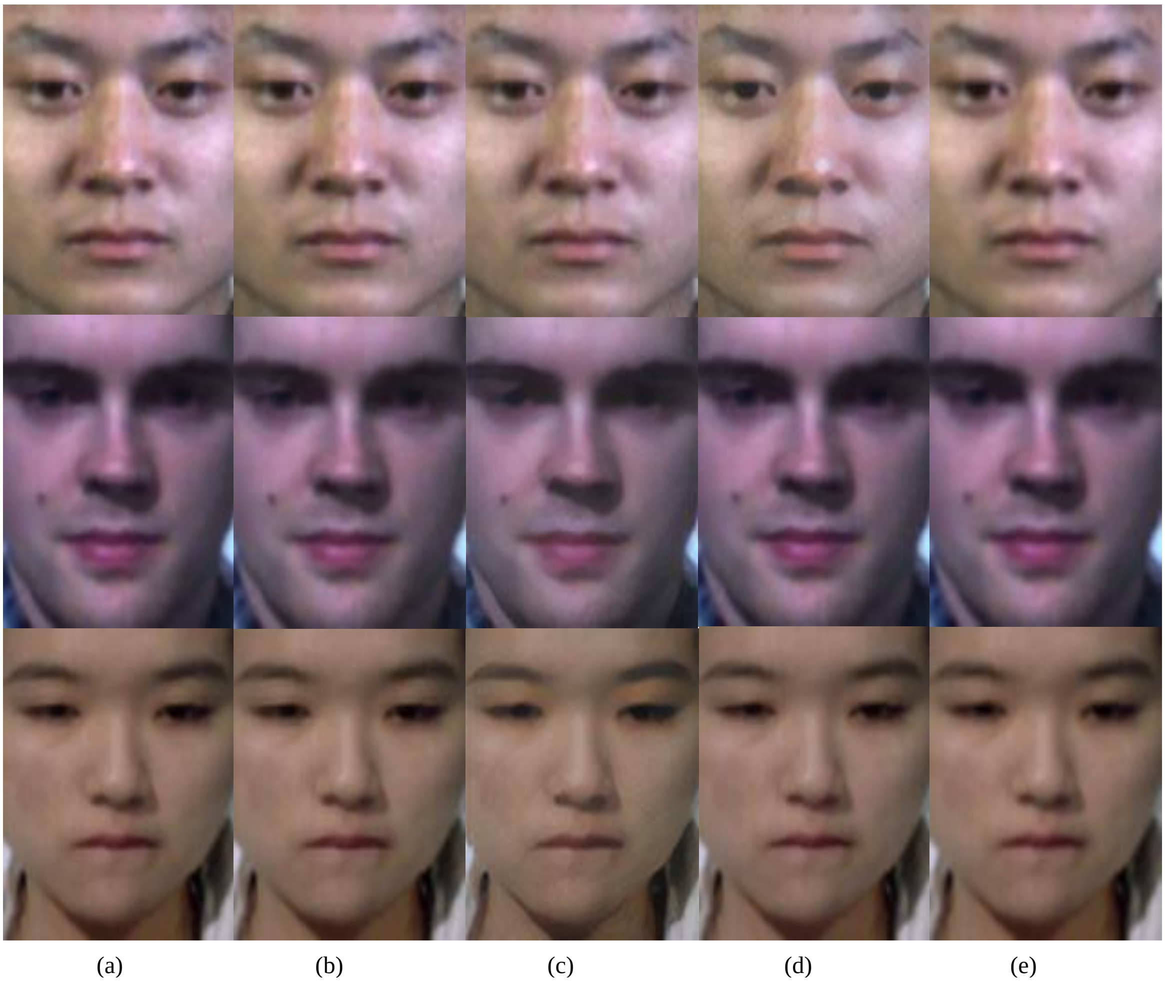


Figure 6.6. Images reconstructed using super resolution algorithms (a) psnr-small, (b) psnr-large, (c) noise-cancel (d) ESRGAN and (e) Bicubic interpolation with scale factor two for CAMSEII(top horizontal layer), SMIC-HS (middle horizontal layer) , SMIC-VIS (bottom horizontal layer).

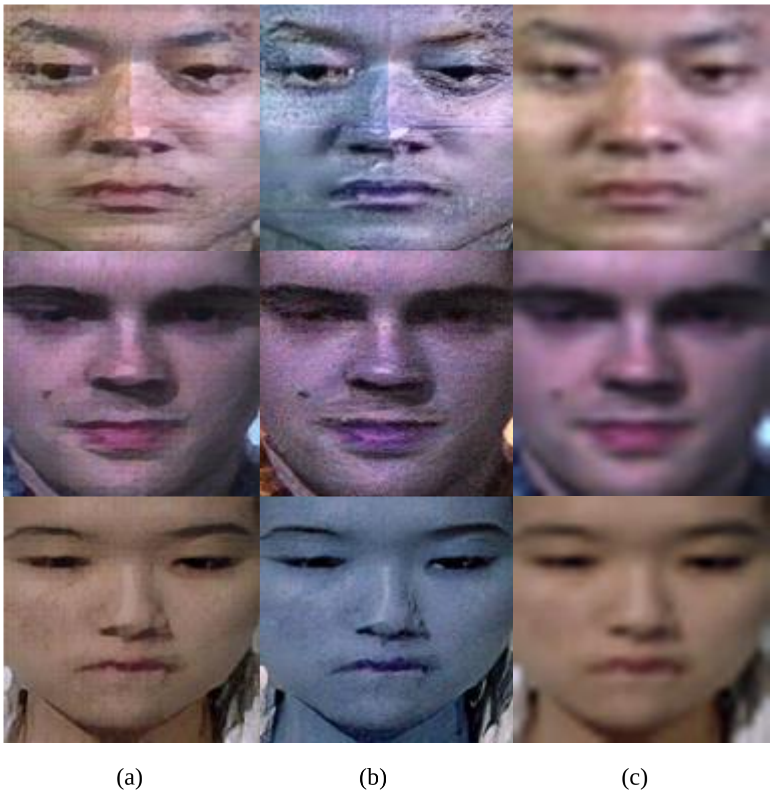


Figure 6.7.
Images reconstructed using super resolution algorithms ESRGAN (a),
nESRGAN+ (b) and
bicubic interpolation (c) ;
with scale factor set to four for CASMEII (top horizontal layer), SMIC-HS(middle horizontal layer) and SMIC-VIS(bottom horizontal layer).

Table 6.5. Accuracy % obtained before introducing super resolution algorithms.

	Accuracy %(ours)					
Database	**CASME II**		**SMIC-HS**		**SMIC-VIS**	
Resolution	**LBP-TOP**	**LPQ-TOP**	**LBP-TOP**	**LPQ-TOP**	**LBP-TOP**	**LPQ-TOP**
HR128	**48.16**	47.17	50.06	**52.43**	53.26	**61.26**
LR64	43.05	43.14	49.2	49.39	49.40	**57.26**
LR32	43.00	**41.05**	44.25	**48.17**	**45.54**	41.18

6.3.5 Recognition Result Analysis Before Super Resolution

In this section the recognition results obtained before applying SR algorithms, at various levels on both LR and HR cases using two different feature extraction methods, are discussed. Therefore, analysis in this section is based on the recognition performance obtained for HR128, LR64 and LR32 i.e., before introducing SR algorithms into the pipeline as depicted in Table 6.5. For all these instances feature extraction was carried out by employing both LPQ-TOP and LBP-TOP separately at different sets of experiments.

Recognition result obtained for HR128 (refer to Table 6.5) reveal that by employing LPQ-TOP technique the recognition framework was able to produce 2.37% and 8% higher recognition accuracy on SMIC-HS and SMIC-VIS database compared to LBP-TOP, but on CASMEII the LBP-TOP performed better by 0.99% over LPQ-TOP method.

For LR64 the recognition performance obtained (refer to Table 6.5) using both the extraction methods (before introducing SR) were almost similar, with LPQ-TOP being higher than the binary method by 0.09% and 0.19% on CASMEII and SMIC-HS database. However, on SMIC-VIS a much better performance was noticed by employing LPQ-TOP with recognition rate higher by 7.86%.

Taking LR32 in consideration performance recorded using LPQ-TOP was higher than LBP-TOP by 2.96% and 3.92% on CASMEII and SMIC-HS, however at same resolution the LPQ-TOP method reported 4.36% dip for SMIC-VIS database. At HR128, 61.26% accuracy was reported compared to 41.18% at LR32 on SMIC-VIS by employing LPQ-TOP and is the best performance recorded at this resolution level in these sets of experiment. At LR64 best performance was also on the same database i.e., SMIC-VIS employing the same method i.e., LPQ-TOP with 57.26% accuracy. For LR32 the highest accuracy recorded was 48.17% and obtained on SMIC-HS database.

Recognition performance comparison at various resolution levels employing both feature extraction techniques on all three databases is illustrated in Figure 6.8. With a gradual decrease in resolution level, a dip in recognition performance can be clearly noticed across all databases before employing SR methods. The results indicate that in general LPQ-TOP method works well on low quality images specially on SMIC variants.

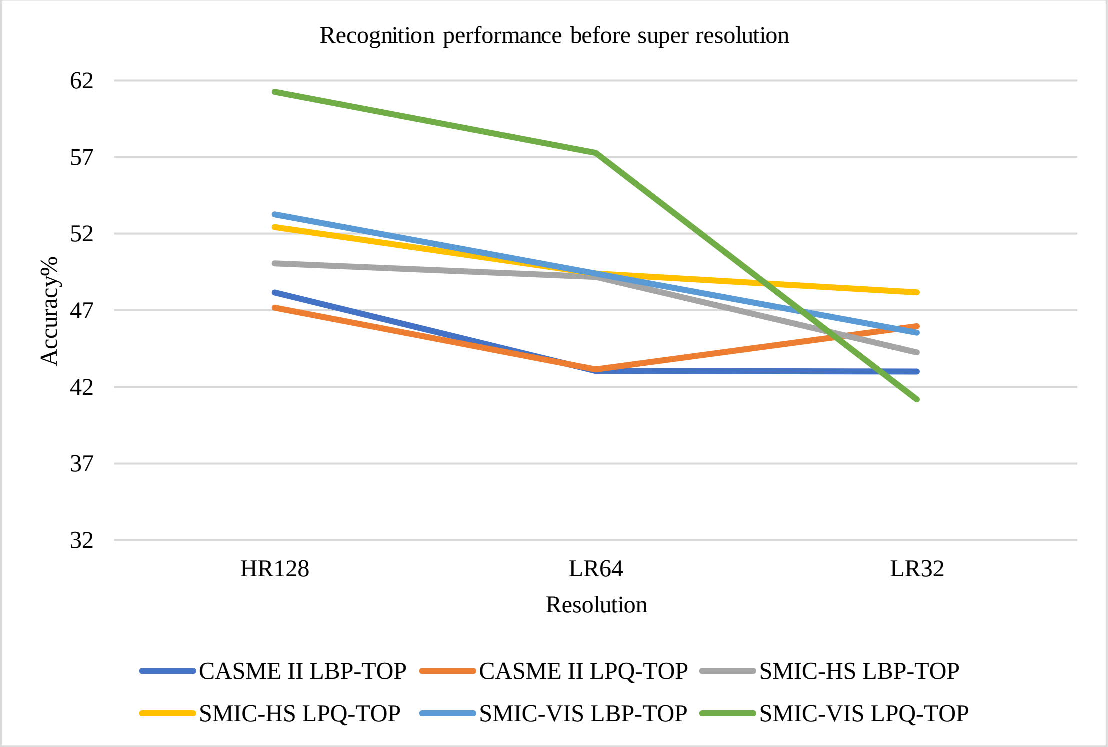


Figure 6.8. Recognition performance analysis on three databases at different resolutions before introducing super resolution.

6.3.6 Recognition Result Analysis Employing Super Resolution

In this section the overall recognition performance of the proposed pipeline after introducing various SR algorithms on three individual databases, recorded in Table 6.6, Table 6.7, and Table 6.8 is discussed. Further a comparative analysis of the recognition performance obtained for all three databases employed in this work is also discussed along with its results recorded in Table 6.9.

Table 6.6. Accuracy% obtained using various super resolution algorithms on SMIC-VIS

Resolution	SR Method	Accuracy %	
		LBP-TOP	**LPQ-TOP**
HR128	-	53.26	**61.26**
SR64	psnr-small	**59.62**	**61.63**
SR64	psnr-large	**56.67**	**61.40**
SR64	noise-cancel	**55.57**	60.33
SR64	ESRGAN	**56.60**	61.01
SR64	bicubic	55.23	61.40
SR32	ESRGAN	51.69	59.15
SR32	nESRGAN+	49.40	56.73
SR32	bicubic	52.11	60.03
LR64	-	49.40	**57.26**
LR32	-	**45.54**	41.18

6.3.6.1 Performance analysis on SMIC-VIS

Utilizing super resolved images, the best recognition performance recorded for the SMIC-VIS database is 61.63% as listed in Table 6.6. This is an increase of 4.37% using the psnr-small, reconstructed images at scale factor 2 with LPQ-TOP method compared to its corresponding LR64. The next best recognition was obtained employing psnr-large with the same extraction method resulting in an increase of 4.14% compared with its corresponding LR64. This was followed by ESRGAN with an increase of 3.75%. The lowest performance boost was for noise-cancel method with an increase of 3.07%.

Employing the LBP-TOP method, the best recognition performance was obtained at 59.62%. This is a boost of 10.22% obtained by employing psnr-small at scale factor 2 compared to its corresponding LR64. For the same extraction method when combined with the psnr-large method, the reconstructed images produced a recognition boost of 7.27% followed by the ESRGAN method with a boost of 7.2%. With a boost of 6.17%, noise-cancel produced the lowest improvement overall. Performance boost is obtained for all cases here but is still lower compared with that obtained employing LPQ-TOP method.

Reconstructing images with a scale factor of 4 with the ESRGAN method obtained a boost of 6.15% and 17.97% using LBP-TOP and LPQ-TOP respectively whereas with nESRGAN+ the accuracy was increased by 3.86% and 15.55% respectively. Therefore, boost in recognition performance obtained after employing SR algorithms at both scale factors is clear for this database. This analysis of recognition performance using the SMIC-VIS database employing various SR and extraction methods is illustrated in Figure 6.9.

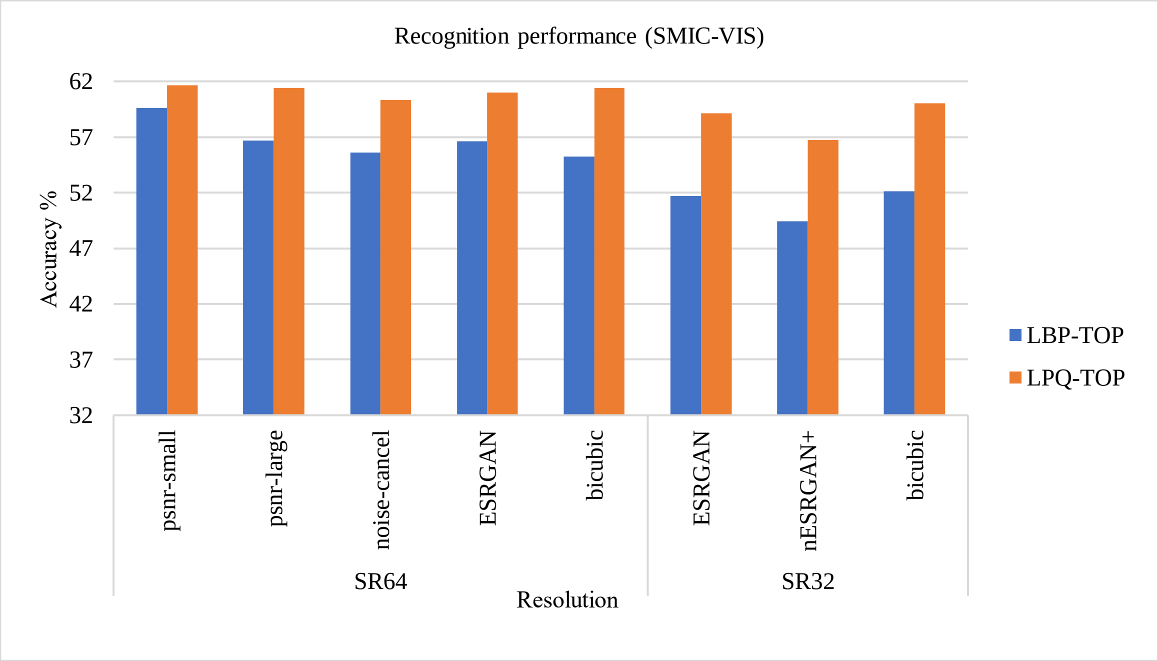


Figure 6.9. Recognition performance analysis on SMIC-VIS database after introducing super resolution.

Table 6.7. Accuracy% obtained for various super resolution algorithms on SMIC-HS.

Resolution	SR Method	Accuracy %	
		LBP-TOP	**LPQ-TOP**
HR128	-	50.06	**52.43**
SR64	psnr-small	**51.45**	**52.43**
SR64	psnr-large	**50.67**	52.00
SR64	noise-cancel	49.39	51.82
SR64	ESRGAN	**51.43**	**52.43**
SR64	bicubic	49.87	**52.43**
SR32	ESRGAN	49.82	50.60
SR32	nESRGAN+	49.24	50.00
SR32	bicubic	49.35	51.02
LR64	-	49.2	49.39
LR32	-	44.25	**48.17**

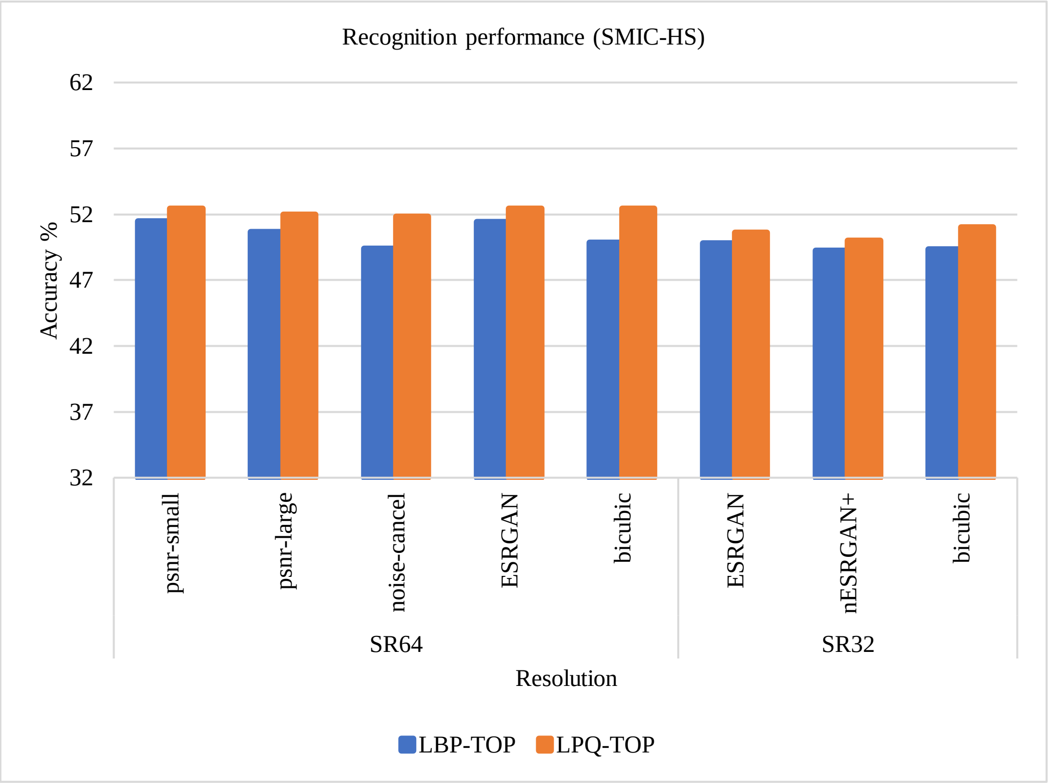


Figure 6.10. Recognition performance analysis on SMIC-HS database after introducing super resolution.

6.3.6.2 Performance analysis on SMIC-HS

The recognition performance analysis obtained using the SMIC-HS database after introducing SR is presented in Figure 6.10. Observing the values from Table 6.7, on SMIC-HS database the best recognition performance achieved after introducing SR is 52.43%. This performance was recorded for both psnr-small and ESRGAN reconstructed images. This value is 3.04% higher that its corresponding LR64 which clearly indicates the recognition performance boost obtained, therefore the overall pipeline has benefited from the SR algorithm. Next, by using psnr-large reconstructed images, a recognition boost of 2.61% was obtained for the same scale factor i.e., 2. Once again at this scale factor, the noise-cancel model generated images resulted in the lowest recognition performance boost of 2.43%. All these performance boosts were obtained when each of the SR methods constructed data were used along with LPQ-TOP method.

For the LBP-TOP method, the highest recognition accuracy obtained is 51.45% by employing ‘psnr-small’ reconstructed images, which is a boost of 2.25% compared with its LR64. The next best performance was obtained using the ESRGAN method which was behind psnr-small method by a mere 0.02% and with a boost of 2.23% for the same scale factor i.e.,2. This was followed by psnr-large generated images which produced a recognition boost of 1.47%. Consistent to the previous observations, the lowest performance was obtained by employing noise-cancel generated images with a recognition boost of only 0.19% compared to its corresponding LR64. Evidently from the results obtained for this database (refer Table 6.7), performance boost has been obtained for all instances across both feature extraction methods, though phase method is slightly better than the binary method.

Moving on to a scale factor four, the best recognition performance obtained is 50.6% which indicates a boost of 2.43% compared to its corresponding LR32. This performance is obtained for phase-based approach. The nESRGAN+ employed images were able to obtain a performance boost of 1.83% for same extraction method. Using binary extraction method, the recognition boost obtained was 5.57% and 4.99% for ESRGAN and its variant respectively. Therefore, for this case too, recognition boost has been achieved for all SR instances employing both feature extraction methods. Once again, these results confirm the benefit of employing SR algorithms for achieving a boost in overall recognition performance at both scale factors on the SMIC-HS database.

Table 6.8. Accuracy% obtained for various super resolution algorithms on CASME II.

Resolution	**SR Method**	**Accuracy %**	
		LBP-TOP	**LPQ-TOP**
HR128	-	**48.16**	**47.17**
SR64	psnr-small	**47.74**	**46.37**
SR64	psnr-large	47.34	**46.01**
SR64	noise-cancel	46.50	43.54
SR64	ESRGAN	**47.93**	**45.96**
SR64	bicubic	47.74	45.56
SR32	ESRGAN	43.05	41.93
SR32	nESRGAN+	40.04	34.67
SR32	bicubic	44.35	42.75
LR64	-	43.05	43.14
LR32	-	43.00	**41.05**

6.3.6.3 Performance analysis on CASME II

Using the CASMEII database, the best recognition performance obtained after introducing the SR model was 47.93% using the ESRGAN super resolution algorithm along with the LBP-TOP extraction method with a scale factor of 2 (refer Table 6.8). This reflects an obvious boost in recognition performance of 4.88% compared with its corresponding LR64 after employing

SR. The next best performance at the same scale factor was given when images reconstructed using psnr-small were used with the LBP-TOP method, a boost of 4.69%. This was followed by images reconstructed by psnr-large, with 4.29% boost in recognition performance. Once again, the lowest performance at this scale factor was obtained using the noise-cancel based images with a boost of 3.45%.

Employing the LPQ-TOP method for images reconstructed using psnr-small produced the best recognition performance of 46.37%, which is an increase of 3.23% over its corresponding LR64 and yet it is still 1.37% lower than the performance obtained using the LBP-TOP method. The boost in recognition obtained employing this extraction method with psnr-large is 2.83%, noise-cancel is 0.4% and ESRGAN is 2.82%. Although performance boost is achieved in all these cases compared with the corresponding LR64 images, the accuracies are lower than those obtained employing LBP-TOP approach. Therefore, for these cases recognition obtained using LBP-TOP seemed better than the phase method.

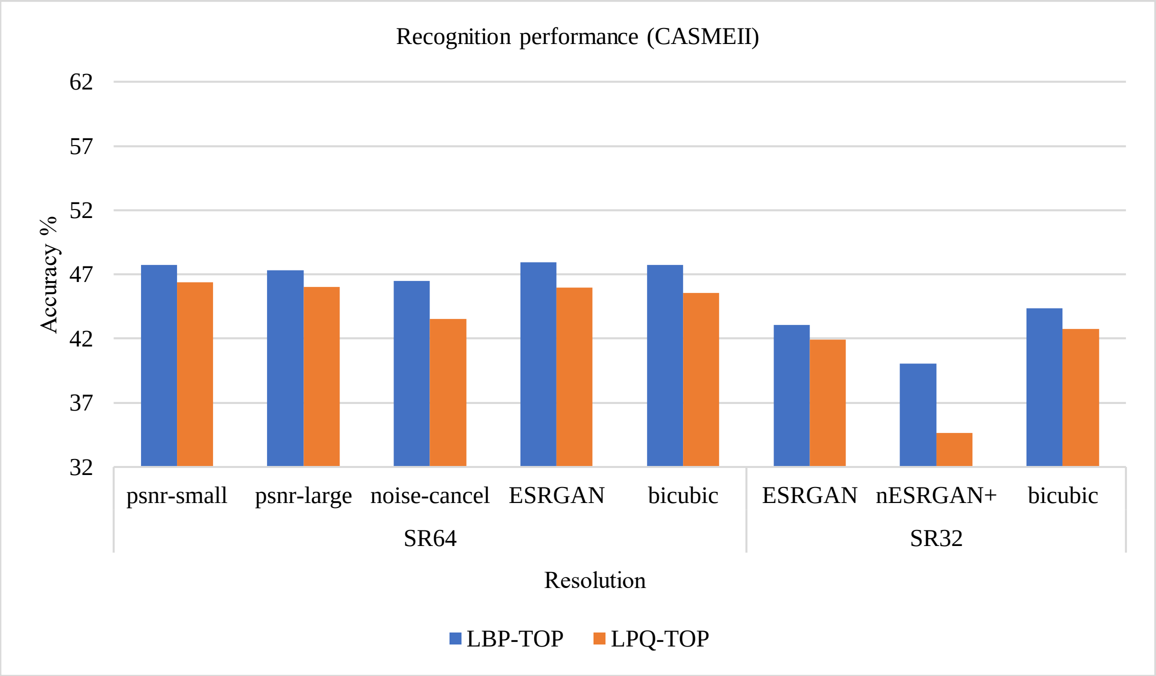


Figure 6.11. Recognition performance analysis on CASME II database after introducing super resolution.

At scale factor four, images reconstructed with ESRGAN method employed with LBP-TOP obtained performance boost of 0.05% whereas with LPQ-TOP it was 0.88%. Though a higher boost is achieved using LPQ-TOP method, the overall recognition performance is still better with LBP-TOP method for this case. The lowest performance was obtained for nESRGAN+ reconstructed images with recognition values recorded below its corresponding LR32 images for both the extraction methods on this database. This analysis of performance using the CASMEII database employing various SR and extraction methods is illustrated in Figure 6.11.

Table 6.9. Accuracy% comparison for various super resolution algorithms across all datasets and methods.

		Accuracy %							
		CASME II			**SMIC-HS**			**SMIC-VIS**	
Resolution	**SR Method (ours)**	**LBP-TOP (ours)**	**LPQ-TOP (ours)**	**Fast LBP-TOP [41]***	**LBP-TOP (ours)**	**LPQ-TOP (ours)**	**Fast LBP-TOP [41]***	**LBP-TOP (ours)**	**LPQ-TOP (ours)**
HR128	-	**48.16**	**47.17**	48.18	**50.06**	**52.43**	50.00	53.26	61.26
SR64	psnr-small	**47.74**	**46.37**	48.18	**51.45**	**52.43**	52.44	**59.62**	**61.63**
SR64	psnr-large	47.34	**46.01**		**50.67**	52.00		**56.67**	**61.40**
SR64	noise-cancel	46.50	43.54		49.39	51.82		**55.57**	60.33
SR64	ESRGAN	**47.93**	**45.96**		**51.43**	**52.43**		**56.60**	61.01
SR64	bicubic	47.74	45.56		49.87	**52.43**		55.23	61.40
SR32	ESRGAN	43.05	41.93	44.53	49.82	50.60	51.83	51.69	59.15
SR32	nESRGAN+	40.04	34.67		49.24	50.00		49.40	56.73
SR32	bicubic	44.35	42.75		49.35	51.02		52.11	60.03
LR64	-	43.05	43.14	44.94	49.2	49.39	50.00	49.40	**57.26**
LR32	-	43.00	**41.05**	44.13	44.25	**48.17**	46.95	**45.54**	41.18

Note: Bold indicates best values obtained in this thesis in general.
For SR methods bold indicates best values obtained when compared with bicubic method.
* Super resolution method used is patch-based and pixel-based regularization which is different from the deep learning-based approach used in this thesis.

6.3.6.4 Performance comparison across all methods and databases

Observing the values presented in Table 6.9, it is noticed that SR images at a scale factor of 2, when used with LPQ-TOP, produced much better recognition performance than the binary method using SMIC-VIS and slightly better for the SMIC-HS database. For instance, examining the best performance on all databases it is seen that recognition is higher by 2.01% using SMIC-VIS and 0.98% using SMIC-HS employing LPQ-TOP, compared with the binary method. However, using the CASMEII database, the LBP-TOP method seems to perform better with 1.37% higher accuracy than the phase quantisation approach. Observing the overall analysis presented in Figure 6.12, for methods utilizing the RDN architecture, employing images reconstructed by the psnr-small method seems to consistently provide the best recognition performance across all databases. The performance of ESRGAN employed images was also at par with images constructed using psnr-small on the SMIC-HS database. However, using the SMIC-VIS database, the psnr-large model constructed images performed slightly better than the ESRGAN model constructed images. The lowest performance was consistently obtained by employing images reconstructed by the noise-cancel approach across all three databases. Therefore, at this scale factor, all three SR approaches, i.e., psnr-small, ESRGAN and psnr-large, seem to be very competitive and performed consistently better than noise-cancel method.

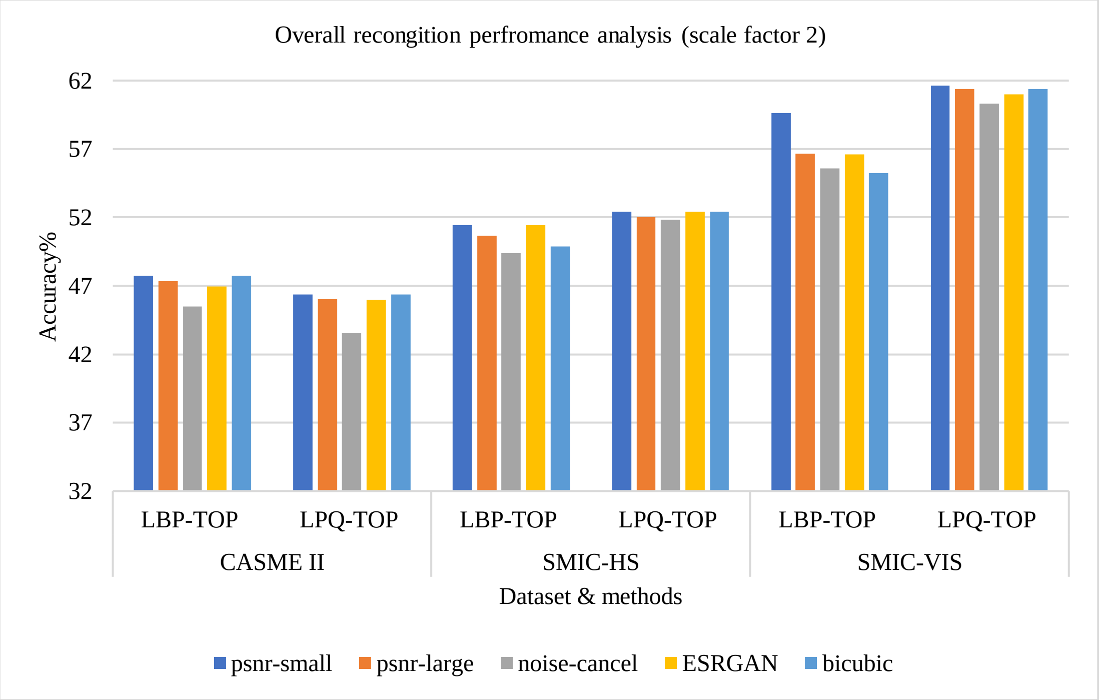


Figure 6.12. Comparison of recognition accuracy employing various instances of super resolution and feature extraction techniques on three micro expression databases for scale factor 2.

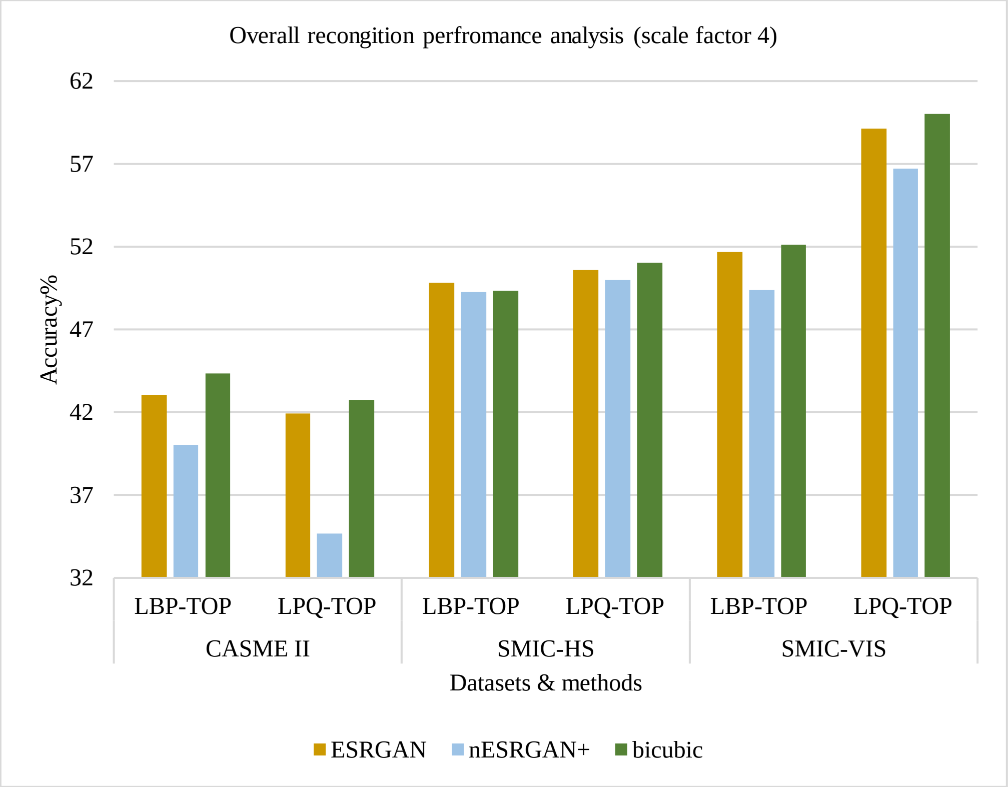


Figure 6.13. Comparison of recognition accuracy employing various instances of super resolution and feature extraction techniques on three micro expression databases for scale factor 4.

Observing the results in Table 6.9 and Figure 6.13, at a higher scale factor of 4, images reconstructed using ESRGAN seem to consistently outperform its corresponding variant i.e., nESRGAN+ across all databases. Once again both methods performed better on SMIC-VIS and SMIC-HS. The lowest performance was obtained on the CASMEII database for both these SR methods.

To have a fair comparison among results obtained using various SR methods in this work, a comparison with bicubic interpolation results is also made. For SMIC-VIS database, SR images reconstructed at scale factor 2 (refer Table 6.9 and Figure 6.12) by all SR methods using LBP-TOP seems to work fairly well with recognition accuracies higher than those obtained using the bicubic method. Employing LPQ-TOP on the SMIC-VIS database, the psnr-small and psnr-large methods performed better than bicubic, whereas the noise-cancel and ESRGAN performances were lower than the bicubic method. At scale factor four recognition performance using images reconstructed by the bicubic method (refer to Table 6.9 and Figure 6.13) was better which is consistent with the image quality metrics obtained for this instance. For the SMIC-HS database, both the psnr-small and ESRGAN methods were able to produce results at par with the bicubic method at scale factor two when combined with the LPQ-TOP method (refer to Table 6.9 and Figure 6.12). Likewise, when combined with the LBP-TOP method the images reconstructed using all SR methods performed better than the bicubic reconstructed images with the exception of noise-cancel which performed lower than bicubic. At scale factor four recognition performances of both SR methods were lower than bicubic method when combined with LPQ-TOP (refer to Table 6.9 and Figure 6.13).

Using the CASMEII database at scale factor two (see Table 6.9 and Figure 6.12), psnr-small and ESRGAN, when used with the LBP-TOP method produced recognition performances better or equal to that of the bicubic method, however using the psnr-large and noise-cancel approaches, recognition was marginally lower than the bicubic method. Most of the SR methods when combined with the LPQ-TOP approach for this database performed similar or better than the bicubic method whereas with the LBP-TOP method only psnr-small and ESRGAN seemed to perform better than bicubic. At scale factor four (refer to Table 6.9 and Figure 6.13) the bicubic method was slightly better than ESRGAN based approach but much better than nESRGAN+ based approach.

The overall comparison of recognition accuracy employing different SR and feature extraction techniques on three ME database is illustrated in Figure 6.12 and Figure 6.13. To summarize, both the psnr-small and ESRGAN methods, when combined with LBP-TOP, perform best using CASMEII at a scale factor of 2, whereas at a scale factor of 4 ESRGAN worked best. Using the SMIC-HS database psnr-small and ESRGAN, when combined with LPQ-TOP, performed the best. Using the SMIC-VIS database psnr-small and psnr-large combined with LPQ-TOP performed well. Therefore, most of the SR methods were able to produce results better than the bicubic method at a scale factor of 2 across all databases, whereas at a scale factor of 4 using the SMIC-HS database and ESRGAN when combined with LBP-TOP produced better recognition results than the bicubic method.

6.4 Summary

As a solution and contribution to the MER system employing low quality images, a new pipeline exploiting DL and GAN based SR approaches is built. The contributions of this pipeline include an extensive analysis of five different SR algorithms particularly for reconstructing ME images. Clearly, all the SR models employed have been able to successfully reconstruct the facial details, though the image quality obtained is varying. Further a comprehensive analysis of two different feature extraction methods employing images produced by each of the SR methods is another contribution of this chapter.

The experiments were performed on publicly available three popular ME database with favourable results. Examining the overall performance, these positive results are a good indicator to ascertain the effects of DL and GAN based SR technique for boosting facial ME image details. Certainly, the classification accuracy was influenced by the size and quality of image reconstructed across all databases, and same is reflected in the results obtained. Two limitations of this work are, first the tests consider two resolutions for LR images, however in the future resolutions lower than these can also be assessed. Second, data imbalance has not been addressed and as such the work can be substantiated by incorporating a suitable approach with more uniform datasets in the future.

Nevertheless, the results achieved are promising and can be extended further by evaluating more SR algorithms with additional scale factors. The results obtained can also be used as a general guideline to widen the usage of suitable SR technique for such specific applications. Acquiring good facial resolution with low-cost surveillance cameras may not always be realistic in day-to-day life especially when faces to be captured are distant from the camera, this directly affects the quality of facial details obtained. Therefore, to overcome resolution issues that may exist in ME obtained in similar unfavourable settings, utilizing such deep learning SR based reconstruction algorithms together with recognition framework seems a feasible option.

Chapter 7

Conclusion

7.1 Introduction

ME analysis has a long history since its discovery in the field of psychology and today has far-reaching applications ranging from medical, security, academic, to business and beyond. To guide any research towards ME analysis, one needs to be able to determine its finer details. Therefore, incremental progress in the research has helped in developing and accumulating a variety of approaches that can achieve this. Meanwhile development of an automated system capable of functioning in a real-world environment will have wider applications and supplement cutting edge technologies. The primary focus of this thesis is designing methods that can effectively model spatial and temporal micro patterns to achieve MER. Modelling a system capable of catering to problems inherent while capturing ME in real world environments is an additional highlight of this thesis.

LPQ based methods have previously been explored in ME for AU detection [134] and designing cross-databases [89], however this thesis investigates it as an ME feature extractor across several independent ME databases solely to accomplish expression recognition. Furthermore, it is combined with temporal interpolation and video magnification and results clearly demonstrate that its performance is as competitive as any other classical feature extraction method. Since the LPQ-TOP method has the ability to detect changes in the temporal as well as spatial domain, employing it for ME analysis in this thesis seemed a good fit.

Significantly improved performance was obtained for the majority of the databases by pooling magnification, TIM, and LPQ-TOP. From the experimental results obtained, use of LPQ-TOP for ME feature extraction is further substantiated. Recognition of ME using this proposed combination is the first investigation successfully realized on seven spontaneous ME databases.

Low quality recordings captured in real-world environments is a major issue, especially within a surveillance setup. This thesis addresses the issue by adopting a more objective approach. Therefore, a novel pipeline that incorporates ISR algorithms (based on DL and GAN) as one of its constituent components, prior to MER, is built. For investigating this concept, five different SR algorithms were tested on three different ME databases for two levels of LR micro expression images.

7.2 Research Findings

This section provides a summary of the research objectives outlined in Section 1.5 and their corresponding outcome. These findings are based on the research and experiments undertaken and provide evidence-based reasoning for the outcome achieved. Investigating the efficiency of recognizing different classes of ME from the available databases employing features from the temporal domain and training models using machine learning was the first objective laid out in the thesis. This was achieved by utilizing the combination of TIM and LPQ-TOP technique for extracting spatio-temporal information from XY, XT and YT planes, training a SVM algorithm on those extracted features and finally classifying the data into relevant classes. LPQ-TOP has already been used for ME action unit detection [134] and designing cross-databases [89], however in this thesis it is employed solely to achieve ME recognition and classification without incorporating the concept of action units and cross database. Experimental results demonstrate its potential as a feature extraction method and suggests it to be as competitive as other classical feature extraction methods. This preliminary experiment was performed using the CASME II database with accuracy of 61.16% which is comparable to those reported in [47,141,142]. This database was chosen for preliminary experiments due to it being one of the most widely used ME datasets. It was concluded that combining the interpolation and phase quantization technique with SVM yields acceptable recognition accuracies for ME.

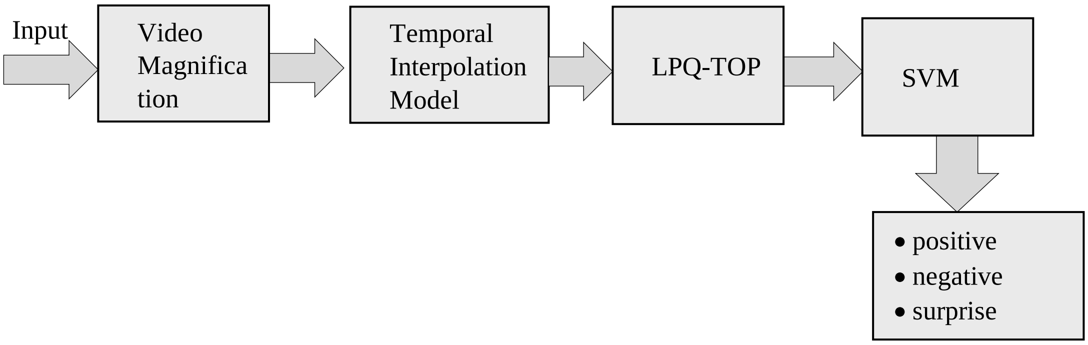


Figure 7.1. Proposed pipeline for micro expression recognition.

Exploring the advantage of utilizing video magnification with interpolation and a phase quantization technique by conducting suitable experiments is the second objective. This was accomplished by further employing the TIM and the LPQ-TOP method along with EVM on seven different spontaneous ME database i.e., SMIC (HS, NIR, &VIS), CASME, CAS(ME)^2, CASMEII and SAMM. The proposed pipeline involving magnification and other methods experimented in this thesis is illustrated in Figure 7.1. The features extracted using the LPQ-TOP technique on all these databases before introducing magnification further ascertained its edge as a feature extraction method. Moreover, utilizing it again after introducing EVM successfully improved the recognition accuracy for all databases, with highest improvement recorded for the CASMEII database at 13.34% and average improvement achieved across all databases at 6.14%. It was concluded that EVM was effective in accentuating the muscle motion for ME enabling more effective feature extraction by LPQ-TOP evident from the boost in recognition accuracy achieved in the majority of the cases. Evidence presented in Section 5.4.2 and Section 5.4.3 clearly substantiates these findings.

The third objective was to explore and compare two feature extraction approaches for describing the ME patterns in LR images. This was achieved by simulating LR micro expression databases for CASMEII, SMIC (HS & VIS) at resolution levels 64x64 and 32 x32. The LBP-TOP and LPQ-TOP methods were then employed to describe the ME patterns within these LR images. A simulated database was created to address the existing limitation of publicly available ME databases where image data have HR and are of good quality, therefore fail to resemble recordings acquired in a real-world environment. The LPQ-TOP method, when combined with SVM, gave best recognition performance for the SMIC-VIS images (LR) at 64x64 resolution compared with LBP-TOP, but same observation could not be seen for 32x32. The LBP-TOP and SVM method seems to work best for LR images taken from the SMIC-HS and CASMEII databases. Experimental results obtained in Section 6.3.5 justify these outcomes.

The fourth objective was to study the impact of low quality and LR during the recognition process. For this purpose, the two LR considered were 64x64 and 32x32, obtained by degrading the HR image of 128x128. The recognition accuracy obtained for ME at 128x128 is compared with those obtained at 64x64 and 32x32 for all three simulated databases i.e., SMIC-HS, SMIC-VIS and CASME II. Knowing that the resolution of an image has direct impact on recognition performance, it is vital to be able to quantitatively demonstrate this. The recognition accuracy percentage obtained at HR128 was 5.11/5.16 , 0.86/5.81, 3.86/7.72 higher than its corresponding LR accuracy percentage at 64/32 levels for CASME II, SMIC-HS and SMIC-VIS databases respectively. As anticipated the worst recognition performance for all the databases was for images at 32x32 level and was at least 5.16% lower than its HR images. This analysis is based on the results obtained using LBP-TOP and SVM on HR and LR images. Similar trends are observed for results obtained using LPQ-TOP and SVM with 4.03/6.12, 3.04/4.26 and 4/20.08 higher than its corresponding LR accuracy percentage at levels 64/32 for CASME II, SMIC-HS and SMIC-VIS databases respectively. Again, decline in the recognition performance is observed with decreasing resolution. Clearly, results obtained in Section 6.3.5 once again justifies these outcomes and confirms the declining recognition performance for MER with diminishing input image resolutions.

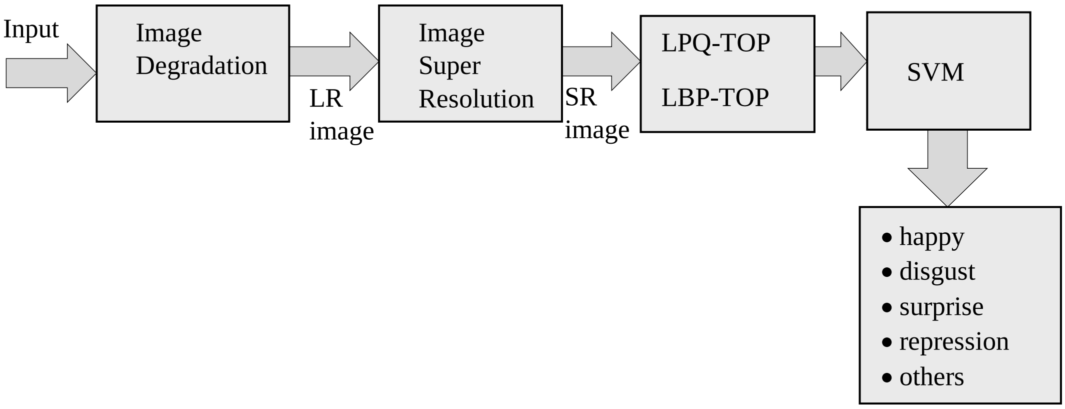


Figure 7.2. Proposed pipeline for micro expression recognition by applying SR algorithm on LR images.

The final objective was to investigate the contribution of SR algorithms and evaluate the performance of the proposed methods on several standard databases. This was achieved in Section 6.3.6 with the creation of databases containing super resolved images. These images were obtained by employing various DL and GAN based SR algorithms. The proposed pipeline tested in this thesis to achieve micro expression recognition utilizing LR images and SR algorithms is illustrated in Figure 7.2. Further, by using the image quality metrics PSNR and SSIM, the quality of the reconstructed images was assessed. This helped to understand the performance of the SR algorithms on ME images before initiating the recognition process. Observing these metrics, among the chosen SR methods utilizing RDN trained with small PSNR gave the best reconstruction performance across all the databases for resolution level 64x64 whereas at 32x32 reconstruction images obtained from ESRGAN method was preferable. Interestingly, reconstruction performance of ESRGAN and psnr-large model were as competitive as psnr-small. The quality of these reconstructed images obtained from various SR methods was clearly reflected on its corresponding recognition results. For instance, at 64x64 the best reconstructed images were produced by the 'psnr-small' method and employing these produced the best recognition performances across all three databases. Similarly, at 32x32 ESRGAN produced best quality SR images and consequently employing them produced better recognition performance compared to other similar methods. Overall, it can be concluded from these SR based ME experiments that utilizing DL and GAN based SR algorithms helps in regaining ME details in an image thereby boosting the recognition performance. Moreover, choosing an appropriate SR approach can significantly favour recognition methods.

7.3 Limitations

Although the techniques and algorithms utilized during research and experiments have successfully met all the objectives, limitations specific to this work should be discussed. First, the phase quantization method is computationally intensive, therefore for real-time processing its optimization maybe essential. Moreover, utilizing DL techniques during the extraction process can deliver better performance. Secondly, the database utilized consists of videos with minimal noise, as such regulating amplification factor delivered videos with magnified noise that could be dealt with easily. However, in real-world situations, obtaining noise free videos is challenging as such utilizing magnification in such scenario may not be appropriate. Thirdly, all the experiments performed by employing images consist of full-frontal faces, but in a natural setup the faces being acquired can have different poses, angles and sometimes suffer from occlusions. Building ME recognition methods that can cater for these unique conditions inherent to natural recordings are required moving forward. Although the low-quality image issue, particularly for ME images, has been addressed in this thesis to some extent, it must be mentioned that every image consists of a single face with empty background. This is unlikely in reality, considering surveillance recordings where most of the time several other objects appear in the background along with faces. Also, recordings consisting of multiple faces in the same frame is highly likely in such real-world scenarios. Developing algorithms that can adapt to such input can give an edge to the overall low quality ME based research and analysis.

7.4 Future Work

Research involving automatic ME analysis is still at an early stage compared with other applications of facial image analysis. The future of ME has a huge potential to continue creating new state-of-the-art solutions. While the work in this thesis represents a significant addition to the current ME domain, much more work is still needed. To expand the knowledge base and develop high end methodologies, some interesting extensions to this work are discussed in this section.

First, utilizing cross-databases to increase the volume of data and making recognition methods more adaptable to varying domains can help to overcome the bias arising due to culture specific databases. This encourages development of a more generic framework that is not limited by the data format and scale.

Second, to have more relevance to the real word situation employing data from "In-the wild" databases can be useful. They often contain data captured at varying pose, with occlusion, positioned at different angle, illumination changes and with multiple background objects. Existing approaches can be extended with appropriate modifications to adapt to such ME databases.

Third, simulating datasets that contain images with resolutions lower than 32x32 can help to verify the robustness of SR based ME recognition pipeline. Also, adding more SR algorithms into the pipeline will give more comprehensive results, further validating the benefit of incorporating the SR concept for ME recognition with low quality images.

In the work described throughout this thesis, three emotion labels were used for the first set of experiments and five emotion labels for the second. Therefore, as a fifth extension to this thesis, a greater number of labels can be incorporated into the list of emotions and combined with verbal and/or vocal data to have all-inclusive emotion analysis.

7.5 Concluding Remarks

Throughout this thesis, the work presented forms contributions to automatic ME recognition, where emphasis was laid on ensuring the algorithms are rigorously tested on several databases to establish robustness of the overall approach. The contributions demonstrate that the results obtained were promising for the future of ME recognition. There is no doubt that research on facial MER has witnessed considerable growth and progress in the last decade, however lack of substantial volume of ME database is still a roadblock. While the research challenges in automatic MER remain, the development of emerging tools encouraging application to real-world problems is growing steadily. Therefore, from the literature and contributions presented through this work, ME recognition in the real world may soon be the dominant research theme in ME analysis. As such designing methodologies with good processing speed might become essential to enable its application to the real world. Furthermore, technical, and academic advancements of the MER field will have a profound influence on various disciplines like medical, business, academic, e-commerce and many more. Therefore, investigating in the field of ME is noteworthy and the findings often earn credit across various disciplines.

Bibliography

[1] Zhang, Y., Zhang, Y., Molina, J., Giordano, R., & Bromley, J. (2011). Face, image, and analysis. In Y. Zhang (Ed.), Advances in Face Image Analysis: Techniques and Technologies (pp. 1-15). IGI Global. https://doi.org/10.4018/978-1-61520-991-0.ch001.

[2] Abdolrashidi, A., Minaei, M., Azimi, E., & Minaee, S. (2020). Age and gender prediction from face images using attentional convolutional network. ArXiv, abs/2010.03791.

[3] Ingole, A.L., & Karande, K.J. (2018). Automatic age estimation from face images using facial features. 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN), 104-108.

[4] Singh, G., & Goel, A.K. (2020). Face Detection and Recognition System using Digital Image Processing. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 348-352.

[5] Ran, P. and Wang, H.(2020). Real-time eye blink detection based on python," The 8th International Symposium on Test Automation & Instrumentation (ISTAI 2020), pp. 98-100, doi: 10.1049/icp.2021.1312.

[6] Woodo Lee, Nokyung Park, Jakyung Koo, Pilgu Kang. (2021). Concentration Levels. IEEE Dataport. https://dx.doi.org/10.21227/mgdg-4z85.

[7] Xin X, Lin X, Yang S, Zheng X (2020) Pain intensity estimation based on a spatial transformation and attention CNN. PLoS ONE 15(8): e0232412. https://doi.org/10.1371/journal.pone.0232412.

[8] Kumar, S., Varshney, D., Dhawan, G., & Jalutharia, H. (2020). Analysing the Effective Psychological State of Students using Facial Features. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 648-653.

[9] Jaiswal, A., Krishnama Raju, A., & Deb, S. (2020). Facial Emotion Detection Using Deep Learning. 2020 International Conference for Emerging Technology (INCET), 1-5.

[10] Simcock, G., McLoughlin, L. T., De Regt, T., Broadhouse, K. M., Beaudequin, D., Lagopoulos, J., & Hermens, D. F. (2020). Associations between Facial Emotion Recognition and Mental Health in Early Adolescence. International Journal of Environmental Research and Public Health, 17(1), 330. https://doi.org/10.3390/ijerph17010330.

[11] Tripathi, A., Ashwin, T.S., & Guddeti, R.M. (2019). EmoWare: A Context-Aware Framework for Personalized Video Recommendation Using Affective Video Sequences. IEEE Access, 7, 51185-51200.

[12] Suguna, R., Devi, M.S., Kushwaha, A., & Gupta, P. (2019). An Efficient Real time Product Recommendation using Facial Sentiment Analysis. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1-6.

[13] Palastanga, N., & Soames, R. (2012). Anatomy and human movement: structure and function (6th ed.). Edinburgh: Churchill Livingstone.

[14] Mehrabian, A. Communication without words (1968). Psychology Today,2(4): 53-56.

[15] Dureha, A. (2014). An Accurate Algorithm for Generating a Music Playlist based on Facial Expressions. International Journal of Computer Applications, 100, 33-39.

[16] Wang, H., & Gu, J. (2018). The Applications of Facial Expression Recognition in Human-computer Interaction. 2018 IEEE International Conference on Advanced Manufacturing (ICAM), 288-291.

[17] Mandal, M. K., Pandey, R., & Prasad, A. B. (1998). Facial expressions of emotions and schizophrenia: A review. Schizophrenia Bulletin, 24(3), 399–412. https://doi.org/10.1093/oxfordjournals.schbul.a033335.

[18] Gehricke, J., & Shapiro, D. (2000). Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. Psychiatry Research, 95, 157-167.

[19] Borod JC, Koff E, Lorch MP, Nicholas M, Welkowitz J.(1988).Emotional and non-emotional facial behaviour in patients with unilateral brain damage. Journal of Neurology, Neurosurgery & Psychiatry 51:826-832.

[20] Owada, K., Kojima, M., Yassin, W., Kuroda, M., Kawakubo, Y., Kuwabara, H., Kano, Y., & Yamasue, H. (2018). Computer-analyzed facial expression as a surrogate marker for autism spectrum social core symptoms. PloS one, 13(1), e0190442. https://doi.org/10.1371/journal.pone.0190442.

[21] Barbosa, J., Lee, K., Lee, S.,Lodhi B.,Cho J-G.,Seo W-K, Kang J.(2016) Efficient quantitative assessment of facial paralysis using iris segmentation and active contour-based key points detection with hybrid classifier. BMC Med Imaging 16, 23. https://doi.org/10.1186/s12880-016-0117-0.

[22] Egger, H.L., Dawson, G., Hashemi, J. et al.(2018) Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study. npj Digital Med 1, 20. https://doi.org/10.1038/s41746-018-0024-6.

[23] Takalkar, M., Xu, M., Wu, Q. & Chaczko Z.(2018). A survey: facial micro-expression recognition. Multimed Tools Appl 77, 19301–19325. https://doi.org/10.1007/s11042-017-5317-2.

[24] de Belen, R.A.J., Bednarz, T., Sowmya, A. et al. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. Transl Psychiatry 10, 333 (2020). https://doi.org/10.1038/s41398-020-01015-w.

[25] Thevenot, J., Lopez, M. B., & Hadid, A. (2018). A Survey on Computer Vision for Assistive Medical Diagnosis From Faces. IEEE journal of biomedical and health informatics, 22(5), 1497–1511. https://doi.org/10.1109/JBHI.2017.2754861.

[26] Bleuler E. (1912).The theory of schizophrenic negativism. New York, NY: The Journal of Nervous and Mental Disease Publishing Company.

[27] Kanner, L. (1943). Autistic disturbances of affective contact. Nervous Child, 2, 217–250.

[28] Park HR, Lee JM, Moon HE, Lee DS, Kim B, Kim J, Kim DG, Paek SH. A Short Review on the Current Understanding of Autism Spectrum Disorders. Exp Neurobiol 2016;25:1-13. https://doi.org/10.5607/en.2016.25.1.1.

[29] Leo M, Carcagnì P, Distante C, Mazzeo PL, Spagnolo P, Levante A, Petrocchi S, Lecciso F.(2019). Computational Analysis of Deep Visual Data for Quantifying Facial Expression Production. Applied Sciences. 9(21):4542. https://doi.org/10.3390/app9214542.

[30] Campbell, K., Carpenter, K. L., Hashemi, J., Espinosa, S., Marsan, S., Borg, J. S., Chang, Z., Qiu, Q., Vermeer, S., Adler, E., Tepper, M., Egger, H. L., Baker, J. P., Sapiro, G., & Dawson, G. (2019). Computer vision analysis captures atypical attention in toddlers with autism. Autism, 23(3), 619–628. https://doi.org/10.1177/1362361318766247.

[31] Coco, M.D., Leo, M., Carcagnì, P., Spagnolo, P., Mazzeo, P.L., Bernava, G.M., Marino, F., Pioggia, G., & Distante, C. (2017). A Computer Vision Based Approach for Understanding Emotional Involvements in Children with Autism Spectrum Disorders. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 1401-1407.

[32] Abirami, S.P., Kousalya, G., Balakrishnan, & Karthick, R. (2019). Varied Expression Analysis of Children With ASD Using Multimodal Deep Learning Technique. Deep Learning and Parallel Computing Environment for Bioengineering Systems.

[33] Zampella, C.J., Bennetto, L., & Herrington, J.D. (2020). Computer Vision Analysis of Reduced Interpersonal Affect Coordination in Youth With Autism Spectrum Disorder. Autism Research, 13.

[34] Guha, T., Yang, Z., Ramakrishna, A., Grossman, R.B., Hedley, D., Lee, S., & Narayanan, S.S. (2015). On quantifying facial expression-related atypicality of children with Autism Spectrum Disorder. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 803-807.

[35] Keating, C.T., & Cook, J.L. (2021). Facial Expression Production and Recognition in Autism Spectrum Disorders: A Shifting Landscape. The Psychiatric clinics of North America, 44 1, 125-139 .

[36] Trevisan, D.A., Hoskyn, M., & Birmingham, E. (2018). Facial Expression Production in Autism: A Meta-Analysis. Autism Research, 11.

[37] Czapinski, P. , & Bryson, S. E. (2003). Reduced facial muscle movements in autism: Evidence for dysfunction in the neuromuscular pathway? Brain and Cognition, 51(2), 177–179.

[38] Leo, M., Carcagnì, P., Mazzeo, P.L., Spagnolo, P., Cazzato, D., & Distante, C. (2020). Analysis of Facial Information for Healthcare Applications: A Survey on Computer Vision-Based Approaches. Inf., 11, 128.

[39] Ekman P,Friesen WV,(1969), Nonverbal leakage and clues to deception.Psychiatry 32: 88–106.

[40] Li, X., Hong, X., Moilanen, A.J., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2018). Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. IEEE Transactions on Affective Computing, 9, 563-577.

[41] Li, G., Shi, J., Peng, J., & Zhao, G. (2019). Micro-expression Recognition Under Low-resolution Cases. VISIGRAPP. doi:10.5220/0007373604270434.

[42] Wood P., Snap: Making the Most of First Impressions, Body Language, and Charisma, Paperback –Oct, 2012, 13 edition, ISBN 978-1577319399.

[43] Ekman P. (2004) Emotional and Conversational Nonverbal Signals. In: Larrazabal J.M., Miranda L.A.P. (eds) Language, Knowledge, and Representation. Philosophical Studies Series, vol 99. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-2783-3_3.

[44] Allaert, B., Bilasco, I.M., & Djeraba, C. (2022). Micro and Macro Facial Expression Recognition Using Advanced Local Motion Patterns. IEEE Transactions on Affective Computing, 13, 147-158.

[45] Liong, S., See, J., Phan, R.C., & Wong, K. (2018). Less is More: Micro-expression Recognition from Video using Apex Frame. Signal Process. Image Commun., 62, 82-92. https://doi.org/10.1016/j.image.2017.11.006.

[46] 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.

[47] Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., Chen, Y. H., & Fu, X. (2014). CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PloS one, 9(1), e86041. https://doi.org/10.1371/journal.pone.0086041.

[48] Ekman P.(2009) Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage, (Revised Edition), WW Norton & Company, ISBN 9780393337457.

[49] Wang, Y., See, J., Oh, Y., Phan, R.C., Rahulamathavan, Y., Ling, H., Tan, S., & Li, X. (2016). Effective recognition of facial micro-expressions with video motion magnification. Multimedia Tools and Applications, 76, 21665-21690., DOI 10.1007/s11042-016-4079-6.

[50] Frank, M.G., Herbasz, M., Sinuk, K., Keller, A., Nolan, C.(2009), I See How You Feel: Training Laypeople and Professionals to Recognize Fleeting Emotions. International Communication Association, Sheraton New York City.

[51] Endres, J., Laidlaw(2009), A. Micro-expression recognition training in medical students: a pilot study. BMC Med Educ 9, 47., vol. 9, no. 1,p. 47, https://doi.org/10.1186/1472-6920-9-47.

[52] Ekman, P. (2002). Microexpression Training Tool (METT). University of California, San Francisco, CA.

[53] Wu, Q. and Shen, X. and Fu, X. (2011) The Machine Knows What You Are Hiding: An Automatic Micro-expression Recognition System, In: D'Mello S., Graesser A., Schuller B., Martin JC. (eds) Affective Computing and Intelligent Interaction. ACII 2011. Lecture Notes in Computer Science, vol 6975. Springer, Berlin, Heidelberg.

[54] Rinn, William E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. Psychological Bulletin, 95(1), 52–77. https://doi.org/10.1037/0033-2909.95.1.52.

[55] Haggard, E.A., Isaacs, K.S. (1966). Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In: Methods of Research in Psychotherapy. The Century Psychology Series. Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-6045-2_14.

[56] Shiffman, M.A. (2012). Muscles Used in Facial Expression. In: Erian, A., Shiffman, M. (eds) Advanced Surgical Facial Rejuvenation. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17838-2_4.

[57] Ekman, P., and Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. California, CA : Consulting Psychologists Press, Palo Alto.

[58] Li, J., Soladie, C. and Seguier, R.(2019), A Survey on Databases for Facial Micro-Expression Analysis. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019), pages 241-248,ISBN: 978-989-758-354-4, DOI: 10.5220/0007309202410248.

[59] Polikovsky, S., Kameda, Y., & Ohta, Y. (2009). Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. ICDP.

[60] Shreve, M., Godavarthy, S., Goldgof, D.B., & Sarkar, S. (2011). Macro- and micro-expression spotting in long videos using spatio-temporal strain. Face and Gesture 2011, 51-56.

[61] Warren, G., Schertler, E. & Bull, P. (2009), Detecting deception from emotional and unemotional cues. J Nonverbal Behav 33, 59–69). https://doi.org/10.1007/s10919-008-0057-7.

[62] Yan, W., Wu, Q., Liu, Y., Wang, S., & Fu, X. (2013). CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1-7.

[63] Qu, F., Wang, S., Yan, W., & Fu, X. (2016). CAS(ME)^2: A Database of Spontaneous Macro-expressions and Micro-expressions. HCI.

[64] Li, X., Pfister, T., Huang, X., Zhao, G., & Pietikäinen, M. (2013). A Spontaneous Micro-expression Database: Inducement, collection and baseline. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),pp.1-6.

[65] Davison, A.K., Lansley, C., Costen, N., Tan, K., & Yap, M. (2018). SAMM: A Spontaneous Micro-Facial Movement Dataset. IEEE Transactions on Affective Computing, 9, 116-129.

[66] Davison, A.K., Merghani, W., & Yap, M. (2018). Objective Classes for Micro-Facial Expression Recognition. J. Imaging, 4, 119.

[67] Husak, P., Cech, J., & Matas, J. (2017). Spotting Facial Micro-Expressions “ In the Wild ”.

[68] Ben, X., Ren, Y., Zhang, J., Wang, S., Kpalma, K., Meng, W., & Liu, Y. (2021). Video-based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms. IEEE transactions on pattern analysis and machine intelligence, PP.

[69] Shen, X. B., Wu, Q., & Fu, X. L. (2012). Effects of the duration of expressions on the recognition of microexpressions. Journal of Zhejiang University. Science. B, 13(3), 221–230. https://doi.org/10.1631/jzus.B1100063.

[70] Pfister, T., Li, X., Zhao, G., & Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. 2011 International Conference on Computer Vision, 1449-1456.

[71] Viola, P.A., & Jones, M.J. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1, I-I.

[72] Cootes, T., Taylor, C.J., Cooper, D.H., & Graham, J. (1995). Active Shape Models-Their Training and Application. Comput. Vis. Image Underst., 61, 38-59.

[73] Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2013). Robust Discriminative Response Map Fitting with Constrained Local Models. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 3444-3451.

[74] Cristinacce, D., & Cootes, T. (2006). Feature Detection and Tracking with Constrained Local Models. BMVC ,Vol. 3, pp.929-938 .

[75] Huang, X., Wang, S., Zhao, G., & Pietikäinen, M. (2015). Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 1-9. doi: 10.1109/ICCVW.2015.10.

[76] Goshtasby, A. (1988). Image registration by local approximation methods. In Image and Vision Computing. Elsevier. Volume 6, Issue 4, Pages 255-261,ISSN 0262-8856.

[77] Zhou, Z., Zhao, G., & Pietikäinen, M. (2011). Towards a practical lipreading system. CVPR 2011, pp. 137-144, doi: 10.1109/CVPR.2011.5995345.

[78] Peng, W., Hong, X., Xu, Y., & Zhao, G. (2019). A Boost in Revealing Subtle Facial Expressions: A Consolidated Eulerian Framework. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 1-5. Lille, France, 2019 pp. 1-5. doi: 10.1109/FG.2019.8756541.

[79] Wu, H., Rubinstein, M., Shih, E., Guttag, J.V., Durand, F., & Freeman, W.T. (2012). Eulerian video magnification for revealing subtle changes in the world. ACM Transactions on Graphics (TOG), 31, 1 - 8.

[80] Ngo, A.C., Johnston, A., Phan, R.C., & See, J. (2018). Micro-Expression Motion Magnification: Global Lagrangian vs. Local Eulerian Approaches. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 650-656. doi: 10.1109/FG.2018.00102.

[81] Takalkar, M.A., & Xu, M. (2017). Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets. 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 1-7. doi: 10.1109/DICTA.2017.8227443.

[82] Liong, S., Gan, Y.S., Zheng, D., Lic, S., Xua, H., Zhang, H., Lyu, R., & Liu, K. (2020). Evaluation of the Spatio-Temporal Features and GAN for Micro-Expression Recognition System. Journal of Signal Processing Systems, 92, 705-725.

[83] Xie, H., Lo, L., Shuai, H., & Cheng, W. (2020). AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. Proceedings of the 28th ACM International Conference on Multimedia. https://doi.org/10.1145/3394171.3414012.

[84] Zhao, G., & Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 29(6), 915–928. https://doi.org/10.1109/TPAMI.2007.1110.

[85] Guo, Y., Xue, C., Yingzi, W., & Yu, M. (2015). Micro-expression recognition based on CBP-TOP feature with ELM. Optik, 126, 4446-4451.

[86] Wang, Y., See, J., Phan, R.C., & Oh, Y. (2014). LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition. ACCV.

[87] Wang, Y., See, J., Phan, R.C., & Oh, Y. (2015). Efficient Spatio-Temporal Local Binary Patterns for Spontaneous Facial Micro-Expression Recognition. PLoS ONE, 10.

[88] Jiang, B., Valstar, M.F., Martínez, B., & Pantic, M. (2014). A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling. IEEE Transactions on Cybernetics, 44, 161-174.doi 10.1109/TCYB.2013.2249063.

[89] Zong, Y., Zheng, W., Hong, X., Tang, C., Cui, Z., & Zhao, G. (2019). Cross-Database Micro-Expression Recognition: A Benchmark. Proceedings of the 2019 on International Conference on Multimedia Retrieval.

[90] Huang, X., Wang, S., Zhao, G., & Pietikäinen, M. (2015). Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 1-9.

[91] Huang, X., Wang, S., Liu, X., Zhao, G., Feng, X., & Pietikäinen, M. (2019). Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection for Spontaneous Facial Micro-Expression Recognition. IEEE Transactions on Affective Computing, 10, 32-47.

[92] Liu, Y., Zhang, J., Yan, W., Wang, S., Zhao, G., & Fu, X. (2016). A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. IEEE Transactions on Affective Computing, 7, 299-310.

[93] Zheng, H. (2017). Micro-Expression Recognition based on 2D Gabor Filter and Sparse Representation.

[94] Lin, C., Long, F., Huang, J., & Li, J. (2018). Micro-Expression Recognition Based on Spatiotemporal Gabor Filters. 2018 Eighth International Conference on Information Science and Technology (ICIST), 487-491.

[95] Patel, D., Hong, X., & Zhao, G. (2016). Selective deep features for micro-expression recognition. 2016 23rd International Conference on Pattern Recognition (ICPR), 2258-2263.

[96] Gan, Y.S., Liong, S., Yau, W., Huang, Y., & Ken, T. (2019). OFF-ApexNet on Micro-expression Recognition System. Signal Process. Image Commun., 74, 129-139.

[97] Shahar, H., & Hel-Or, H. (2019). Micro Expression Classification using Facial Color and Deep Learning Methods. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 1673-1680.

[98] Song, B., Li, K., Zong, Y., Zhu, J., Zheng, W., Shi, J., & Zhao, L. (2019). Recognizing Spontaneous Micro-Expression Using a Three-Stream Convolutional Neural Network. IEEE Access, 7, 184537-184551.

[99] Li, Q., Zhan, S., Xu, L., & Wu, C. (2018). Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. Multimedia Tools and Applications, 1-16.

[100] Yu, J., Zhang, C., Song, Y., & Cai, W. (2020). ICE-GAN: Identity-aware and Capsule-Enhanced GAN for Micro-Expression Recognition and Synthesis. ArXiv, abs/2005.04370.

[101] Quang, N.V., Chun, J., & Tokuyama, T. (2019). CapsuleNet for Micro-Expression Recognition. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 1-7.

[102] Zhou L., Shao X., Mao Q.(2021), A survey of micro-expression recognition, Image and Vision Computing, Volume 105, 2021, 104043, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2020.104043.

[103] Oh, Y., See, J., Ngo, A.C., Phan, R.C., & Baskaran, V.M. (2018). A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges. Frontiers in Psychology, 9.

[104] Guo, Y., Tian, Y., Gao, X., & Zhang, X. (2014). Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. 2014 International Joint Conference on Neural Networks (IJCNN), 3473-3479.

[105] Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., 2, 27:1-27:27.

[106] Dutta, P., & M, N. (2018). Facial Pain Expression Recognition in Real-Time Videos. Journal of Healthcare Engineering, 7961427, https://doi.org/10.1155/2018/7961427.

[107] Chen, Z., Ansari, R., & Wilkie, D.J. (2018). Automated Pain Detection from Facial Expressions using FACS: A Review. ArXiv, abs/1811.07988.

[108] Matsumoto, D., & Hwang, H.C. (2018). Microexpressions Differentiate Truths From Lies About Future Malicious Intent. Frontiers in Psychology, 9.

[109] Mao, L., Wang, N., Wang, L., & Chen, Y. (2019). Classroom Micro-Expression Recognition Algorithms Based on Multi-Feature Fusion. IEEE Access, 7, 64978-64983.

[110] Pei, J., & Shan, P. (2019). A Micro-expression Recognition Algorithm for Students in Classroom Learning Based on Convolutional Neural Network. Traitement du Signal, 36, 557-563.

[111] Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., & Zhang, L. (2016). Image super-resolution: The techniques, applications, and future. Signal Process., 128, 389-408.

[112] Merghani, W., Davison, A.K., & Yap, M. (2018). A Review on Facial Micro-Expressions Analysis: Datasets, Features and Metrics. ArXiv, abs/1805.02397.

[113] Tsai, R.Y., & Huang, T.S. (1984). Multiframe image restoration and registration. Advances in Computer Vision and Image Processing. 317-339.

[114] Dong, C., Loy, C.C., He, K., & Tang, X. (2014). Learning a Deep Convolutional Network for Image Super-Resolution. ECCV.

[115] Dong, C., Loy, C.C., He, K., & Tang, X. (2016). Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 295-307.

[116] Dong, C., Loy, C.C., & Tang, X. (2016). Accelerating the Super-Resolution Convolutional Neural Network. ECCV.

[117] Anwar, S., Khan, S.H., & Barnes, N. (2019). A Deep Journey into Super-resolution: A survey. ArXiv, abs/1904.07523.

[118] Lim, B., Son, S., Kim, H., Nah, S., & Lee, K.M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1132-1140.

[119] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y.R. (2018). Residual Dense Network for Image Super-Resolution. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2472-2481.

[120] Ledig, C., Theis, L., Huszár, F., Caballero, J., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 105-114.

[121] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., & Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. ECCV Workshops.

[122] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., & Bengio, Y. (2014). Generative Adversarial Nets. NIPS.

[123] AlQahtani, H., Kavakli-Thorne, M., & Kumar, G. (2019). Applications of Generative Adversarial Networks (GANs): An Updated Review. Archives of Computational Methods in Engineering, 28, 525-552.

[124] Rakotonirina, N.C., & Rasoanaivo, A. (2020). ESRGAN+ : Further Improving Enhanced Super-Resolution Generative Adversarial Network. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3637-3641.

[125] Russell, W.S. (1995). Polynomial interpolation schemes for internal derivative distributions on structured grids. Applied Numerical Mathematics, 17, 129-171.

[126] Oh, B., Choi, B., & Toh, K. (2009). Fusing horizontal and vertical components of face images for identity verification. 2009 4th IEEE Conference on Industrial Electronics and Applications, 651-655.

[127] Ojansivu, V., & Heikkilä, J. (2008). Blur Insensitive Texture Classification Using Local Phase Quantization. ICISP.

[128] Heikkila, J., & Ojansivu, V. (2009). Methods for local phase quantization in blur-insensitive image analysis. 2009 International Workshop on Local and Non-Local Approximation in Image Processing, 104-111.

[129] Päivärinta, J., Rahtu, E., & Heikkilä, J. (2011). Volume Local Phase Quantization for Blur-Insensitive Dynamic Texture Classification. SCIA.

[130] Wang, Z., & Ying, Z. (2012). Facial Expression Recognition Based on Local Phase Quantization and Sparse Representation. ICNC.

[131] Zhang, B., Liu, G., & Xie, G. (2016). Facial expression recognition using LBP and LPQ based on Gabor wavelet transform. 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 365-369.

[132] Kherchaoui, S., & Houacine, A. (2018). Facial expression identification using gradient local phase. Multimedia Tools and Applications, 78, 16843-16859.

[133] Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 1113-1133.

[134] Li, Y., Huang, X., & Zhao, G. (2021). Micro-expression action unit detection with spatial and channel attention. Neurocomputing, 436, 221-231.

[135] Pawar, S.S., Moh, M., & Moh, T. (2019). Micro-expression Recognition Using Motion Magnification and Spatiotemporal Texture Map. IMCOM.

[136] Zhao, Y., & Xu, J. (2019). A Convolutional Neural Network for Compound Micro-Expression Recognition. Sensors (Basel, Switzerland), 19.

[137] Wei, J., Lu, G., Yan, J., & Liu, H. (2022). Micro-expression recognition using local binary pattern from five intersecting planes. Multimedia Tools and Applications.

[138] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[139] Cardinale, F., Tran, D., “Image Super Resolution”, (2018), https://github.com/idealo/image-super-resolution.

[140] Horé, A., & Ziou, D. (2010). Image Quality Metrics: PSNR vs. SSIM. 2010 20th International Conference on Pattern Recognition, 2366-2369.

[141] Wang, S., Yan, W., Li, X., Zhao, G., & Fu, X. (2014). Micro-expression Recognition Using Dynamic Textures on Tensor Independent Color Space. 2014 22nd International Conference on Pattern Recognition, 4678-4683.

[142] Lu, H., Kpalma, K., & Ronsin, J. (2018). Motion descriptors for micro-expression recognition. Signal Process. Image Commun., 67, 108-117.

[143] Wardhani N. W. S., Rochayani M. Y., Iriany A., Sulistyono A. D. and Lestantyo P., "Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data," 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2019, pp. 14-18, doi: 10.1109/IC3INA48034.2019.8949568.