Facial Expression Recognition: Feature Based Approaches To Deep Learning Techniques

by

SUJATA
201521003

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

December, 2020

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Doctor of Philosophy at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

SUJATA

Certificate

This is to certify that the thesis work entitled FACIAL EXPRESSION RECOGNITION: FEATURE BASED APPROACHES TO DEEP LEARNING TECHNIQUES has been carried out by SUJATA for the degree of Doctor of Philosophy at *Dhirubhai Ambani Institute of Information and Communication Technology* under my supervision.

Suman K. Mitra
Thesis Supervisor

Acknowledgments

First and above all, I praise God, the almighty for providing me this opportunity and granting me the capability to proceed successfully. Although the work described here was performed independently, I never would have been able to complete it if not for the support of many wonderful people. I would like to offer my sincere thanks to them.

I would like begin by expressing my sincere gratitude to my supervisor Prof. Suman K Mitra, with whom I have learned immensely and has had a strong influence in my development as a researcher. He has constantly encouraged me to ensure that I remain focused on achieving my goal. I am grateful to him for patiently supervising and directing my work, fruitful discussions, providing learning opportunities on a number of occasions, and helping me throughout all the different steps of my doctoral research endeavor for the past few years. He not only guiding me in my research but for teaching me many invaluable life lessons.

On a broader note, I wish to acknowledge all the professors of DA-IICT who have inspired me directly or indirectly. I would like to thank faculty members Prof. V. Sunitha, Prof. Manish Narwaria, Prof. Srimanta Mandal for contributing in many ways to my thesis. Thanks for all the insightfull suggestions and interesting observations made during my research progress seminar. A special thanks to Prof. Ranendu Ghosh for refining my work by their valuable suggestions during the synopsis.

I made a lot of new friends at DA-IICT, who helped me in many steps of my study. I thank all of them for everything that they did for me. A very warm thanks to my senior PhD scholars Dr. Purvi Koringa, Dr. Gitam Shikkenawis, Dr. Nupur Jain and Dr. Ashish Phophaliya, for their motivation and support. A

special thanks to fellow researchers Nidhi, Archana, Madhu, Rishi, Pankaj, Purvi, Prashant, Anjali, Miral and Pranav for being amazing buddies. Thanks to DA-IICT staff for all the support.

I am especially grateful to my better half (Chandra Prakash), my daughter (Elakshi Singh), my parents and my in laws for their understanding of my goals, constant love and moral support. I am also thankful to all my dearest friends for their support and encouragement.

Contents

Abstract

Facial expression recognition (FER) is a problem of pattern recognition that invites the attention of computer vision researchers for the last three decades. However, the problem is still alive due to challenges such as – blurring, illumination variation, pose variation, face image captured in the unconstrained environment, and so on. In the beginning, hand-crafted features followed by classical classification mechanism through a classifier have been studied for various features as well as various classifiers. The hand-crafted features that are associated with changes in expression are hard to extract due to the individual distinction and variations in emotional states. With the induction of deep neural network (DNN) and convolution neural network (CNN), a change in the techniques of facial expression recognition is observed both in terms of efficiency and handling various challenges mentioned above. The modular approach presented here mimics the capability of the human to identify a person with a limited facial part. Facial parts like eyes, nose, lips, and forehead contribute more to the expression recognition task. In this thesis, we have addressed classical feature-based approaches to deep learning techniques.

This thesis presents approaches for Facial Expression Recognition (FER). Firstly, we propose two dimensional Taylor expansion for the facial feature extraction as well as to handle the local illumination. Most procedures just used the arrangement with global illumination varieties and thus yielded more unsatisfactory recognition performances within the case of natural illumination variations that are usually uncontrolled within the globe. Hence, to address the brightening variety issue, at that point we presented the (LL) Laplace-Logarithmic area in this article for further improving the exhibition. We applied the proposed 2D

Taylor expansion theorem in the facial feature extraction phase and formulated the 2DTFP method.

In our second FER approach, we propose a histogram of second-order gradients (HSOG) for the feature extraction. Most of the popular local image descriptors in the literature, such as SIFT, HOG, DAISY, LBP and GLOH, only use the first-order gradient information related to slope and elasticity, e.g., length, area, etc. of a surface, and therefore partly characterize the geometric properties of an image. We exploit the local image descriptor that extracts the histogram of second-order gradients (HSOG), which capture the local curvatures of differential geometry, i.e., cliffs, ridges, summits, valleys, basins, etc. That gives us a different shape index. The shape index is computed from the curvatures, and its different values correspond to different shapes. That different shape corresponds to different expressions of the face.

Much work has been done in this field where local texture, features have been extracted and used in the classification. Due to the very local nature of this information, the dimension of the feature vector achieved for the full image is very high, posing computational challenges in real-time expression recognition. In recent times, Dimensionality Reduction methods have been successfully used in image recognition tasks. Here we propose two Dimensionality Reduction methods E-PCA (Euler Principal Component Analysis) and CS-ONPP (Orthogonal Neighborhood Preserving Projection with Class Similarity-based neighborhood). It proved to be gaining huge margin in terms of feature vector length while maintaining the same recognition accuracy.

Classical FER methods do well in certain well-controlled cases. The fundamental issue with hand-crafted features based arrangement approaches is that they require space learning and not generalize well like in the complex dataset. Deep learning is fast becoming a go-to tool for many artificial intelligence problems due to its ability to overcome other approaches and even humans in many problems. DNN has millions of parameters. To get an optimal set of parameters, we need to have a lot of data to train. Even if we have a lot of data, training generally requires multiple iterations, and it takes a toll on the computing resources.

The task of fine-tuning a network is to tweak the parameters of an already-trained network so that it adapts to the new task at hand. Here we propose two deep learning-based methods. The first method is DNNFG (DNN based on Fourier transform followed by Gabor filtering), where we used pre-trained model VGG16 with fine tuning for extracting the facial features. VGG16 is chosen due to the fact of its effective performance in visible detection and speedy convergence. It's concerning 138 million parameters and contains 13 convolutional layers, followed by 3 fully-connected layers (FCs). Since the VGG framework not designed for the FER tasks, so we modified the framework according to our requirements. And the second is 2DNN (Double-channel based Deep Neural Network). Where we utilized VGGFace architecture, VGGFace is trained on 2.6M face images from 2.6k different people. VGGFace architecture is the same as VGG16. Input images are just different in VGGFace other architecture is the same as VGG16. Adapt VGGFace to FER problem, VGGFace is fine-tuned. It easily utilized local and global information about the expressions. DNN based methods improved recognition accuracy compared to classical approaches.

Facial expression recognition (FER) experiments are performed on a number of the benchmark FER databases. Here experiments performed on the four benchmark databases, which are JAFFE, VIDEO, CK+, OULU-CASIA. Basically thesis addresses the classical facial expression recognition approaches and its shortcomings, then moved to deep learning-based approaches to handle these shortcomings. It performed well compared to handcrafted methods. Also, experimentally proved in the thesis that a modular approach is to perform better than holistic approach.

Keywords: LBP, 1D Taylor expansion, 2D Taylor expansion, SVM, K-NN, HOG, HSOG, Facial Expression Recognition, PCA, KPCA, e-PCA, Dimensionality Reduction, ONPP, CNN, DNN, VGG16, VGGFace, TAYLOR SERIES.

List of Tables

List of Figures

CHAPTER 1

Introduction

Meaningful communicating with each other is utmost important for human being. Spoken language is the best means of communication. However there are other means also. One such is facial expression which is a visual communication. So What is Facial Expression? A facial expression produced by the movement of muscles below the face. Usually, human facial expressions are direct and natural means of representing their feelings and intentions. Facial expressions are the salient features of non-verbal communication.

Humans have always had the innate ability to recognize and distinguish the faces of fellow′s faces. With the exception of fingerprints, the facial expression is one of a person's most distinctive and distinguishable visible features. Emotions allow humans to perceive what they feel for something. Recently, there has been a growing interest in improving human-computer interaction (HCI) using natural modalities. Indeed, it is argued [147] that in order to achieve effective intelligent human-computer interaction, computers must be able to interact naturally with the user, similar to the way humans interact with each other. Human beings interact mainly through the speech, but the interaction also takes place through single or both hands gesture and even the gesture of the whole body [157] (to emphasize some parts of the speech or convey the sign language, for example) and through the visualization of emotions [13]. Emotions can be displayed visually or verbally through the speech, and one of the most important ways in which humans show emotions is through facial expressions. For example, humans can clearly identify fear and disgust in the voice, joy, and surprise in facial expressions. Therefore, the recognition of facial expressions, in particular, to convey emotions, has aroused

considerable interest in the scientific community, with conveying applications in HCI, computer animation (virtual characters that transmit emotions), surveillance and security, medical diagnosis, law enforcement, and awareness systems. and therefore, it has been an active research topic in multiple areas such as psychology, cognitive science, human-computer interaction, and the recognition of models in image processing [149].

Common uses of facial expression recognition are: 1. Intelligent entertainment systems for children; 2. Interactive computers; 3. Intelligent sensors; 4. Social robots. In the field of HCI based on language and language technology, recognition of emotional speech is also a challenge. A related interdisciplinary area is affective computing, which studies the development of systems and devices capable of recognizing, processing, and simulating human affections. Affection and emotion are not the same but are deeply connected. Affective qualities include beauty, shape and structure, characteristics that evoke emotions, while emotions include the different feelings that a human being can have, such as anger, fear, disgust, sadness, happiness, and surprise. A system capable of recognizing and understanding emotions, through the recognition of facial expressions, in the Human-Computer Interaction (HCI) process, would facilitate interaction with users: it should be able to "perceive, interpret, express and regulate emotions " [124]. Therefore, recognizing the user's emotional state is "one of the main requirements for computers to successfully interact with humans" [18]. It would also enable "distinguish between user satisfaction and dissatisfaction of the user in a given computer-aided task" [171]. The range of potential applications is wide: "emotional and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments" [171]. For example, an affective system could "calm a crying baby or prevent a strong feeling of loneliness and negative emotion." [181] and an emotion recognition system could be used to help "autistic children learn and elicit emotional responses. And they could also help autistic children learn to distinguish between emotions ".

Facial expressions can be described at different levels, and two of the main

Figure 1.1: Basic emotions as identified by Ekman and Friesen [32]

methods are facial affect (emotion) and facial muscle action (action unit) [171]. The leading study by Ekman and Friesen [32] identified six basic emotions: anger, disgust, fear, happiness, sadness, and surprise, as well as a neutral state, as shown in Fig. 1.1. Their studies suggest that each of these emotions bases corresponds to basic prototypical facial expressions recognized universally.

1.1 Motivation

Although recognizing facial expressions is a relatively easy task for the majority of people, it is a very challenging task for a computer. One reason is that it has been observed that the variations between the images representing the same terms are almost always larger than variations in the facial expressions due to the change in lighting and viewing directions. These fluctuations are compounded by additional factors such as occlusion, gender, and even ethnic origin. Such appearance variations make it difficult to locate facial regions and extract the inherent facial expression features. In unrestricted environments, these variations are even more difficult to model than in well-controlled environments.

The ultimate goal of this work is to study facial expression detection algorithms is restricted and unconstrained environments. Despite the researcher's efforts over a few decades, this problem has remained largely unresolved. To achieve the goal, the objectives of this work can be divided into three parts, which are pursued separately.

The first goal is to investigate novel methods for feature extraction. Because of the large differences between classes and similarities between classes, effective feature extraction is critical for facial expression recognition. The extracted features should represent different types of facial expressions in a manner that is not materially affected by the subject's age, gender, or appearance. It is also desirable

to have features that are robust to localization errors and occlusions.

The second goal is the investigation of feature selection and combination methods for the recognition of facial expressions. It is generally accepted that facial expression recognition performance can benefit from a combination of multiple features. However, there is often no obvious way to select and combine different types of features.

The last objective is to deal with the problem of lacking training data. The majority of the existing facial expression datasets were collected under controlled environments, which can not represent the diverse set of variations found in the real world. Since it is often costly to collect a large amount of training examples. They also suffer from common limitations of being small in terms of both the number of human subjects and images which can lead to over-fitting problems in many learning algorithms and result in poor recognition performance. In this thesis, we increase the number of images of the dataset using the Data Augmentation approach. Then apply the DNN (Deep Neural Network), which can train a system with more than 2 or 3 nonlinear hidden layers. DNN has achieved success in fields such as computer vision, natural language processing, and automatic speech recognition. One of the main strengths of using DNN techniques is that there is no need to feature engineering. Algorithms can learn the features themselves on the basic representations. For example, in image recognition, it is possible to feed a DNN with representations of images in pixels. Then, the algorithm will determine whether a certain combination of pixels represents a particular feature, which is repeated throughout the image. As the data is processed through the levels, the characteristics will change from very abstract forms to meaningful representation of facial expressions.

1.2 Thesis contributions

Having provided a brief introduction to Facial Expression Recognition (FER), we now summarize the important contributions of this thesis, the details of which are discussed in the subsequent chapters. We also highlight how these contributions

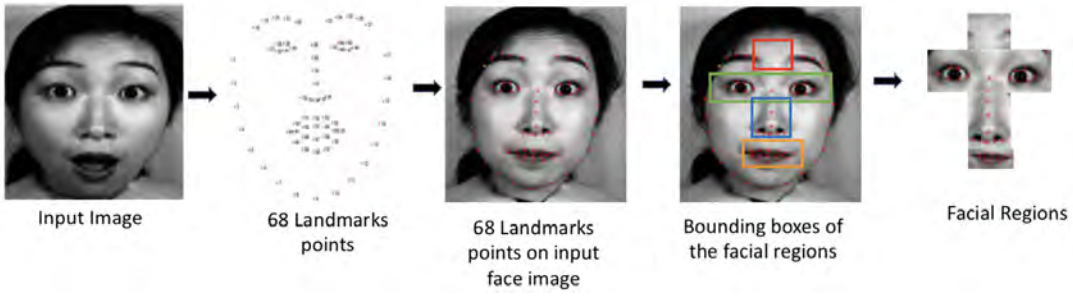

Figure 1.2: Extraction of four facial parts of the given face image

are different from other works. The thesis is to study and implement classical handcrafted approaches and deep learning techniques. Humans have the capability to identify a person with a limited facial part. To extract these facial parts from the face, we have used the Facial Landmark Detection algorithm offered by Dlib which is an open-source machine learning library provided by King [75] The facial landmark detection algorithm offered by Dlib is an implementation of the Ensemble of Regression Trees. It utilizes the technique of pixel intensity difference to directly estimate the landmark positions. The algorithm has a very fast response rate and detects a set of 68 landmarks on a given face. As shown in Fig. 1.2.

The landmarks (key points) of our interest are those that describe the eyes, nose, lips, and eyebrows. Using the landmarks of the eyes, eyebrows, and nose, we find the upper patch between two eyes which we have called the forehead, using the landmarks of the eyes we have extracted the eye region and similarly the nose and lips. These patches are cropped out for each face image and saved. These regions are selected as they give most of the information about the expressions, as proved in [169].

Chapter wise major accomplishments of the thesis are listed down:

1. We developed the facial expression features extraction technique, namely 2DTFP (Two dimensional Taylor Feature Pattern). In this system (FER), the Logarithm-Laplace (LL) is used to handle illumination variation present in the face image. We introduce 2D Taylor Feature Pattern (TFP) in light of the proposed Two Dimensional Taylor Expansion. 2DTFP strategy technique will acquire an efficient facial features feature vector from the given face

images and work efficiently for recognition tasks.

2. We proposed the local image descriptor that extracts the histogram of second-order gradients (HSOG), which capture the local curvatures of differential geometry. The shape index is computed from the curvatures, and its different values correspond to different shapes. An attempt has been made to replicate the same on machines by only considering some of the informative regions of the face like eyes, nose, lip, and forehead. It is observed that the combination of these regions is useful enough to distinguish facial expressions of different persons or the same persons in most of the cases.

3. Due to the very local nature of this information, the dimension of the feature vector achieved for the full image is very high, posing computational challenges in real-time expression recognition. In recent times, Dimensionality Reduction methods have been successfully used in image recognition tasks. Though being high dimensional data, natural images such as face images lie in low dimensional subspace, and Dimensionality Reduction methods try to learn this underlying subspace to reduce the computational complexity involved in the classification stage of image recognition task. We propose E-PCA (Euler Principal Component Analysis) and CS-ONPP (Class Similarity-based Orthogonal Neighborhood Preserving Projection) for expression recognition. Local Features based methods have been successfully applied to Facial Expression Recognition problems; the resulting feature vector lengths usually are of order 10^5, which slows down the classification process.

4. Classical FER methods a do well in certain well-controlled cases. The fundamental issue with hand-crafted features based arrangement approaches is that they require space learning and not generalize well like in the complex dataset. Fortunately, Deep Neural Network (DNN) is giving a satisfactory solution to these issues which were not able to deliver by the hand-crafted techniques. So proposed DNNFG (DNN based on Fourier transform followed by Gabor filtering) where the input images are processed

by VGG16_ft and 2DNN (Two-channel based Deep Neural Network) easily utilized the local and global information about the expressions. Two pre-processing approaches, Histogram equalization (handle the illumination) and Data Augmentation (increase number of facial images), is implemented to restrain the regions used for Facial expression recognition (FER). Double channel architecture used for the implementation, one channel takes input as a grayscale facial image, processed by VGGFace_ft, and other channel takes input as Taylor feature pattern (TFP) facial image, processed by proposed CNN model and extract the features accordingly. DNN based methods improved recognition accuracy compared to classical approaches.

1.3 Thesis Organization

The contents of this thesis are organized as follows. The literature review is presented in .chapter 2. Where we discussed about the existing hand-craft based FER and deep neural network based FER techniques. In chapter 3 , We introduce 2D Taylor Feature Pattern (TFP) in light of the proposed Two Dimensional Taylor Expansion. In this facial expression recognition system (FER), the Logarithm-Laplace (LL) is used to handle illumination variation present in the face image. Ding Yuanyuan [30] deals with one dimensional Taylor series approximation for images however digital images which is considered as a two dimensional signals where each pixel is having dependencies in neighbouring pixels in both horizontal and vertical directions. With this aim in mind we formulated two dimensional Taylor series approximation for digital images. In this formulation we derived how the pixel in concern can be approximated from its neighbours pixel in both horizontal and vertical direction, to the best of our knowledge the Two dimensional Taylor series approximation for digital image is not in the literature. Here, we newly proposed Two dimensional Taylor series approximation for the modular FER as well as holistic FER.

Other hand-craft based feature extraction technique is furnished in chapter 4. We proposed the feature extraction technique which is based on Histogram of

second order gradients of the Image. HSOG is variant of the HOG (Histogram of oriented gradient). HOG [23] is a local image descriptor for feature extraction, which is mainly used for object recognition. This proposed method characterizes the local shape changes of face (that is called the facial expression) by encoding the second oder gradient from the first order oriented gradient. First order gradient delivers the slope where as Second order gradients compute the curvature at the point, that curvature gives shape index and different shape index corresponds to different local shapes. Proposed HSOG is applied on most informative regions of the face i.e. eyes, nose, lip and forehead. It is observed that combination of these regions are useful enough to distinguish facial expressions of different persons or same persons in most of the cases.

Much work has been done in this field where local texture, features have been extracted and used in the classification. Due to the very local nature of this information, the dimension of the feature vector achieved for the full image is very high, posing computational challenges in real-time expression recognition. In last decade it has been observed that dimensionality reduction for face and FER tasks attain maximum attention from the researchers. In the same direction we proposed two dimensionality reduction techniques in chapter 5. First method is based on the Euler Principal Component Analysis (EPCA). of helpful details obtainable within the collected face images. Euler Principal Component Analysis uses a difference live to extend the variations between objects although the face images area unit below the influence of visual variation. Other dimensionality we proposed is CS–ONPP (Class Similarity-based Orthogonal Neighbor- hood Preserving Projection) for expression recognition. Proposed methods are tested on benchmark databases and proved to be gaining huge margin in terms of feature vector length while maintaining same recognition accuracy.

Starting from conventional feature based approach researchers in the current era are mostly relying on deep and convolutional neural network for the mentioned task. In the same direction we proposed a couple of Deep Neural Networks techniques in chapter 6. DNN has millions of parameters. To get an optimal set of parameters, we need to have lot of data to train. Even if we have a lot of data,

training generally requires multiple iterations and it takes a toll on the computing resources. The task of fine-tuning a network is to tweak the parameters of an already trained network so that it adapts to the new task at hand. As explained here, the initial layers learn very general features and as we go higher up the network, the layers tend to learn patterns more specific to the task it is being trained on. Thus, for fine-tuning, we want to keep the initial layers intact (or freeze them) and train the later layers for our task. Original VGG16 has been trained on the IMAGE NET dataset, but for the better classification we need to train the network in our datasets. This has been be efficiently done by the VGG16 with fine-tuning. So Our first method is proposed DNNFG (DNN based on Fourier transform followed by Gabor filtering) which utilized VGG16_ft. And other method 2DNN (Two-channel based Deep Neural Network) easily utilized the local and global information about the expressions. For this task we utilized VGGFace, which is trained on 2.6M face images from 2.6k different people. VGGFace architecture is the same as the VGG16. To adapt VGGFace to FER problem, the VGGFace fine-tuned (denoted as VGGFace_ft) by freezing four blocks of VGGFace and tuning the parameter of the last block. DNN based methods improved recognition accuracy compared to classical approaches.

Finally, in chapter 8 we conclude the thesis by summarizing the main contributions and by listing out future research directions.

CHAPTER 2

Literature Survey

Most traditional methods have used handcrafted features or shallow learning (e.g., Local Binary Patterns (LBP) [150], LBP in three orthogonal planes (LBP-TOP) [212], non-negative matrix factorization (NMF) [215]) and sparse learning [216]) for FER. However, since 2013, emotion recognition competitions such as FER2013 [43] and Emotion Recognition in the Wild (EmotiW) [27] have collected relatively sufficient training data from real world scenarios, which implicitly promote the transition of FER from laboratory controlled environments to in-the-wild settings. Meanwhile, due to the significant increase in chip processing capabilities (for example, GPU units) and a well-designed network architecture, studies

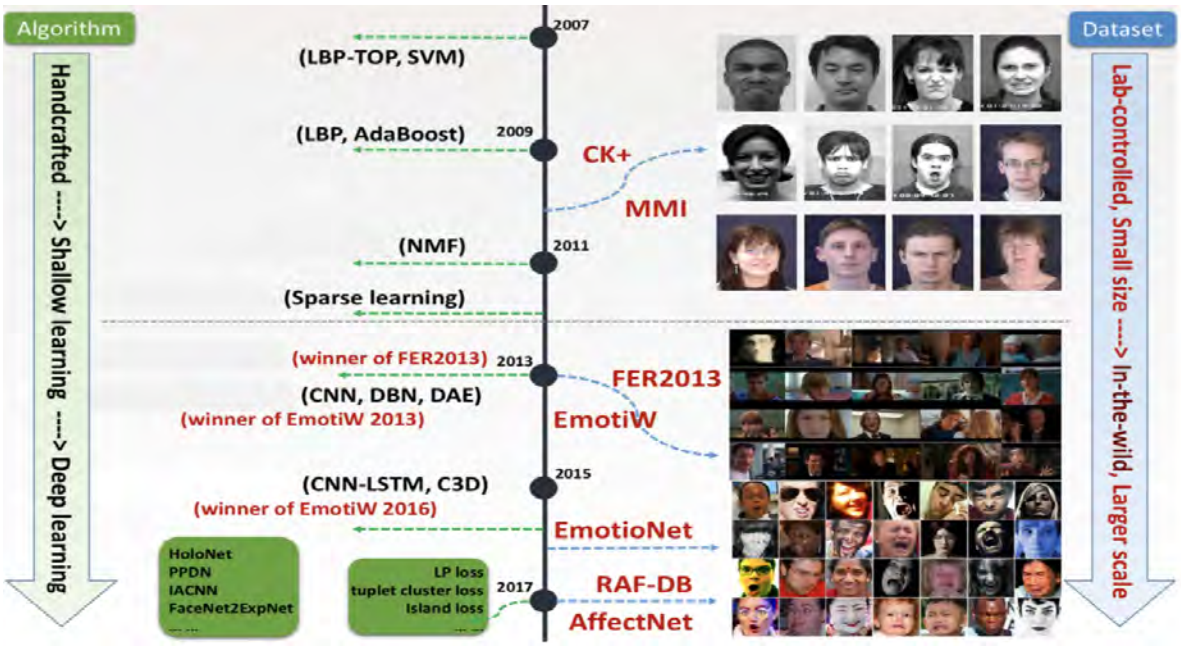

Figure 2.1: The evolution of facial expression recognition in terms of datasets and methods. [88]

in various fields have begun to transfer to deep learning methods, which have achieved the state-of-the-art recognition accuracy and exceeded previous results by a large margin. Similarly, more effective data on facial expression training, deep learning techniques have been increasingly implemented to manage stimulating factors for the recognition of emotions in wild. Fig. 2.1 illustrates this evolution of FER in the aspect of algorithms and data sets.

So In this chapter, we provide a review of the literature for Hand-crafted features extraction and Deep learning based Facial Expression Recognition Techniques, highlighting insights into the current state of research in these areas. We first looked at the literature for Hand-craft features techniques in section 2.1, followed by the Deep Learning based techniques in section 2.2.

2.1 Hand-Crafted Based FER Approaches

Facial expression is one of the most powerful, natural and universal signals for human beings in transmitting their emotional states and their intentions [24], [172]. Numerous studies have been conducted on automatic facial expression analysis due to its practical importance in social robotics, medical treatment, monitoring driver fatigue and many other human-computer interaction systems. In the field of machine vision and machine learning, various facial expression recognition systems (FER) have been explored to encode information about expressions from facial representations. As early as the 20th century, Ekman and Friesen [35] defined six basic emotions based on intercultural study [33], indicating that humans perceive certain basic emotions in the same way, regardless of culture. These prototypical facial expressions are anger, disgust, fear, happiness, sadness and surprise. Contempt was subseqequently added as one of the basic emotions [107]. Recently, advanced research in neuroscience and psychology has argued that the basic six emotion model is culture specific and not universal [56].

Although the basic model of emotion-based affectivity is limited in the ability to represent the complexity and subtlety of our daily affective displays [202], [139], [106] and other models of emotion description, such as facial action coding

system (FACS) [34] and the continuous model that uses the dimensions of affect [45], considered to represent a wider range of emotions, the categorical model that describes emotions in terms of discrete basic emotions remains the most popular for FER, thanks to its pioneering research combined with the direct and intuitive definition of facial expressions. The FER systems can be divided into two main categories based on the feature representations: static image FER and dynamic sequence FER. In static-based methods [150], [95], [112], the feature representation is encoded with only spatial information from the current single image, whereas dynamic based methods [212], [62], [214] consider the temporal relationship between contiguous frames in the facial expression input sequence. Based on these two vision-based methods, other modalities, such as physiological and audio channels, have also been used in multi modal systems [14] to facilitate the recognition of expressions.

The Facial Expression Recognition (FERA) [179] evaluated the detection of Action Unit (AU) and the classification of discrete emotions for four basic emotions and one non-basic emotion. The Audio / Visual Emotion Challenges (AVEC) [145] , [144], [178] evaluated dimensional affect models. FERA demonstrated the substantial progress made in subject-dependent emotion recognition and highlighted open issues in subject-independent emotion recognition; while the challenges of AVEC have highlighted the limitations of existing techniques when it comes to spontaneous affective behavior. Most traditional methods have used handcrafted features or shallow learning (e.g., Local Binary Patterns (LBP) [150], LBP in three orthogonal planes (LBP-TOP) [212], non-negative matrix factorization (NMF) [215] and sparse learning [216]) for Facial Expression Recognition. Low-level histogram representations first extract local features and encode them into a transformed image, then group local features into uniform regions, and finally pool the features of each region with local histograms. The representations are obtained by concatenating all the local histograms. Low-level representations, especially local binary models (LBP) [2] and local phase quantification (LPQ) are very popular. LBP was used by the winner of the word-level challenge AVEC [141] and FERA AU the detection challenge [148], LPQ was used by prominent

systems in FERA [194] and AVEC [19]. The LPQ descriptor was proposed for the classification of blur-insensitive textures through the local Fourier transformation [120]. Similar to an LBP, an LPQ describes a local neighborhood with an integer at [1, 256]. Local histograms simply count LPQ patterns and the dimensionality of each histogram is 256 [120]. The Histogram of Gradients (HoG) approach [23] represents images based on the directions of the edges contained in them. HoG extracts local features by applying gradient operators across the image and encoding their output in terms of gradient magnitude and angle. HoG has been used by a prominent system in the emotional challenge FERA [20].

Another low-level histogram representation is Quantized Zernike Local Moments (QLZM), which describes a neighborhood by calculating its local Zernike moments [140]. Each moment coefficient describes the variation on a unique scale and orientation and the information transmitted by different moment coefficients does not overlap [170]. The QLZM descriptor is obtained by quantifying all the coefficients of the moment in an integer and the local histograms count the integers QLZM.

Low-level representations can be compared from various perspectives. LBP and HoG are compared in terms of sensitivity to registration errors and the results suggest that LBP histograms are generally less sensitive [44]. LBP and LPQ are compared in terms of overall affect recognition performance in various studies and LPQ generally exceeds LBP [59], [60], [178], [194]. This may be due to the size of the local description, as LBPs are usually extracted from smaller regions with 3 pixel diameter [150], whereas LPQs are extracted from larger regions of 7×7 pixels [2], [59], [60]. LBPs cause information to be lost when extracted from larger regions, as they ignore the pixels that remain within the circular region. In contrast, LPQ integers describe the regions as a whole. The QLZM also describe the local regions as a whole and the larger regions as 7×7 were more useful, in particular for recognizing the naturalistic effect [140]. Another comparison that can be useful for low level representations is dimensionality. While the local histograms of LBP and LPQ are relatively higher dimensional (due to their pattern size), QLZM and HoG can be adjusted to obtain smaller histograms that have

been successful respectively on AVEC data [140] and FERA challenge [20].

Another feature-based low-level representation is the Gabor representation, which is used by various systems, including the winner of the FERA AU [90], [186] and AVEC [42] detection challenge. A representation of Gabor is obtained by convolving the input image with a set of Gabor filters of various scales and orientations[81], [184]. Gabor filters encode component information and, depending on the registration scheme, the general representation can implicitly convey configuration information. The high dimensionality of the convolution output makes a phase of reduction of dimensionality essential. As the pixels in the images filtered with Gabor contain information about the neighboring pixels, it is possible to use simple dimensional reduction techniques such as minimum, maximum and average pooling. Gabor filters are differential and localised in space, providing tolerance to illumination variations to a degree [65], [184]. Similar to the representations of low-level histograms, Gabor's representation suffers from identity distortions, in that it favors clues rather than expressions [166]. It is robust for registration errors to some extent, since the filters are uniform and the size of the filtered images is robust for small translations and rotations [44], [81]. The robustness of registration errors can be further enhanced by pooling. The Gabor filter is computationally expensive due to the convolution with a large number of filters [184].

The Bag-of-Words (BoW) representation used in the FER [156] describes local neighborhoods by extracting local features (ie SIFT) densely from fixed positions and then measuring the similarity of each of these featureswith a set of features(for example visual words) in a data set (for example, visual vocabulary) using linear coding with position limitation [156]. The representation inherits the robustness of the SIFT features against illumination variations and minor registration errors. The representation uses spatial pyramid [83] a technique that performs the pooling of histograms and increases the tolerance for registration errors. This matching scheme encodes the information on the components on various scales and the layer that does not divide the image into subregions transmits holistic information. This representation can have a very high dimensionality

and, therefore, its generalization to spontaneous data requires further validation. Although the SIFT descriptors are computationally simple, the visual word count is based on a search for the visual vocabulary and, depending on the size of the vocabulary and the search algorithm used, it can be computationally expensive. Vocabulary training also has a one-off training cost.

So far we describe the local texture. Implicitly or explicitly, its features encode the distribution of the edges. Instead, recent approaches aim to obtain representations based on higher level data to encode features that are semantically interpretable from the point of view of the recognition of expressions. Two methods that generate these representations are NMF [117], [215] and sparse coding [17], [139], [199]. Alternatively, various features learning approaches can also be used [135]. NMF methods decompose a matrix into two non-negative matrices. Decomposition is not unique and can be designed to have multiple semantic interpretations. An NMF-based technique is the NMF conservation graph (GP-NMF) [215], which divides the faces into spatially independent components through a spatial sparseness constraint [52]. The decomposition into independent parts encodes the information on the components and possibly the configural information.

Another NMF-based approach is the subclass discriminant NMF (SD-NMF) [117], which represents an expression with a multimodal projection (rather than assuming that an expression is distributed unimodally). Unlike GP-NMF, SD-NMF does not explicitly impose decomposition into spatially independent components. The basic images provided [101] suggest that the encoded information may be holistic, modular or configurational.

NMF creates several basis images and the features of the NMF-based representations are the coefficients of each basis image. The method performs the minimization to calculate the coefficients, therefore its computational complexity varies according to the optimization algorithm and the number and size of the basis images. Since NMF depends on training, its tolerance against illumination variations and registration errors depends on training data; the NMF's ability to handle both issues NMF concurrently is limited as NMF is a linear technique

[174]. NMF-based representations can address identity bias by learning identity free basis images. This depends on the number of identities provided during the training, as well as on the ability of the technique to deal with interpersonal variations. The dimensionality of NMF-based representations is low: their performances are saturated with less than 100 [215] or 200 features [117].

In addition, another feature extraction method used is the vertical backward time (VTB) which also extracts the texture features from the facial images. The moments descriptor extracts the features relating to the shape of the significant facial components. Both VTB and moments descriptor are effective on the spatiotemporal planes [58]. Weber Local Descriptor (WLD) is a feature extraction technique that extracts highly discriminating texture features from segmented face images [16]. Feature extraction is performed in three stages using the supervised descent method (SDM). Initially, the main facial positions are extracted. Next, the related positions are selected. Finally, estimate the distance between the various components of the face [138]. Weighted projection based LBP (WPLBP) is also a feature extraction, but is based on the instruction regions extracted from LBP features. Subsequently, depending on the significance of the instructive regions, these features are weighted [79]. DCT (Discrete Contour Transformation) extracts the texture features that can be performed by decomposition with two key stages. The stages are Laplacian Pyramid (LP) and Directional Filter Bank (DFB) used in the transformed domain. In the LP stage, partitions the image into low-pass, band-pass and limits the position of the discontinuities. The DFB stage processes the band transition and forms the linear composition that associates the position of the discontinuities [8].

Descriptors that extract features based on edge-based methods are, Line Edge Map (LEM) descriptor that is a facial expression descriptor that improves the geometric structural features using the dynamic two-stripe algorithm (Dyn2S) [41]. Based on the analysis of motion, two types of facial features are extracted, such as non-discriminatory and discriminatory facial features [118]. The Active Shape (GASM) model based on the Graphics processing unit is the feature extraction method that can be performed with edge detection, enhancement, local appear-

ance model matching. After that the image ratio features are extracted from the expressed face images [161].

Descriptors that extract features based on global and local features based methods are, Principal Component Analysis (PCA) method that is used for feature extraction. Extract global and low dimensional features Independent component analysis (ICA) is also a feature extraction method that extracts local features using multichannel observations [155]. Stepwise Linear Discriminant Analysis (SWLDA) is the feature extraction technique that extracts localized features with back and forth regression models. Depending on the class labels, the F test values are estimated for both regression models [154]. Descriptors that extract features based on methods based on geometric features based methods are, Local Curvelet Transformation (LCT) is a feature descriptor which extracts the geometric features that depend on the wrapping mechanism. The extracted geometric features are mean, entropy and standard deviation [176]. In addition to these geometric features are energy, kurtosis are extracted through the use of a three-stage steerable pyramidal representation [105].

Descriptors that extract features based on patch-based methods are, Facial movement features are extracted as patches based on distance characteristics. These are done through the use of two processes, such as patch extraction and patch matching. Patch matching is performed by translating the extracted patches into distance characteristics [205]. The texture Feature descriptors are a more useful feature extraction method than others because it extracts the appearance-related texture features provided by important feature vectors for FER. Also Local Directional Number (LDN) [128], Local Ternary Directional model (LDTP) [137], KL Transform Extended LBP(K-ELBP) [46] and (DWT) [116] texture feature based descriptors are used as feature descriptors in recent years FER. Several extracted features have high dimensional vectors. In general, these feature vectors are reduced using various dimensionality reduction algorithms such as PCA, linear discriminant analysis, Whitened Principle Component Analysis and important features are also selected with different algorithms such as Adaboost and similarity scores.

2.2 Deep Learning Based FER Approaches

Existing FER approaches based on hand-crafted features demonstrate a limited recognition performance. Efforts should be made to manually extract effective features related to expression changes. Many studies have recently studied FER issues based on deep learning consideration of FER's great success in pattern recognition, particularly with the development of the Emotion recognition in Wild Challenge (EmotiW) [26]. A complete review of deep learning is beyond the scope of this study; however, readers can refer to [85], [142]. Here we only deals with some deep networks that can be used to implement FER tasks.

We have divided the works presented in the literature into two main groups based on the type of data: Deep FER networks for static images and Deep FER networks for dynamic image sequences.

2.2.1 Deep FER networks for static images

A large volume of existing studies conducted image-based static expression recognition activities without considering temporal information due to the convenience of data processing and the availability of relevant training and test material. We first introduce specific pre-training and fine-tuning skills for FER, then review the novel deep neural networks in this field.

Pre-training and Fine-Tuning

Direct training of deep networks in relatively small facial expression data sets is prone to overfitting. To mitigate this problem, many studies have used additional activity-oriented data to pre-train their self-built networks from scratch or fine-tunined pre-trained models (e.g., AlexNet [78], VGG [160], VGG-face [123] and GoogleNet [167]). Kahou et al. [64], [67] have indicated that the use of additional data can help to obtain models with high capacity without overfitting, thus improving the performance of FER. To select the appropriate auxiliary data, large-scale face recognition data set (FR) (e.g., CASIA WebFace [196], Celebrity Face in the Wild (CFW) [209], FaceScrub data set [114]) or Relatively large FER datasets

(FER2013 [43] and TFD [165]) are adequate. Kaya et al. [69] suggested that VGG-Face, which was trained for Face Recognition overwhelmed ImageNet, which was developed for objected recognition. Another interesting result observed by Knyazev et al. [76] is that pre-training on larger FR data positively affects the accuracy of emotion recognition and further adjustment with additional FER data sets can help improve performance. Instead of directly using pre-trained or fine-tuned models to extract features in the target data set, a multi-stage fine tuning strategy [113] can achieve better performance: after the first stage fine-tuning with FER2013 in pre-trained models, a second stage fine-tunung based on the training portion of the target data set (EmotiW) is used to refine the models to fit a more specific data set (i.e. the target data set).

Although pre-training and fine-tuning on external Face Recognition data can indirectly avoid the problem of small training data, the networks train separately from the FER and the information dominated by the face remains on the learned features which can weaken the ability of networks to represent expressions . To eliminate this effect, a two-phase training algorithm FaceNet2ExpNet has been proposed [28]. The fine-tuned face net serves as a good initialization for the expressions net and is used to guide learning only of the convolutional layers. And the fully connected layers are trained from scratch with expression information to regularize the training of the target FER net.

Diverse network input

Traditional practices commonly use the entire aligned face of RGB images as network inputs to learn FER features. However, these raw data lack important information, such as homogeneous or regular textures and invariance in terms of image scaling, rotation, occlusion and illumination, which can be confounding factors for FER. Some methods have used various hand-craft features and their extensions as network inputs to alleviate this problem.

The low-level representations encode the features of small regions in the given RGB image, then cluster and pool these features together with the local histograms, which are robust for illumination variations and small registration errors. A novel

mapped LBP feature [86] for FER invariant illumination has been proposed. invariant feature transform (SIFT) [98] features which are robust against image scaling and rotation are employed [208] for multi-view FER activities. The combination of different descriptors in outline, texture, angle and color as input data can also help improve the deep network performance [201], [102].

Part-based representations extract features based on the target activity, which removes the non-critical parts of the entire image and exploit key parts that are sensitive to the activity [11] indicated that three regions of interest (ROI), namely eyebrows, eyes and mouth, are strongly correlated to changes in facial expression and have cut these regions as input for DSAE. Other research has proposed to automatically learn the key parts of facial expression. For example, [108] used a deep multilayer network [15] to detect the saliency map that put intensities on parts that required visual attention. And [185] applied the near center difference vector (NCDV) [100] to obtain characteristics with more intrinsic information.

Auxiliary blocks & layers

Based on the basic architecture of CNN, several studies have proposed the addition of well-designed auxiliary blocks or layers to improve the expression-related representation capability of the learned features. A novel CNN architecture, HoloNet [195], was designed for FER, where CReLU [152] was combined with the powerful residual structure [51] to increase the depth of the network without reducing efficiency and an initial residual block [168], [166] has been uniquely designed for FER to learn multi-scale features to capture variations in expressions. Another CNN model, the Supervised Scoring Ensemble (SSE) [53], was introduced to improve the degree of supervision for FER, in which three types of supervised blocks were embedded in the early hidden layers of mainstream CNN for shallow, intermediate and deep supervision, respectively. And a feature selection network (FSN) [213] was designed by embedding a feature selection mechanism within AlexNet, which automatically filters out irrelevant features and emphasizes related features according to learned feature maps of facial expression. . Interestingly, Zeng et al. [201] noted that inconsistent annotations between different FER

databases are inevitable, which would damage performance when the training set is expanded by merging multiple data sets. To address this problem, the authors proposed a framework of Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) . In IPA2LT, an end-to-end trainable LTNet is designed to discover the latent truths of human annotations and machine annotations trained by different datasets maximizing the log-likelihood of these inconsistent annotations.

The traditional softmax loss layer in CNN simply forces the features of different classes to remain separate, but the FER in real world scenarios suffers from not only a high inter-class similarity, but also high intra-class variation. Therefore, several works have proposed novel loss layers for FER. Inspired by the center loss [183], which penalizes the distance between the deep features and their corresponding class centers, two variants have been proposed to assist the supervision of the softmax loss for the most discriminative features for FER: (1) the loss of the island [9] was formalized to further increase the pairwise distances between the different class centers and (2) locality-preserving loss (loss of LP) [89] was formalized to bring locally neighboring features of the same class together so that the intra-class local clusters of each class are compact. Furthermore, based on the triplet loss [143], which requires a positive example to be closer to the anchor than a negative example with a fixed gap, two variants have been proposed to replace or assist the supervision of the softmax loss: (1) exponential triplet-based loss [47] has been formalized to give difficult samples more weight when updating the network and (2) (N + M)-tuples cluster loss [96] has been formalized to alleviate the difficulty in selecting the anchor and validating the threshold on the loss of a triplet for an invariant of identity- invariant FER. In addition, a feature loss has been proposed [200] to provide complementary information for the deep feature during the initial training phase

Network ensemble

Previous research has suggested that assemblies of multiple networks may outperform a single network [12]. Two key factors should be taken when implementing network ensembles: (1) sufficient diversity of networks to ensure comple-

mentarity and (2) an appropriate ensemble method that can effectively aggregate committee networks.

In terms of the first factor, different types of training data and various network parameters or architectures are considered to generate various committees. Various preprocessing methods [72], such as deformation and normalization, and the methods described can generate different data to train various networks. By changing the size of the filters, the number of neurons and the number of layers of the networks and applying more random seeds for the initialization of the weight, the diversity of the networks can also be improved [73], [126]. In addition, different network architectures can be used to improve diversity.

Multitask networks

Many existing FER networks focus on a single activity and learn features that are sensitive to expressions without considering interactions between other latent factors. However, in the real world, FER is interwined with several factors, such as head posture, illumination and subject identity (facial morphology). To solve this problem, multitasking learning is introduced to transfer knowledge from other relevant tasks and to disentangle nuisance factors. Reed et al. [132] built a higher-order Boltzmann machine (disBM) to learn manifold coordinates for relevant factors of expressions and the proposed training strategies to dis- entangling so that hidden units related to expression are invariant to face morphology. Other works [25], [127] have suggested that FER performed concurrently with other activities, such as identifying facial landmarks and facial AU's [36] detecting, could jointly improve FER performance.

Furthermore, several works [110], [204] used multitasking learning for identity invariant FER. In [110], an identity-aware CNN (IACNN) with two identical sub-CNNs was proposed. One sequence used expression-sensitive contrastive loss to learn the discriminating features of the expression and the other sequence used identity sensitive contrast Loss of to learn identity features for identity invariant FER. In [204], a multi-signal CNN (MSCNN) was proposed, which was trained under the supervision of FER and face verification activities, to force the model

to focus on expression information. In addition, an all-in-one CNN model [131] has been proposed to simultaneously solve a diverse set of face analysis tasks, including smile detection. The network was initially initiated using pre-trained weight on face recognition, so activity-specific subnetworks branched out from several layers with domain-based regularization by training on multiple datasets. In particular, since smile detection is a subject independent activity that relies more on local information available from the lower layers, the authors proposed to combine the lower convolutional layers to form a generic representation for smile detection. Conventional supervised multitasking learning requires labeled training samples for all activities. To relax this, [210] proposed a novel attribute propagation method that can leverage the inherent correspondences between facial expression and other heterogeneous attributes despite disparate distributions of different data sets.

Cascaded networks

In a cascade network, different modules for different tasks are combined sequentially to bulid a deeper network, in which the outputs of the former modules are used by the subsequent modules. Related studies have proposed combinations of different structures to learn a hierarchy of features through which variation factors unrelated to expressions can be gradually filtered out. Very often, different networks or learning methods are combined sequentially and individually and each of them contributes differently and hierarchically. In [103], DBNs were trained to first detect faces and areas related to expression. These analyzed facial components were then classified by a stacked autoencoder. In [133], a multi-scale contractive convolutional network (CCNET) was proposed to obtain Local Translation Invariant (LTI) representations. Then, contractive autoencoder was designed to hierarchically separate the emotional related factors from subject identity and pose. In [91], [92], excessively complete representations were first learned using the CNN architecture, so a multilayer RBM was used to learn higher level features for FER. Rather than simply concatenating different networks, Liu et al. [95] introduced a boosted DBN (BDBN) which performed iteratively feature represen-

tation, feature selection and construction of classifiers in a unified loopy state. Compared to concatenation without feedback, this loopy framework propagates backward the classification error to start the process of selecting the feature alternately until convergence. Therefore, the discriminatory ability of RES can be significantly improved during this iteration.

Generative adversarial networks (GANs)

Recently, GAN-based methods have been used successfully in image synthesis to generate extraordinarily realistic faces, numbers and a variety of other types of images, useful for training data augmentation and corresponding recognition activities. Numerous works have proposed novel GAN based models for pose-invariant RES and identity invariant RES. For post-invariant FER, Lai et al. [82] proposed a face frontalization frame based on GAN, in which the generator frontalises the images of the input face while preserving the identity and expression characteristics and the discriminator distinguishes the real images from the generated frontal face images. And Zhang et al. [203] proposed a GAN-based model capable of generating images with different expressions in arbitrary poses for multi-view FER. For identity invariant FER, Yang et al. [193] proposed a Adaptive-Generation (IA-gen) model with two parts. The upper part generates images of the same subject with different expressions using cGAN, respectively. The lower part then leads FER for each individual identity subspace without involving other individuals, so identity variations can be well alleviated. Chen et al. [10] proposed a Privacy Preserving Learning Variational GAN (PPRL-VGAN) that combines VAE and GAN to learn an identity invariant representation that is explicitly disentangled from the identity information and generative for the expression-preserving face image synthesis. Yang et al. [192] proposed an De-expression Residue learning (DeRL) procedure to explore expressive information, which is filtered out during the de-expression process, but still integrated into the generator. Then the model extracted this information from the generator directly to mitigate the influence of subject variations and improve FER performance.

2.2.2 Deep FER networks for dynamic image sequences

Although most of the previous models focus on static images, the recognition of facial expressions can benefit from the temporal correlations of consecutive frames in a sequence. First, we introduce the existing frames aggregation techniques that strategically combine the deep features learned from static-based FER networks. So, taking into account the fact that in a videostream people usually show the same expression with different intensities, let's review the methods that use images in different expression intensity states for for intensity invariant FER. Finally, we present deep FER networks that consider the spatio-temporal movement patterns in the video frames and the learned features derived from the temporal structure.

Frame aggregation

Since the frames in a given video can vary in expression intensity, direct measurement of the error per frame does not produce satisfactory performance. Various methods have been proposed to aggregate the network output for frames in each stream to improve performance. We divide these methods into two groups: aggregation of frames at the decision level aggregation and feature-level frame aggregation.

For decision-level frame aggregation, the n-class probability vectors of each frame in a sequence are integrated. The most convenient way is to directly concatenate the output of these frames. However, the number of frames in each sequence can be different. Two aggregation approaches were considered to generate a fixed length feature vector for each sequence [64], [63]: average of the frames and expansion of the frames. An alternative approach whose dose does not require a fixed number of frames in the applying statistical coding. The average, max, average of square, the average of maximum suppression vectors, etc. they can be used to summarize the probabilities per frame in each sequence.

For feature-level frame aggregation, the learned features of frames in the sequence are aggregate. Many statistics-based coding modules can be applied in this scheme. A simple and effective way is to concatenate the mean, variance,

minimum and maximum of features over all frames [6]. Alternatively, matrix-based models such as eigenvector, covariance matrix and multidimensional Gaussian distribution can also be used for aggregation [29], [94]. Furthermore, multi-instance learning has been explored video-level rendering [188], in which the centers of the cluster are calculated from auxiliary image data and therefore bag-of-words representation for each bag in video frame is obtained.

Expression Intensity network

Most of the methods focus on recognizing the peak high expression and ignore the subtle expressions of lower intensity. Here, we present expression intensity invariant networks that take training samples with different intensities as inputs to exploit the intrinsic correlations between the expressions of a sequence that varies in intensity. In expression intensity invariant network , image frames with intensity labels are used for training. During the test, data that vary in expression intensity are used to verify the intensity-invariant ability of the network. Zhao et al. [214] proposed a Deep Peak Piloted Network (PPDN) that captures a pair of peak and non-peak images of the same expression and the same subject as input and uses the loss L2 norm to minimize the distance between two images. During back propagation, a peak gradient suppression (PGS) was proposed to drive the learned feature of non- peak expression towards that of peak expression while avoiding the inverse. Thus, the network discriminat ability on lower intensity expressions can be improves. Based on PPDN, Yu et al. [74]proposed a deeper cascaded peak-pilot network (DCPN) which used a deeper and large architecture to improve the discriminative ability of the learned features and used an integration training method called cascade fit-tuning to avoid an excess of adaptation. In [198] more intensity states were used (onset, onset to apex transition, apex, apex to offset transition and offset) and five loss functions were adopted to regulate the training of the network by minimizing the error of expression classification, variation of intra-class expression, intensity classification error and intra-intensity variation and encoding of intermediate intensity, respectively.

Deep spatio-temporal FER network

Although frame aggregation can integrate frames into the video stream, crucial time dependency is not explicitly exploited. By contrast, the spatio-temporal FER network takes a series of frames in a temporal window as a single input without a prior knowledge of the expression intensity and uses both textural and temporal information to encode more subtle expressions.

RNN and C3D: RNN can robustly derive information from sequences by taking advantage of the fact that feature vectors for subsequent data are semantically connected and are therefore interdependent. The improved version, LSTM, is flexible to manage variable length sequential data with lower computation costs. Derived from RNN, an RNN composed of ReLUs and initialized with Identity Matrix (IRNN) [84] has been used to provide a simpler mechanism for dealing with vanishing and exploding gradient problems [64]. And bidirectional RNNs (BRNNs) [146] have been used to learn temporal relationships in both the original and inverse directions [204], [191]. Recently, a Nested LSTM was proposed in [197] with two sub-LSTM. That is, T-LSTM models the temporal dynamics of the learned features and C-LSTM integrates the outputs of all T-LSTMs to encode the multilevel features encoded in the intermediate layers of the network.

Compared to RNN, CNN is more suitable for computer vision applications; hence, its derivative C3D [173], which uses 3D convolutional kernels with shared weights along the time axis instead of the traditional 2D kernels, has been widely used for dynamic FER(for example [1], [37], [122]) to capture the spatiotemporal features. Based on C3D, many derived structures have been designed for FER. In [93], 3D CNN was incorporated with the DPM-inspired [38] deformable facial action constraints to simultaneously encode dynamic movements and discriminations part-based representations. In [62], a deep temporal appearance network (DTAN) was proposed that used 3D filters without sharing the weight along the time axis; hence, each filter can vary in importance over time. Similarly, a weighted C3D [180] has been proposed, in which several consecutive frame windows of each sequence have been extracted and weighted based on their prediction scores. Instead of using C3D directly for classification, [115] employed C3D

for spatiotemporal feature extraction and then cascaded with DBN for prediction. In [125], C3D was also used as a feature extractor, followed by a NetVLAD layer [4] to aggregate the temporal information of the movement features by the centers of the learning cluster.

Facial Landmark Trajectory:

Related psychological studies have shown that expressions are invoked by dynamic motions of some parts of the face (eg. Eyes, nose and mouth) which contain the most descriptive information to represent the expressions. To obtain more accurate facial actions for FER, facial landmark trajectory models of the facial reference point have been proposed to capture dynamic variations of the facial components from consecutive frames. To extract the representation of the landmark trajectory, the most direct way is to concatenate facial landmark points from frames over time with normalization to generate a one-dimensional trajectory signal for each sequence [62] or to form an image-like map as the input of CNN [191]. Besides, relative distance variation of each landmark in consecutive frames can also be used to capture temporal information [74]. In addition, the part-based model that divides the facial landmarks into various parts based on the physical structure of the face and then feeds them separately into networks hierarchical has proven effective in both local low-level and global high level feature encoding [204]. Instead of extracting the trajectory features separately and then insert them into the networks, Hasani et al. [49] incorporated the trajectory features by replacing the shortcut in the residual unit of the original 3D Inception-ResNet with element-wise multiplication of the facial landmarks and the input tensor of the residual unit. Therefore, the landmark based network can be end-to-end trained.

Cascaded networks:

By combining the powerful perceptual vision representations learned by CNN with the strength of the LSTM for variable length inputs and outputs, Donahue et al. [31] have proposed a both spatially and temporally deep model that cascades CNN outputs with LSTM for various vision tasks involving time-varying inputs and outputs. Similar to this hybrid network, many cascade networks have been proposed for FER (eg [74], [37]).

Instead of CNN, [5] used a convolutional sparse autoencoder for sparse and shift invariant features; then, an LSTM classifier was trained for temporal evolution. [122] used a more flexible network called ResNet-LSTM, which allows nodes in the lower layers of CNN to contact LSTM directly to acquire spatio-temporal information. In addition to concatenating LSTM with the fully connected layer of CNN, a hypercolumn-based system [68] has extracted the latest convolutional layer features such as LSTM input for long-range dependencies without losing overall consistency. Instead of LSTM, the conditional random fields (CRFs) model citelafferty2001conditional which is effective in recognizing human activities in [50] has been used to distinguish temporal relationships from input sequences.

Network ensemble:

A two-stream CNN for action recognition in videos, which trained one stream of the CNN on the multi-frame optical flow for temporal information and the other CNN stream on still images for appearance features and then combined outputs of two streams, was presented by Simonyan et al. [159] Inspired by this architecture, various network ensemble models have been proposed for FER. Sun et al. [164] proposed a multichannel network that extracted spatial information from faces expressing emotions and temporal information (optical flow) from changes between emotional and neutral faces and investigate three feature fusion strategies: score average fusion , SVM-based fusion and neural network-based fusion.Zhang et al. [204] combined the temporal network PHRNN and the spatial network MSCNN to extract partial-complete, geometry-appearance and static-dynamic information for FER. Instead of combining network outputs with different weights, Jung et al. [62] proposed a joint fine-tuning method that co-trained the DTAN, the DTGN and the integrated network, which outperformed the weighted sum strategy.

Having discussed the current research status in Facial Expression Recognition, we present our first FER approach in the following chapter 3.

CHAPTER 3

2DTFP: Two Dimensional Taylor Feature Pattern

In any successful facial expression system, the most critical aspect is to locate the proficient features of the given face image or image sequences. The extracted facial feature may be regarded as an efficient representation, Which aims to maximize changes between class and reduce the within-class variations of expressions. In literature, LBP is a good feature method that is widely accepted collectively of the most straightforward options to capture the natural form and edge data. The brief description of the LBP is given in the next subsection.

3.1 Local Binary Pattern (LBP)

At first, LBP was planned to be applied for texture evaluation [119]. Later it was applied in many other fields. Overview of LBP and its coding process is given in Fig .3.1. In each pixel of a given image/image sequences, LBP allots a label within N-neighborhood with the aid of thresholding its value with the center pixel value (p_c), at that point changing over these thresholded values into the decimal number by Eq. 3.1. N is similarly separated pixel esteem inside the range R and meant as (p_n).

$$LBP_{N,R}(X_c, Y_c) = \sum_{n=0}^{N-1} S(p_n - p_c)2^n \tag{3.1}$$

$$S(p_c, p_n) = \begin{cases} 1; & \text{if } p_c \geq p_n \\ 0; & \text{if } p_c < p_n \end{cases}$$

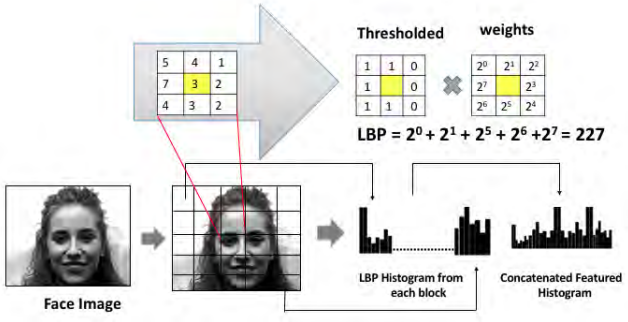

Figure 3.1: Overview of Local Binary Pattern (LBP)

LBP has the capability of strong texture discrimination so it can quickly obtain impressive accuracy in the fields of pattern recognition. However, real time applications of LBP have some limitations, so if we are talking about facial expression recognition, LBP is sensitive to local illumination variations. It also cannot correctly identify the texture of facial muscles and other types of local deformations. That is why we are moving towards the other patterns for recognizing facial expression.

3.2 Taylor Series Theorem

A function f on [c,d], is n times differentiable. Let ϕ and ψ be distinct points on [c,d] ans we define,

$$R(t) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\phi)}{k!}(t-\phi)^k \tag{3.2}$$

Then some point **x** exists in between ϕ and ψ such that

$$f(\psi) = R(\psi) + \frac{f^{(n)}(x)}{n!}(\psi-\phi)^n \tag{3.3}$$

Taylor's theorem is regularly truncated through a limited maximum point of confinement for the summation that is based on the specific scenario in which the Taylor expansion is being implemented. Taylor's expansion is utilized pervasively in all fields to help take care of issues in a compliant manner. Taylor's

hypothesis demonstrates that the function f can be an estimation of degree n polynomial.

In writing Taylor's theorem has been viewed as the proficient apparatus for contemplating the SAR images [134]. In 2017 Ding Yuanyuan [30] came with the theory of one dimensional Taylor expansion for FER (facial expression recognition) related tasks. The detail description about the one dimensional Taylor expansion is given in the next section.

3.3 One Dimensional Taylor Expansion

Ding Yuanyuan [30] used one dimensional Taylor expansion for feature extrication for the FER tasks. In the feature extrication stage, per pixel feature is expanded into various term accordance to Taylor's theorem to generate a strong description of a given image. It used some expanded term to portray the feature of an individual pixel, named as Pixel Taylor feature.

3.3.1 Pixel Taylor Feature

Let $f_n(p_c)$ be the $n^{th}-$ order Taylor pixel feature whose focal pixel is p_c. According to Taylor's theorem as given in *eq*. (3.2) and (3.3), $f_n(p_c)$ is outlined as:

$$f_n(p_c) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\phi)(p_c - \phi)^k}{k!} + \frac{f^{(n)}(\phi)(p_c - \phi)}{n!} \tag{3.4}$$

Fig 2 shows the TU (Texture unit) with 3×3 of 1^{st}order Taylor pixel feature $f_1(p_c)$ and 5×5 of the 2^{nd} order Taylor pixel feature $f_2(p_c)$. Grey implies the first layer of texture unit, and dark grey indicates the second layer of the texture unit.

According to Eq. 3.4 1^{st} order taylor pixel feature $f_1(p_c)$ is expressed as:

$$f_1(p_c) \approx \frac{f^{(0)}(\phi)(p_c - \phi)^0}{0!} + \frac{f^{(1)}(\phi)(p_c - \phi)^1}{1!} \tag{3.5}$$

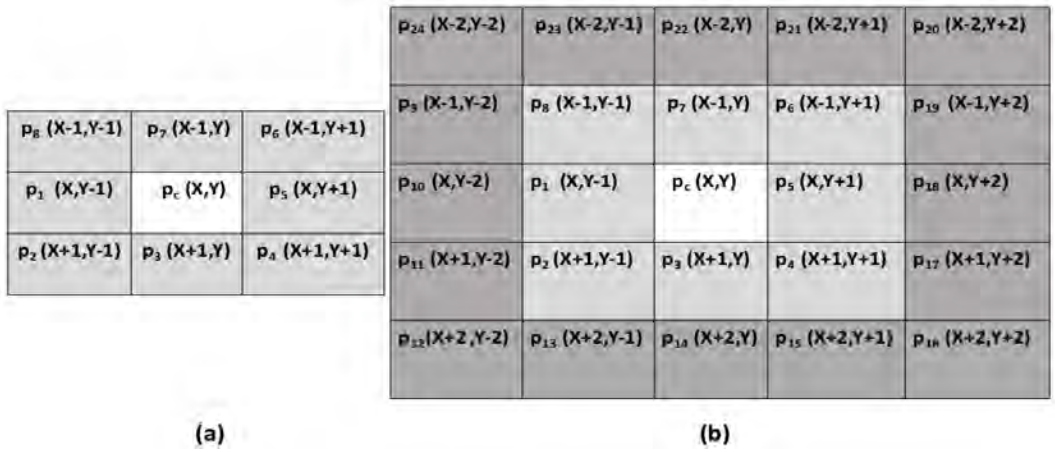

Figure 3.2: (a) 1^{st} order Pixel taylor feature $f_1(p_c)$ (Texture1 (T1) with 3×3 pixels) (b) 2^{nd} order Pixel taylor feature $f_2(p_c)$ (Texture2 (T2) with 5×5 pixels)

where

$$f^{(1)}(\phi) = \begin{cases} 1; & \text{if } p_c - \phi \geq 0 \\ -1; & \text{if } p_c - \phi < 0 \end{cases}$$

Here, $f^{(0)}(\phi)$ is communicated as the average of all the pixels value 1^{st} order Texture1(T1), while ϕ is the average of all the pixels value which comes under the 1^{st} layer (noted as grey) of T1 (as shown in Fig .3.2(a).

2^{nd}order taylor pixel feature $f_2(p_c)$ can be expressed as:

$$\begin{aligned} f_2(p_c) \approx & \frac{f^{(0)}(\phi)(p_c - \phi)^0}{0!} + \frac{f^{(1)}(\phi)(p_c - \phi)^1}{1!} \\ & + \frac{f^{(2)}(\phi)(p_c - \phi)^2}{2!} \end{aligned} \tag{3.6}$$

where

$$f^{(1)}(\phi) = \begin{cases} 1; & \text{if } p_c - \phi \geq 0 \\ -1; & \text{if } p_c - \phi < 0 \end{cases}$$

$$f^{(2)}(\phi) = \begin{cases} 1; & \text{if } (p_c - \phi_1)(p_c - \phi_2) \geq 0 \\ -1; & \text{if } (p_c - \phi_1)(p_c - \phi_2) < 0 \end{cases}$$

Here, $f^{(0)}(\phi)$ is additionally characterized as the average of all the pixels value 2^{nd} order Texture2(T2). while ϕ_1 is the average of all the pixels value which comes under the 1^{st} layer (noted as grey) of T2, ϕ_2 is the average of all the pixels value of the 2^{nd} layer in the T2 (noted as dark grey) and ϕ is the average of all the pixels

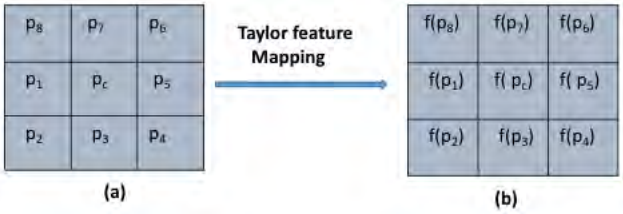

Figure 3.3: (a) Texture with 3×3 pixels in an image I (b) Same Texture in TFM (Taylor feature map)

value that comes in the 1^{st} layer and 2^{nd} layer of T2 (as shown in Fig .3.2(b)).

Similarly, we can get Pixel Taylor feature of higher-order $f_n(p_c)$, $n > 2$. After every pixel of given face images is mapped to Pixel Taylor feature space, we get the TFM (Taylor Feature Map) for further analysis.

3.3.2 Taylor Feature pattern (TFP)

In order to contemplate any real application, we need to reduce the dimension of the feature. Hence, we compute Taylor Feature Pattern (TFP) of Taylor Feature Map (TFM). As shown in Fig. 3.3(a) that is Texture with 3×3 in the given image and Fig. 3.3(b) texture in the TFP (Taylor Feature map). The TFP of $f_n(p_c)$ is expressed as:

$$TFP = \sum_{j=1}^{8} S(f_n(p_c), f_n(p_j).2^{j-1} \tag{3.7}$$

$S(f_n(p_c), f_n(p_j))$ computed the same as defined in LBP.

Finally, a feature vector of the given facial image using Taylor Feature Pattern (TFP) histogram of the TFM (Taylor feature map) was constructed. 2^{nd} order Taylor Feature Pattern with one dimensional provided a satisfactory result. But we tried to find a more accurate result compared to LBP and 2^{nd} order one dimensional Taylor expansion. So we moved forward towards the 2^{nd} order two dimensional Taylor theorem expansion. In the subsequent section, we discuss the proposed two dimensional Taylor expansion.

3.4 Proposed Two Dimensional Taylor Expansion

In this segment, we essentially depict the premise of our technique and propose the expansions to the Taylor series approach, which improves the efficiency and accuracy of the facial expression recognition. This paper is influenced by the LBP and Taylor expansion [30] and utilizes some theory of these two image descriptor in its way, which is beneficial for the FER (Facial Expression Recognition) application. Based on the above analysis of LBP and Taylor expansion we describe the Two Dimensional Taylor expansion for the facial expression recognition.

Proposed 2D Taylor expansion can extract the advantageous features from the given face image and be utilized for the facial expression recognition task. But it may be possible that the recognition rates of these techniques vary a lot with changing illuminations. To handle the illuminations variations, Logarithm-Laplace (LL) strategy was proposed. In Logarithm-Laplace (LL), initially, the input face image is transferred to LL space. That Logarithm-Laplace (LL) space is the invariable illumination space. So in the first step, the raw face image/ image sequences are transferred to the LL domain. After that process, the 2D Taylor expansion of the converting LL domain image into R blocks to induce the natural features of the face image. R is associated with the recognition rate and the time of the recognition step. Here, R is set to 6×6, 8×8, 12, 16×16 respectively. Finally, the sub Taylor feature pattern (TFP) histogram of each block in the resulting image are computed. Then the long histogram was made by concatenating all the sub histograms. That long histogram acts as the Taylor Feature Pattern (TFP) feature vector of the input face image. The procedure works in three overlays for computing the descriptor. The overview of the process is given in Fig 3.4. The means are talked about in the accompanying subsections.

Step 1: Converting in Logarithm-Laplace (LL) Domain The proposed 2D Taylor expansion can get excellent outcomes for FER applications compared with different state-of-art hand-craft based feature extrication methods. In some uncon-

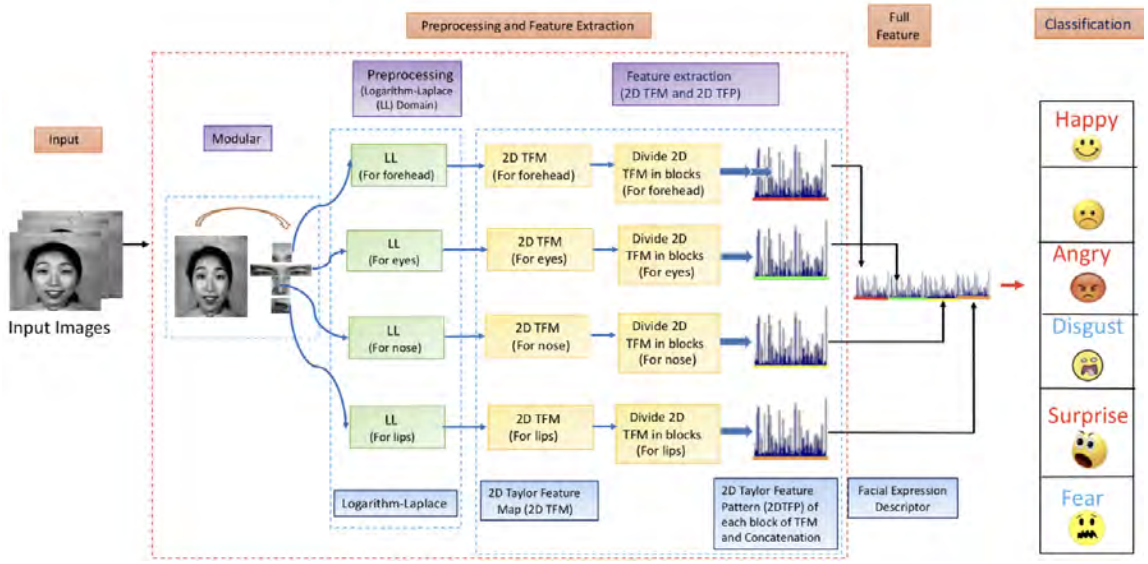

Figure 3.4: Overview of the proposed 2D Taylor Expansion for extracting the facial features

trolled environment like varying illuminations, postures, and noise, FER frameworks performance is drastically influenced. Considering the influence of the illumination variations and different challenges within the real applications, Logarithm-Laplace (LL) is further proposed to assist in inducing an additional strong facial expression related feature. As we know LL space is an invariable illumination space, so the details description of this LL space is given below:

Input image I at point (x_1, y_1) in reflectance model can be defined as:

$$I(x_1, y_1) = H(x_1, y_1).K(x_1, y_1) \tag{3.8}$$

Here, $H(x_1, y_1)$ is reflectance segment which is managed by the qualities of the object. $K(x_1, y_1)$ demonstrates the illumination segment; it depends upon the lighting (brightening) source. As a rule, there is a typical suspicion [136, 207] that differs gradually while H fluctuates unexpectedly. Firstly, this model transformed to logarithmic space to get the illumination invariant features. The logarithmic domain is computed as:

$$logI(x_1, y_1) = logH(x_1, y_1) + logK(x_1, y_1) \tag{3.9}$$

Eq. (3.9) can be also written as:

$$i(x_1, y) = h(x, y) + k(x, y) \tag{3.10}$$

Here $i(x_1, y_1) = logI(x_1, y_1), h(x_1, y_1) = logH(x_1, y_1)$ and $k(x_1, y_1) = logK(x_1, y_1)$

We know that the Laplace space, takes advantage of the connections among component points in a neighborhood may get new facial expression than the pixel domain. Thus, $i(x, y)$ is then remodeled into the Laplace domain:

$$\bigtriangledown^2 i(x_1, y_1) = \bigtriangledown_x^2 i(x_1, y_1) + \bigtriangledown_y^2 i(x_1, y_1) \tag{3.11}$$

Here $\bigtriangledown_x^2 i(x_1, y_1)$ and $\bigtriangledown_y^2 i(x_1, y_1)$ indicate Laplacian value of $i(x_1, y_1)$ in each x axis and y axis. By using the Laplace transformation Eq. 3.11 can be communicated as:

$$\bigtriangledown^2 i(x_1, y_1) \;\; = \;\; i(x_1 + 1, y_1) + i(x_1 - 1, y_1) + i(x_1, y_1 + 1) + i(x_1, y_1 - 1) - 4.i(x_1, y_1)$$

So Eq. 3.10 substitute the into the Eq. 3.11. $k(x_1 + 1, y_1), k(x_1 - 1, y_1), k(x_1, y_1 + 1)$ and $k(x_1, y_1)$ are practically equivalent since K differs gradually. Along these lines, Eq. 3.12 is adjusted as:

$$\bigtriangledown^2 i(x_1, y_1) \;\; \approx \;\; j(x_1 + 1, y_1) + j(x_1 - 1, y_1) + j(x_1, y_1 + 1) + j(x_1, y_1 - 1) - 4.j(x_1, y_1)$$

So we can say that $\bigtriangledown^2 i(x_1, y_1)$ is only dependent on the reflectance part in Eq. 3.8, it is viewed as illumination-invariable space of given face image and space is called a "LL domain". After transferring the raw image/image sequences into the LL domain, the Taylor 2D expansion is applied on it to get the powerful features from it. Subsequent stage depicts the Taylor 2D extension in detail.

Step 2: Two Dimensional (2D) Taylor Expansion Of Image In this section, we suggest the FER (facial expression recognition) technique. It is much like the 2^{nd} order one dimensional Taylor expansion. As shown in Fig 3.5 T1 (Texture 1) with 3×3 of 1^{st}order Pixel taylor feature $f_1(p_c)$ and T2 (Texture 2) with 5×5 pixels of the 2D 2^{nd} order Texture . So 2D 2^{nd} order Texture feature extraction is

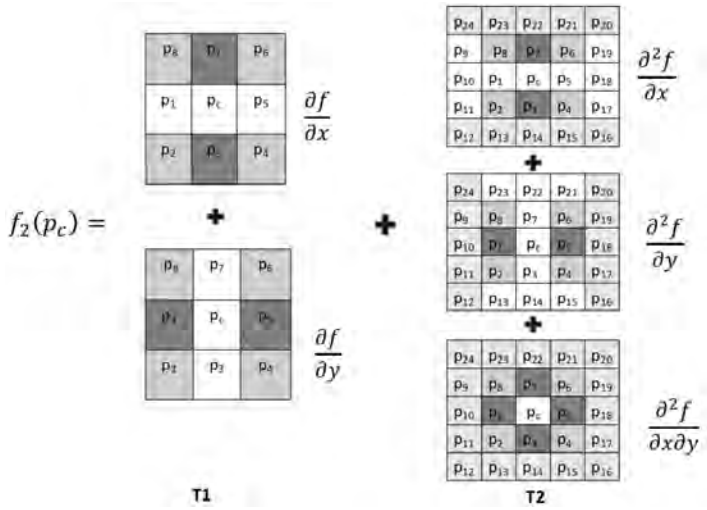

Figure 3.5: Illustration of the proposed 2D 2^{nd} order Taylor feature extraction. T1(Texture 1) with 3×3 pixels of the 2D 1^{st} order Pixel taylor feature $f_1(p_c)$, T2 (Texture 2) with 5×5 pixels of the 2D 2^{nd} order Texture. So 2D 2^{nd} order Texture feature extraction is $f_2(p_c) = T1 + T2$

$f_2(p_c) = T1 + T2$. Detail about the Computation of 2D Pixel taylor feature is given below.

2D Pixel taylor feature extraction Suppose 1^{st} order 2D Pixel taylor feature of the focal pixel p_c is $f_1(p_c)$. As indicated by one dimensional Taylor theorem expansion Eqs. 3.4 and 3.5, Two dimensional may be or so outlined as:

$$f_1(p_c) \approx f(\phi, \psi) + [(p_c - \phi)\frac{\partial f}{\partial x} + (p_c - \psi)\frac{\partial f}{\partial y}] \tag{3.12}$$

where

$$\phi = \frac{\frac{1}{\sqrt{2}}p_8 + p_7 + \frac{1}{\sqrt{2}}p_6 + \frac{1}{\sqrt{2}}p_4 + p_3 + \frac{1}{\sqrt{2}}p_2}{\frac{4}{\sqrt{2}} + 2}$$

This p_8, p_7, p_6, p_4, p_3, p_2 are the pixels of T1 which are in x direction. As shown in Fig 3.5.

$$\psi = \frac{\frac{1}{\sqrt{2}}p_8 + p_1 + \frac{1}{\sqrt{2}}p_2 + \frac{1}{\sqrt{2}}p_4 + p_5 + \frac{1}{\sqrt{2}}p_6}{\frac{4}{\sqrt{2}} + 2}$$

This p_8, p_1, p_2, p_4, p_5, p_6 are the pixels of T1 which are in y direction. As appeared in Fig 3.5.

$$f(\phi, \psi) = \frac{\phi + \psi}{2}$$

$$\frac{\partial f}{\partial x} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \phi \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \phi < 0 \end{cases} \qquad \frac{\partial f}{\partial y} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \psi \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \psi < 0 \end{cases}$$

2^{nd}order pixel taylor feature $f_2(p_c)$ can be expressed as:

$$\begin{aligned} f_2(p_c) \approx\ & f(\phi, \psi) + [(p_c - \phi)\frac{\partial f}{\partial x} + (p_c - \psi)\frac{\partial f}{\partial y}] + \frac{1}{2}[(p_c - \phi)^2\frac{\partial^2 f}{\partial x^2} \\ & + 2(p_c - \phi)(p_c - \psi)\frac{\partial^2 f}{\partial x \partial y} + (p_c - \psi)^2\frac{\partial^2 f}{\partial y^2}] \end{aligned} \tag{3.13}$$

In the computation of ϕ and ψ there are some term like ϕ_1, ψ_1, ϕ_2, ψ_2 used. These terms defined as

$$\phi_1 = \frac{\frac{1}{\sqrt{2}}p_8 + p_7 + \frac{1}{\sqrt{2}}p_6 + \frac{1}{\sqrt{2}}p_4 + p_3 + \frac{1}{\sqrt{2}}p_2}{\frac{4}{\sqrt{2}} + 2}.$$

These p_8, p_7, p_6, p_4, p_3, p_2 are the pixels of T2 which are in x direction. As shown in Fig 3.5.

$$\psi_1 = \frac{\frac{1}{\sqrt{2}}p_8 + p_1 + \frac{1}{\sqrt{2}}p_2 + \frac{1}{\sqrt{2}}p_4 + p_5 + \frac{1}{\sqrt{2}}p_6}{\frac{4}{\sqrt{2}} + 2}$$

These p_8, p_1, p_2, p_4, p_5, p_6 are the pixels of T2 which are in y direction. As appeared in Fig .3.5.

$$\phi_2 = \frac{1}{10}[p_{24} + p_{23} + p_{22} + p_{21} + p_{20} + p_{16} + p_{15} + p_{14} + p_{13} + p_{12}]$$

Here, ϕ_2 is the average of all the pixels value of the 2^{nd} layer in the T2 which are in x direction. Pixels $p_{24}, p_{23}, p_{22}, p_{21}, p_{20}, p_{16}, p_{15}, p_{14}, p_{13}, p_{12}$ will came here.

$$\psi_2 = \frac{1}{10}[p_{24} + p_9 + p_{10} + p_{11} + p_{12} + p_{16} + p_{17} + p_{18} + p_{19} + p_{20}]$$

Here, ψ_2 is the average of all the pixels value of the 2^{nd} layer in the T2 which are in y direction. Pixels $p_{24}, p_9, p_{10}, p_{11}, p_{12}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$ will came here. Finally compute the ϕ and ψ like

$$\phi = \frac{\phi_1 + \phi_2}{2} \quad and \quad \psi = \frac{\psi_1 + \psi_2}{2}$$

Then, compute

$$f(\phi, \psi) = \frac{\phi + \psi}{2}$$

Here, evaluate the derivative of x and y like

$$\frac{\partial f}{\partial x} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \phi_1 \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \phi_1 < 0 \end{cases} \qquad \frac{\partial f}{\partial y} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \psi_1 \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \psi_1 < 0 \end{cases}$$

$$\frac{\partial^2 f}{\partial x^2} = \begin{cases} \frac{1}{4}; & \text{if } (p_c - \phi_1)\,(p_c - \phi_2) \geq 0 \\ -\frac{1}{4}; & \text{if } (p_c - \phi_1)\,(p_c - \phi_2) < 0 \end{cases} \quad \frac{\partial^2 f}{\partial y^2} = \begin{cases} \frac{1}{4}; & \text{if } (p_c - \psi_1)\,(p_c - \psi_2) \geq 0 \\ -\frac{1}{4}; & \text{if } (p_c - \psi_1)\,(p_c - \psi_2) < 0 \end{cases}$$

$$\frac{\partial^2 f}{\partial x \partial y} = \begin{cases} \frac{1}{4}; & \text{if } (p_c - \phi_2)\,(p_c - \psi_2) \geq 0 \\ -\frac{1}{4}; & \text{if } (p_c - \phi_2)\,(p_c - \psi_2) < 0 \end{cases}$$

Likewise, we are able to get the Pixel Taylor feature $f_n(p_c)$ (when $n > 2$) as indicated by previously mentioned equations. At the point when every one of the pixels of given face images is mapped to Pixel Taylor feature space, we get the TFM (Taylor Feature Map) for any analysis. After that, we move towards the last step.

Step 3: Taylor Feature Pattern Generation After all the face images are mapped to Pixel Taylor feature space, then the resulting image was divided into R blocks to induce the natural features of the face image. Here, R is set to 6×6, 8×8, 10×10, 12×12 respectively. Experiment section demonstrates the effect of different block numbers on the algorithm' s performance on various datasets. Subsequently, Taylor Feature Pattern of the Taylor feature map was computed. Description of the Taylor Feature Pattern are given below:

2D Taylor Feature Pattern (2DTFP) Here, compute the Taylor Feature Pattern of each Pixel Taylor Feature $f_n(p_c)$ in each block of the TFM (Taylor feature map). Computation of Two Dimensional TFP (2DTFP) is same as the one dimensional TFP. The 2DTFP of $f_n(p_c)$ is characterized as

$$2DTFP = \sum_{j=1}^{8} S(f_n(p_c), f_n(p_j).2^{j-1} \tag{3.14}$$

$$S(p_c, p_j) = \begin{cases} 1; & \text{if } p_c \geq p_j \\ 0; & \text{if } p_c < p_j \end{cases}$$

By the above mentioned equation, 2D Taylor Feature Pattern of each pixel of each block in the Taylor Feature map is computed. Then concatenate all the histogram of each block to make a long feature vector. Finally, a feature vector of the given facial image using 2D Taylor Feature Pattern (2DTFP) histogram of the TFM (Taylor feature map) was constructed. Summarization of our Technique given in Table 3.1.

3.4.1 Effect of the 2D Pixel Taylor Feature Order

Higher-order $f_n(p_c)$ will extract a lot of essential texture data from the given face image, and this is the main reason behind the enhancement of the recognition accuracy. But it is not always necessary that the recognition rate would not increase all the time alongside the order. Here,we also compare the $1^{st}order$ 2DTFP, $2^{nd}order$ 2DTFP, $3^{rd}order$ 2DTFP as well as $4^{th}order$ 2DTFP recognition rates. Recognition rates on the JAFFE data shown in Table 3.2. Experimental con-

Table 3.1: 2DTFP (Two Dimensional Taylor Feature Pattern)

	Input: Face image I, the block size $R \times R$ for the 2DTFP **Output:** 2DTFP histogram for the given input face image I
1:	Compute TFM (Taylor feature map) of the input face image I as given in Eq. (6.5).
2:	Divide the TF map into $R \times R$ blocks
3:	**for** Every block within the TFM (Taylor Feature Map)
4:	**for** Pixel Taylor Feature $f_n(p_j)$ in each block
5:	Calculate 2DTFP operator of$f_n(p_j)$ as Eq.(3.14)
6:	**end for**
7:	Every block in TFM built the sub-2DTFP histogram
8:	**end for**
9:	Then generate a feature vector by concatenation of all sub-2DTFP histogram
10:	**Return** Histogram OF 2DTFP feature of given face image I

sequences demonstrate that the $2^{nd} order$ 2DTFP (92.78%) accomplishes the most effective recognition rate. In this article, 2D Taylor expansion order n are continually fixed to 2 within the accompanying experiments to induce the best performance.

Table 3.2: Comparison recognition rates of 2DTFP with various order of JAFFE dataset (In %)

	1^{st} Order **2DTFP**	2^{nd} Order **2DTFP**	3^{rd} Order **2DTFP**	4^{th}Order **2DTFP**
Recognition rate	87.78	**92.78**	89.87	87.90

3.5 Experiments and Results

Facial expression recognition (FER) experiments are performed on a number of the benchmark FER databases. Face images are in extremely high dimensions. Dealing with such extensive information turns out to be exceptionally trying for the machines. Hence the modular approach is applied where only some informative regions of the face are considered. Facial expressions offer cues of the emotional state of the person even while not communicating verbally. Eyes are the most communicative part of someone's face and reveal sufficient information regarding the sentiments. Aside from the eyes, forehead, nose, lips, and so forth

additionally are instructive locales. Throughout the task of expression analysis, we discovered that apart from the eyes, forehead, nose, lips/mouth additionally plays a significant role as far as expressions are involved [169]. Now, most of the facial expression recognition technique was applied to full-face images. This paper focuses on only some informative regions of the face, as discussed.
To approve the hypothetical conclusion Two Dimensional Taylor expansion, experiments were performed on the some real datasets. For the classification we used SVM [182] and K nearest-neighbor (K=1,2,3) [70] classifier with various distance measures. Euclidean distance, Chi-square distance, as well as histogram intersection (HI) are utilized in our experiments. Which are defined as in Eq. 6.7, 6.8 and 6.9

$$d(x_1, y_1) = \sqrt{\sum_{i=0}^{n} (x_{1i} - y_{1i})^2} \tag{3.15}$$

$$\chi^2 = \sum_{i,j} \frac{(x_{1i,j} - y_{1i,j})^2}{(x_{1i,j} - y_{1i,j})} \tag{3.16}$$

$$D_{HI}(x_1, y_1) = -\sum_{i,j} min(x_{1i,j}, y_{1i,j}) \tag{3.17}$$

Always SVM gives better results compare to K-NN with different distance measures on 8×8 block size. So the given tables in next section shows the comparison between the holistic approach and modular approach only on the SVM as the classifier because SVM is performed better than other classifiers. Similarly graphs of all datasets show the comparison of the results with different classifiers. Experiments on the datasets are below:

3.5.1 JAFFE Dataset [104]

JAFFE dataset consists 213 face images of seven facial expressions presented by 10 Japanese female models. All the face images are of size 256×256 which are cut as appeared in Fig 3.6. The sizes of informative areas are : forehead 54×44, eyes 39×117, nose 46×55, lips 28×74. Out of 213 images, random 70% images were

chosen for training and the remaining 30% were used for testing. Table 3.3 reports average recognition result of 20 such iterations. The results of the comparison between all the above mentioned classifier is shown in Fig 3.7.

Figure 3.6: Examples of facial expressions from JAFFE dataset: (a) angry (b) disgust, (c) fear, (d) happy, (e) neutral, (f) sad, (g) surprise

Table 3.3: Comparison of the 1D Taylor Expansion with the 2D Taylor Expansion (Hoslistic Vs Modular (both ways)) in the light of SVM as a classifier for JAFFE dataset

JAFFE Database								
Block size	**1D Taylor Expanison**				**2D Taylor Expansion**			
	Holistic		**Modular**		**Holistic**		**Modular**	
	TF Map	TFP	TF Map	TFP	TF Map	2DTFP	TF Map	2DTFP
With out blocking	75.07	79.68	76.59	80.68	78.59	83.02	78.07	83.97
6×6	78.65	80.12	82.12	85.79	80.78	85.71	85.21	90.41
8×8	80.17	83.72	85.76	88.78	82.67	88.78	88.07	**92.78**
10×10	78.65	73.76	76.75	86.68	75.23	78.76	80.78	88.08
12×12	75.62	78.67	80.65	86.08	78.93	80.68	83.79	89.08

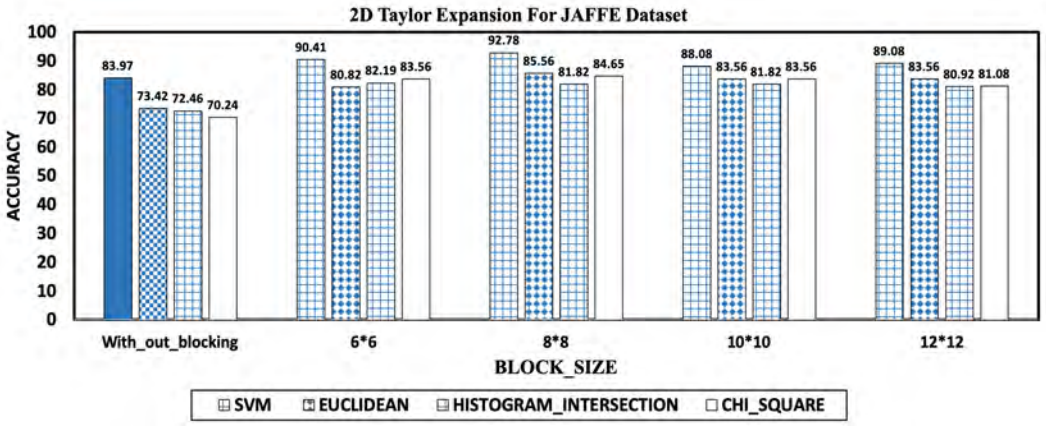

Figure 3.7: Compare the results of SVM and K-NN with different distance measures on JAFFE database

3.5.2 VIDEO (DA-IICT) Dataset [153]

Video (DA-IICT) dataset has been made within which videos of eleven subjects are recorded. Each video contains four different expressions: Normal, Smiling, Angry, and Open mouth, as shown in Fig 3.8. The sizes of informative areas are : forehead 51×60, eyes 40×126, nose 49×38, lips 40×60. Out of 6668 images, 70% images were indiscriminately chosen for training and rest images were utilized for testing. Average and best recognition results of 20 such iterations are reported in Table 3.4. The results of the comparison between all the above˘mentioned classifier are shown in Fig 3.9.

Figure 3.8: Examples of facial expressions from VIDEO database: (a) Normal (b) Smiling (c) Angry (d) open mouth.

Table 3.4: Comparison of the 1D Taylor Expansion with the 2D Taylor Expansion (Hoslistic Vs Modular (both ways)) in the light of SVM as a classifier for VIDEO dataset

VIDEO Database								
Block size	**1D Taylor Expanison**				**2D Taylor Expansion**			
	Holistic		**Modular**		**Holistic**		**Modular**	
	TF Map	TFP	TF Map	TFP	TF Map	2DTFP	TF Map	2DTFP
With out blocking	75.06	81.35	80.28	87.78	85.65	85.78	85.76	89.08
6×6	78.79	82.18	83.79	89.88	87.79	88.09	88.78	88.78
8×8	80.79	87.78	85.79	90.08	88.79	90.08	90.08	**94.86**
10×10	81.35	85.28	82.18	87.78	87.79	88.08	90.08	92.86
12×12	82.58	80.35	81.59	85.69	86.78	89.65	88.78	90.78

3.5.3 CK+ Dataset [66]

There are 593 sequences across 123 persons giving 8 facial expressions. All sequences are captured from the neutral face to the peak expression. Participants were eighteen to fifty years of age, 81%, Euro-American, 13% Afro-American, 69%

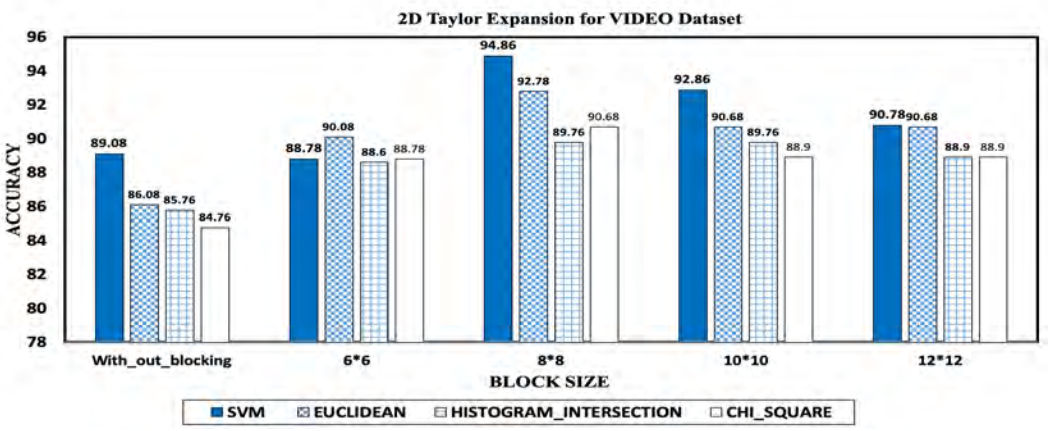

Figure 3.9: Compare the results of SVM and K-NN with different distance measures on VIDEO dataset

female and 6% other groups. Image sequences for frontal views and 30-degree views.

Figure 3.10: Examples of facial expressions from CK+ dataset: (a) sad, (b) happy, (c) fear, (d) surprise, (e) disgust, (f)neutral, (g) angry

For experiment we uses image sequences of 99 subjects with 7 facial expressions. So out of 921 images 70% images were indiscriminately chosen for training and rest images were utilized for testing. The face images of CK+ are cut into four informative regions, as shown in Fig 3.10.

Table 3.5: Comparison of the 1D Taylor Expansion with the 2D Taylor Expansion (Hoslistic Vs Modular (both ways)) in the light of SVM as a classifier for CK+ dataset

	CK+ Database							
Block size	**1D Taylor Expanison**				**2D Taylor Expansion**			
	Holistic		**Modular**		**Holistic**		**Modular**	
	TF Map	TFP	TF Map	TFP	TF Map	2DTFP	TF Map	2DTFP
With out blocking	80.15	85.12	83.25	87.78	83.25	87.78	85.76	89.08
6×6	83.25	87.78	85.76	89.02	87.79	88.67	87.78	90.08
8×8	85.12	89.76	88.12	90.08	88.67	90.12	89.02	**93.78**
10×10	84.76	87.78	88.12	88.12	87.79	89.02	88.12	91.68
12×12	81.25	85.12	85.12	87.78	86.78	88.67	87.78	90.78

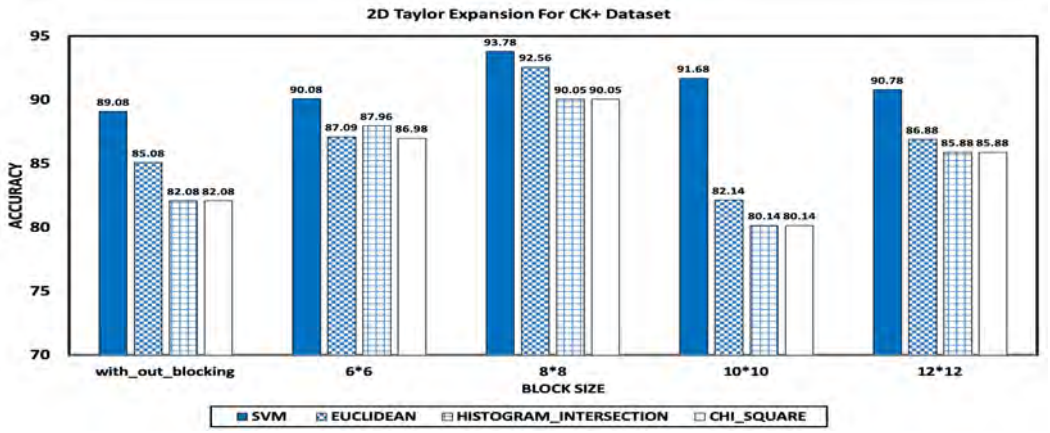

Figure 3.11: Compare the results of SVM and K-NN with different distance measures on CK+ dataset

3.5.4 Oulu - Casia Dataset [211]

Oulu-Casia has 6 facial expressions (anger, happiness, surprise, fear, disgust and sad) form 80 different subjects between 23 to 58 years of age. 73.8% of the persons are males. Out of 3360 images randomly 70% images were chosen for training and remaining 30% images used as testing. Table 3.6 reports average recognition result of 20 such iterations. The results of the comparison between all the above mentioned classifier is shown in Fig 3.12.

Table 3.6: Comparison of the 1D Taylor Expansion with the 2D Taylor Expansion (Hoslistic Vs Modular (both ways)) in the light of SVM as a classifier for Oulu-Casia dataset

Oulu-Casia Database								
Block size	**1D Taylor Expanison**				**2D Taylor Expansion**			
	Holistic		**Modular**		**Holistic**		**Modular**	
	TF Map	TFP	TF Map	TFP	TF Map	2DTFP	TF Map	2DTFP
With out blocking	82.51	87.34	85.43	89.90	85.25	89.67	87.42	90.18
6×6	84.12	88.69	87.94	90.41	89.10	90.10	89.86	91.42
8×8	86.21	90.78	90.42	91.27	90.89	91.42	91.00	**96.87**
10×10	87.68	88.45	90.48	90.32	89.98	90.41	90.90	93.86
12×12	83.52	86.14	87.43	88.68	87.46	89.98	89.45	91.52

We also did some experiments that is based on combination of proposed method and existing hand-craft based features methods as reported in Table.3.7. Where as Table .3.7 reported the the existing methods which are the combination of different hand crafted features.

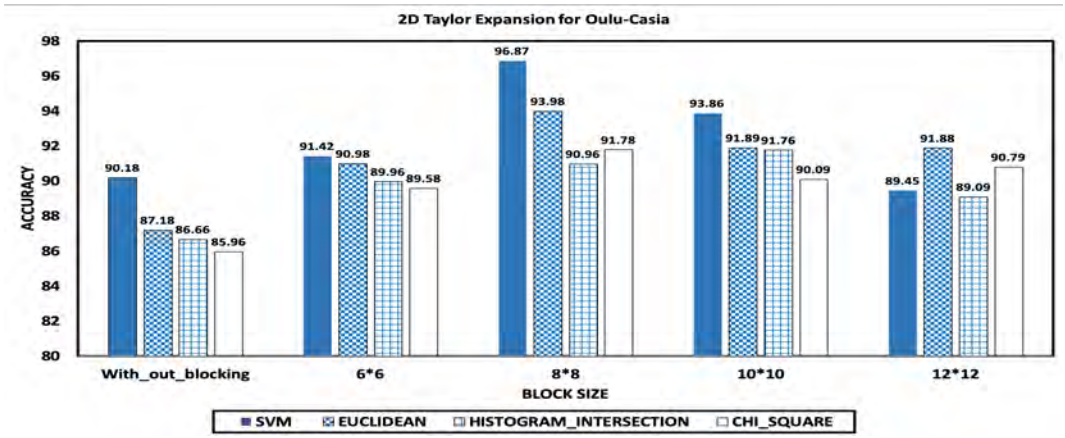

Figure 3.12: Compare the results of SVM and K-NN with different distance measures on Oulu-Casia dataset

Table 3.7: Results of Combination of proposed method with other existing hand craft features based methods in the light of SVM as a classifier

Methods	JAFFE	CK+
LBP+1DTFP	85.93	80.45
LBP+2DTFP	88.38	84.45
1DTFP+2DTFP	89.93	87.45

3.6 Conclusion

In our designed system, we proposed two dimensional Taylor expansion for the facial feature extraction as well as to handle the local illumination. Most procedures just used the arrangement with global illumination varieties and thus yielded more unsatisfactory recognition performances within the case of natural illumination variations that are usually uncontrolled within the globe. Hence, to address the brightening variety issue, we at that point presented the (LL) Laplace-Logarithmic area in this article for further improving the exhibition. We applied the proposed 2D Taylor expansion theorem in the facial feature extraction phase and formulated the 2DTFP method. Results in experimental section demonstrate that the proposed 2DTFP method can obtain an effective facial expression feature vector from the facial images which work best compared with some other state-of-the-art methods.

CHAPTER 4

HSOG: Histogram Of Second Order Gradients for Feature Extraction

In chapter 2, we have discussed the conventional feature extraction technique for facial expression recognition. Local image descriptors is a Scale Invariant Feature Transform (SIFT) proposed by Lowe [99]. SIFT has been widely studied and has played a dominant role in expression recognition. Its descriptor is represented by a 3D histogram of the gradient locations and orientations whose contributions are weighted by their gradient magnitudes. Mikolajczyk and Schmid [40] extended SIFT to the Gradient Location and Orientation Histogram (GLOH) descriptor to increase both distinctiveness and robustness. Dalal and Triggs [23]presented the Histogram of Oriented Gradient (HOG) descriptor. HOG combines both the properties of SIFT and GLOH. This method that was initially developed for person detection is used in more general object detection algorithms. Using the concept of HOG Mohamed Dahmane and Jean Meunier [21] comes with HOG for emotion detection. But in the facial expressions, HOG not give a good result as compare. It may be possible the main reason behind this is HOG is working on the 1^{st} order oriented gradient, which only computes the slope, but in the case of facial expression need to compute other factors too.

In this chapter, we discuss the feature extraction technique based on a Histogram of second-order gradients of the image. HSOG is a variant of the HOG (Histogram of the oriented gradient). HOG [23] is a local image descriptor for feature extraction, mainly used for object recognition. HOG counts the appearance of gradient orientation in the local region of the image. HOG gives better

results compared to other existing image descriptors like SIFT [99], GLOH [40] etc. HOG performs better in each stage of implementation; it gives fine gradients, coarse spatial binning, and fine orientation binning, high-quality normalization in each overlapping block of the descriptor.

4.1 Histogram of Oriented Gradients (HOG)

HOG [23] is a local image descriptor for feature extraction, which is mainly used for object recognition. HOG counts the appearance of gradient orientation in the local region of the image. HOG gives better results compared to other existing image descriptors like SIFT, GLOH, etc. HOG performs better in each stage of implementation; it provides fine gradients, coarse spatial binning, and fine orientation binning, high-quality normalization in each overlapping block of the descriptor. Overview of feature extraction technique by HOG is given in Fig 4.1.

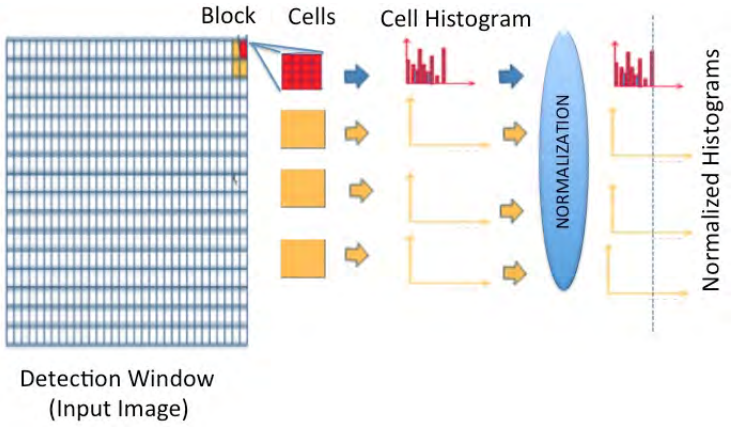

Figure 4.1: Overview of HOG Image descriptor

HOG mainly focuses on the normalized local histogram of the gradient orientation of the image is a dense grid. The key idea is that image gradients or edge directions characterize any object present. This is implemented by dividing the image window or detection window into small connected regions called cells. Each cell computes a histogram of gradient directions or edge orientations for the cell's pixels. We discretized each cell into angular bins according to the gradient orientation. Each cell's pixel contributes a weighted gradient to its corresponding angular bin. Groups of adjacent cells are considered as spatial regions called

blocks. The grouping of cells into a block is the basis for grouping and normalization of histograms. A normalized group of histograms represents the block histogram. The set of these blocks histograms represents the descriptors.

The use of orientation histograms has many precursors [109], but reached maturity only in combination with the local spatial histogram and normalization in Lowe's Scale Invariant Feature Transformation (SIFT) approach to the broad correspondence of the baseline image matching [99], where the description of the image patch below is provided to match the scale-invariant key points. The SIFT-style approaches the extraordinary application well in this application [99][111].

The Shape Context [7] work was studied as an alternative to cells and blocks. Initially, it was used only as an edge pixel count without the orientation histogram that makes the representation so effective. The success of these scattered representations has, in any way, overshadowed the power and simplicity of HOG as a dense image descriptors. Dalal and Triggs presented the Histogram of Oriented Gradient (HOG) descriptor. HOG combines the properties of SIFT and GLOH because it is also represented by the 3D histogram of the positions and orientations of the gradient and uses both rectangular and log-polar grids. The main difference between HOG and SIFT is that HOG is calculated on a dense grid of equidistant cells, with overlapping local contrast normalization.

We focus on the discriminatory power of local image descriptors and study a new one based on second-order gradient clues, i.e., second-order gradient histograms (HSOG), capable of simulating the visual characteristics perceived by the human being. Indeed, some more recent studies on human vision suggest that the neural image is a landscape or a surface, made up of elements such as cliffs, ridges, peaks, valleys or basins, whose geometric properties can be uniquely and accurately characterized by the local curvatures of the differential geometry through the information related to the second-order gradient. While the gradients of the first order measure only the slope of the luminance profile at each point and, therefore, give the amount of elasticity of a surface, e.g., length, area. In the theory of differential geometry, slope and curvature are different geometric traces that can be measured at each point on a 1D curve. As we know, the first-

order gradient calculated at a point delivers the slope or velocity of the curve at that point that encodes the metrics, for example, the length of that curve, whereas the second-order gradient at that point is the quantity relative to the local curvature or how much the curve bends. Now the retinal image is a landscape or surface embedded in 3D space and the local shape around a point. On a smooth 2D surface embedded in a 3D space, one can compute the two principle curvatures, i.e., the maximum and minimum curvatures, which can be calculated from the second fundamental form, which is closely related to the second-order gradient cue. Their joint variations, example, through the value of the shape index, define various local shapes.

Following the insights conveyed by recent research on human vision, as well as the existing differential geometry tools, we hypothesize that local image descriptors calculated on a point should exploit the second-order gradient information to account for its local shape attributes of a retinal image in curvature terms and therefore provide additional discriminating power with respect to their first order gradient based counterparts. However, since first-order gradients are quantities related to surface metrics, such as length, angle and area, while second-order gradients correspond to curvatures, these two categories of quantities must have some complementarity in the description of a local surface shape. Here we implemented a local image descriptor, namely Histograms of Second Order Gradients abbreviated as HSOG, to characterize local shape changes for images. HOG image descriptor is unable to find the local shape changes because it is only working on the 1^{st} order oriented gradient. But for the facial expressions it is needful to find out local shape changes of images. We computes the second order gradient that gives the curvature at point, helps to find the local shape changes of image. That shapes basically corresponds to different facial expression images. In the following section proposed image descriptor defined in detailed.

4.2 Proposed HSOG (Histogram of second order gradients) Image Descriptor

Histogram of Oriented Gradients (HOG) is a window-compatible feature descriptor that uses the gradient filter. The extracted features are based on the information on the edges of the registered face images. It extracts the visual features; for example, a smile expression means curved eyes. The Histogram of Gradients (HOG) approach represents images based on the directions of the edges contained in them. HOG extracts local features by applying the gradient operators on the image and encoding the output in terms of gradient magnitude and angle. First, local magnitude – angle histograms are extracted from the cells, then these local histograms are combined into larger entities (blocks): the dimensionality increases when the blocks are overlapping [31]. HOG has been used by a prominent system in the FERA emotional challenge [20]. HOG features capture global and detailed information from facial images and, therefore, reflect an individual's expression. However, these features are drawn from the whole facial region, and local regions that are closely related to changes in expression, such as eyes, nose, and mouth, are ignored. Therefore, the geometric features, represented by the geometric relationships of the facial landmarks detected from local regions that are closely related to changes in expression, are used for FER tasks. Furthermore, the combination of different features is a promising trend.

The FER system introduced by Yan [190] is one of the most recent methods that has used both visual and audio information. They proposed a new collaborative discriminative multimeric learning (CD-MML) to recognize facial expressions in videos. For the visual feature, there were two types of functionality: 3D-HOG and geometric wrap feature. By extending the traditional 2D HOG and obtaining three orthogonal planes, a HOG feature was extracted from each block located on each plane. These HOG functions were then combined to form a descriptor for each frame, and a high-dimensional feature vector finally described the whole face. The extensions of HOG can be found in Co-occurrence histograms of oriented gradients (CoHOG [129]) and Coherence Vector of Oriented Gradients (CVOG

[130]).

Our proposed method is inspired by the HOG, which characterizes the local shape changes of the face (called the facial expression) by encoding the second-order gradient from the first order oriented gradient. The first order gradient delivers the slope, whereas Second-order gradients compute the curvature at the point that curvature gives shape index, and different shape index corresponds to different local shapes. In the concept of differential geometry, slope and curvature are distinct geometric cues that can be measured at each point in the one-dimensional curve. It can be observed that first-order gradient computed at each point of the curve, which gives the slope or can say that velocity of the curve at that point, that encodes the matrices, that delivers, for example, length of that curve. Whereas second-order gradient at that point measures the amount by which a geometric object such as surface deviates from being a flat plane or how much the curve bends that is called curvature at that point. To find the local shape around a point, it needs to compute the minimum and maximum curvatures, which can be calculated by the second-order gradient cue. Their joint variations give a different shape index; its different values correspond to different local shapes.

After computing the second-order gradient, a simple concatenation strategy is applied, like other image descriptors also do. We divide the image into different block sizes like 8×8, 12×12, 16×16 that enables the slight displacement in the second-order gradients in the neighborhood at the point. The process works in three-fold for computing the descriptor. The overview of the process is given in Fig .4.2 for a face image. Details of these steps are available in [54]. We are redefining a few steps and mainly the pooling strategy. The steps are discussed in the following subsections.

4.2.1 Computation Of First Order Oriented Gradient Maps (OGMs)

Image descriptor starts from computing the 1^{st} order oriented gradient maps (OGMs). For a given image region I, there are specific gradient maps G_1, G_2, ..., G_M for each pixel (x,y) in one of the quantized direction. The G_i OGMs is defined as:

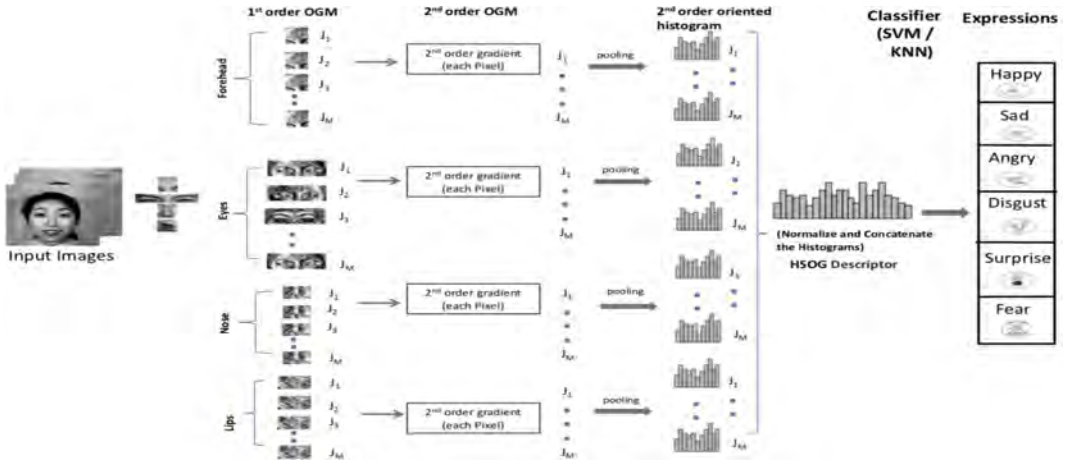

Figure 4.2: Overview of the proposed method for computing HSOG for face image

$$G_i = (\frac{\partial I}{\partial i})^+; i = 1, 2, ...M \tag{4.1}$$

Here, $'+'$ sign indicate that only the positive values are considered, While $'-'$ values set to 0. Positive values sufficient to retain the local information that is needed for the methodology. Negative values are automatically considered in the filter rotated by 180°.

Each G describes the gradient norms of I region in a direction $'i'$ for each pixel location. After that the convolution of these gradient maps G with the Gaussian kernal G is performed, denoted by

$$\rho = G_i * G \tag{4.2}$$

A crucial issue to be dealt with when computing the second order gradients is the sensitivity of the resultant local image descriptor with respect to noise. The fact of using the Gaussian kernel to simulate human simple cells and smooth first order gradients by 4.2.1 gives descriptor a desirable robustness to noise.

The purpose of convolving Gaussian kernal is to shift the gradient to its neighborhood without any unexpected changes. Collect all the values of convolved gradient maps at each pixel (x,y) in M quantized directions and build a vector ρ^r

of these values.

$$\rho^r = [\rho_1^r(x,y), \rho_2^r(x,y).....,\rho_M^r(x,y)]^T \tag{4.3}$$

In this ρ^r vector, further unit normalization is performed, that is denoted by as $\underline{\rho}^r$. For example

$$\underline{\rho}_1^r(x,y) = (\frac{\rho_1^r(x,y)}{\|\rho^r(x,y)\|})$$

For each given image I, the First Order Oriented Gradient Maps (OGMs) J_i for each orientation i is computed and is defined as

$$J_i(x,y) = \underline{\rho}_i^r(x,y) \tag{4.4}$$

Due to the processing of all these steps, OGMs are having to invariant affine lighting transformations property. According to equation 4.2.1 OGMs J_i is convolved normalized gradient map at each quantized direction i. Any brightness change not affected by the gradient computation only adds a constant intensity value. Change in the image contrast in which the intensity values are multiplied by the constant will result in the multiplication of the gradient computation. Change of image contrast will be canceled by the normalization of the response vector $\underline{\rho}^r$. These properties will be necessary for implementing the image descriptor for facial expression recognition. Using those 1^{st} order OGMs compute the 2^{nd} Order Gradient in the next step.

4.2.2 Computation Of Second Order Gradient

Once the 1^{st} order OGMs is computed in each quantized directions i, they are used as inputs to the 2^{nd} order gradient calculation over the image region I. For each OGMs $J_i(x,y), i = 1,2,.....M$, calculate the gradient magnitude Mag_i and gradient orientation Φ_i at each pixel location, Defined Mag_i and Φ_i defined as

follows

$$Mag_i(x,y) = \sqrt{(\frac{\partial J_i(x,y)}{\partial x})^2 + (\frac{\partial J_i(x,y)}{\partial y})^2} \tag{4.5}$$

$$\Phi_i(x,y) = \arctan(\frac{\partial J_i(x,y)}{\partial y} / \frac{\partial J_i(x,y)}{\partial x}) \tag{4.6}$$

where i=1,2,.....M

$$\frac{\partial J_i(x,y)}{\partial x} = J_i(x+1,y) - J_i(x-1,y) \tag{4.7}$$

$$\frac{\partial J_i(x,y)}{\partial y} = J_i(x,y+1) - J_i(x,y-1) \tag{4.8}$$

Each orientation (denoted as Φ_i) is then mapped from the range of $[-\pi\backslash 2, \pi\backslash 2]$ to that of $[0, 2\pi]$, then quantize into M orientation which are persistent with the number of 1^{st} oriented gradient maps. The value of n_i in each quantized direction is computed as

$$n_i(x,y) = mod(\lfloor(\frac{\Phi_i(x,y)}{2\pi/M}) + \frac{1}{2}\rfloor, M) \tag{4.9}$$

4.2.3 Concatenation

Here we implements a concatenation in which given image/image sequence is divided into different block size with 50% overlapping. Block sizes i.e 8×8, 12×12, 16×16.

Let total number of divided blocks in a given image be D. With in each block D, j= 1, 2,.....,D and each 1^{st} order oriented gradient maps (OGMs) J_i, i=1,2,.....,M and second order gradient histogram h_{ij} is formulated by assembling gradient magnitude Mag_i of all pixels with same quantized orientation entry n_i.

$$h_{ij}(k) = \sum_{(x,y)\in\mathcal{D}_j} f(n_i(x,y) == k) * Mag_i \tag{4.10}$$

where k=1,2,.....M-1, i=1,2,.....M, j=1,2.....D

$$f(p) = \begin{cases} 1; & \text{if p is true} \\ 0; & \text{otherwise} \end{cases} \tag{4.11}$$

For each 1^{st} order OGMs, its second order gradient histogram h_i is computed by concatenating all the histogram from D blocks.

$$h_i = [h_{i1}, h_{i2}, ..., h_{iD}]^T \tag{4.12}$$

where i is number of quantized orientation direction $i = 1, 2, ..., M$. The descriptor H is obtained by concatenating all M histograms of second order of gradients as

$$H = [h_1, h_2, ..., h_M]^T \tag{4.13}$$

4.3 Experiments and Results

Descriptor H has been used for the extracting the facial expression features. The second-order gradient histogram technique applied to some of the informative regions of the face image taken instead of full face. Face images generally comprise of very high dimensions. Handling such large data becomes very challenging for the machines; hence, the modular approach is applied where only some of the face's informative regions are considered like eyes, nose, lips, forehead. Till now, most of the facial expression recognition techniques applied for full-face images. It seems that eyes, nose, lips, and forehead are more informative for identifying a person. By having a look at only one of these face parts or a combination of these parts, the facial expression can be identified. The extraction of these facial regions is already described earlier. Descriptor H is applied to these regions, and classification using these different parts is carried out. Experiments using separate parts are performed on three databases having different facial expressions.

To approve the hypothetical conclusion of the proposed framework, experiments were performed on the four facial datasets. 1) JAFFE database [104] 2) The

Video database [153] 3) In CK+ [101] 4) Oulu-CASIA [211] and the same classifier as we used in chapter 2.

The current work is more emphasized for extracting the strong features in different datasets rather than classification. After evaluation found results of NN and SVM in different datasets is presented in Table 4.1. It is clearly seen on the table accuracy of KNN is keep high in most of the cases. The confusion matrix of the jaffe database using 8*8 block size is given in Table 4.2. For the VIDEO, CK+ and Oulu-casia is given in Table 4.3 4.4 4.5. A comparison of the recognition rates achieved by proposed second order histogram method with existing Histogram of Oriented Gradient (HOG) for facial expressions recognition is depicted at Table 4.6. Comparison of HSOG with previously implemented 2DTFP is reported in Table 4.7.

Table 4.1: Comparison The Performance Of SVM and KNN in different datasets

Database	Blocksize	1-NN	2-NN	3-NN	SVM
JAFFE	8	**94.15**	90.41	87.67	90.41
	12	90.41	87.67	89.04	89.04
	16	91.78	84.93	86.32	89.04
VIDEO	8	**95.98**	95.32	94.94	91.49
	12	94.89	95.30	95.14	88.24
	16	92.80	93.33	93.80	86.46
CK+	8	**93.89**	82.78	82.40	82.50
	12	82.14	76.07	76.07	70.35
	16	80.00	75.30	74.64	67.14
Oulu-CASIA	8	**97.00**	85.86	85.04	84.50
	12	92.61	86.05	84.07	80.15
	16	92.89	88.43	87.34	77.33

Table 4.2: confusion matrix of JAFFE database (Block size 8*8)(In%)

Label	Happy	Disgust	Fear	Angry	Sad	Surprise	Neutral
Happy	81.81	0.0	0.0	0.0	9.09	0.0	9.09
Disgust	0.0	100	0.0	0.0	0.0	0.0	0.0
Fear	0.0	8.3	91.66	0.0	0.0	0.0	0.0
Angry	0.0	0.0	0.0	100	0.0	0.0	0.0
Sad	9.09	0.0	9.09	0.0	81.81	0.0	0.0
Surprise	0.0	0.0	0.0	0.0	0.0	100	0.0
Surprise	0.0	0.0	0.0	0.0	0.0	0.0	100

Table 4.3: confusion matrix of Video database (Block size 8*8) (In%)

Label	Angry	Normal	Smile	Open Mouth
Angry	94.83	2.58	0.0	2.58
Normal	2.55	96.16	1.27	0.0
Smile	0.63	0.63	95.36	3.36
Open Mouth	2.17	0.24	1.45	96.12

Table 4.4: confusion matrix of CK+ database (Block size 8*8) (In%)

Label	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	86.79	5.66	1.88	1.88	1.88	1.88
Disgust	6.25	79.16	2.08	8.33	0.0	4.16
Fear	0.0	7.14	71.42	7.14	0.0	14.28
Happy	0.0	0.0	1.49	91.04	0.0	7.46
Sad	0.0	5.5	0.0	0.0	83.33	11.11
Surprise	2.5	7.5	3.75	7.5	2.5	76.25

Table 4.5: confusion matrix of Oulu-Casia database (Block size 8*8) (In%)

Label	Happy	Disgust	Angry	Fear	Sad	Surprise
Happy	98.00	0.00	0.00	0.00	0.00	2.00
Disgust	0.00	96.00	1.00	2.00	1.00	0.00
Angry	0.00	1.00	98.00	1.00	0.00	0.00
Fear	0.00	1.00	0.00	97.00	2.00	0.00
Sad	0.00	1.00	1.00	1.00	97.00	0.00
Surprise	3.00	1.00	0.00	0.00	0.00	96.00

Table 4.6: Results of HOG in different datasets

DataBase	**1-NN**	**2-NN**	**3-NN**	**SVM**
JAFFE	**84.93**	79.45	80.82	70
VIDEO	**90.25**	88.79	87.67	87.09
CK+	**78.21**	74.64	76.07	76.78
Oulu-CASIA	**88.21**	86.74	86.07	84.76

Table 4.7: Compare HSOG with 2DTFP (Two dimensional Taylor Feature Pattern)

DataBase	**2DTFP** [chapter 3]	**HSOG**
JAFFE	92.78	94.15
VIDEO	94.86	95.98
CK+	93.78	93.89
Oulu-CASIA	96.87	97.00

4.4 Conclusion

Here we present an effective local image descriptor, namely HSOG, using histograms of the second-order gradient to capture local geometric properties related to curvature. These experimental results were achieved on four benchmark datasets. This descriptor has the ability of the histogram of the second-order gradient descriptor (HSOG) to recognize a person using partial information from the entire face image. Furthermore, the information transmitted by HSOG is complementary to that acquired by local image descriptors based on the state-of-the-art first-order gradient, for example, HOG, SIFT, CS-LBP, and DAISY. A comparison of HSOG with previously implemented 2DTFP is reported in Table 4.7.

CHAPTER 5

Dimensionality Reduction Based Feature Extraction Techniques

Facial expression recognition is a big problem in the field of Human behavioral analysis. Much work has been done in this field where local texture, features have been extracted and used in the classi- fication. Due to the very local nature of this information, the dimension of the feature vector achieved for the full image is very high, posing computational challenges in real-time expression recognition. In recent times, Dimensionality Reduction methods have been successfully used in image recognition tasks. Though being high dimensional data, natural images such as face images lie in low dimensional subspace, and Dimen- sionality Reduction methods try to learn this underlying subspace to reduce the computational complexity involved in classification stage of image recognition task. Here we proposes the Euler Principal Component Analysis (e-PCA) and Orthogonal Neighborhood Preserving Projection with Class Similarity-based neighborhood (CS-ONPP) for expression recognition.

5.1 PCA and KPCA

Principal component analysis (PCA), is a very prominent technique for dimensionality reduction and feature extraction. Fundamental thought of PCA is to discover the vector which best account for distribution of face images with in whole image space as stated in [175] . In PCA the faces are basically represented as a linear combination of weighted eigen vectors that is called as eigen faces. These

eigen faces are nothing but the dissemination, of faces or can state the eigen vectors of covariance matrix of the set of facial images, where image size is m×n is considered as a point in the *mn* dimensional space.

Principal component Analysis (PCA) is that the most well liked technique. The most advantage of PCA is in a position to capture the direction of the training images with most variance by selecting and manufacturing the orthonormal eigenvectors, however, the PCA coefficients within the subspace don't seem to be associated. During this manner, it will simply hold the global structure. Additionally PCA cannot catch the most straightforward invariance except if the data is explicitly fed to the training system [187]. Likewise, the l_2-norm among the quality PCA is barely best for the case of independent and identically distributed Gaussian noise. it's additionally not sturdy to outliers, like illumination variations and occlusions [97] because of the linear transformation. So as to handle the a lot of sophisticated structure among the data, Kernel Principal Component Analysis (KPCA) is employed. KPCA turn out the non-linear kind of PCA.

In order to reveal a lot of difficult structure at intervals the information, Kernel Principal Component Analysis (KPCA) is employed. KPCA speaks the non-linear kind of PCA. It comes the non-linear feature vector into a high dimensional space to separate the features linearly seperable. Therefore, it will reveal the non-linear structure at intervals the images, and encodes higher order statistics [187]. Still, despite the very fact that KPCA is in a position to beat the constraints of linear transformation, it doesn't contemplate the outliers issues.

5.2 Proposed Euler-PCA based Facial Expression Expression

e-PCA is as late planned by [97] to resolve the outlier issues. It's primarily kernel PCA that has complex number using Euler representation. It utilizes l_1-norm [39] rather than l_2 -norm [175]. This methodology has closely connected with standard PCA thanks to the mapping of the dissimilarity measure between the pixel

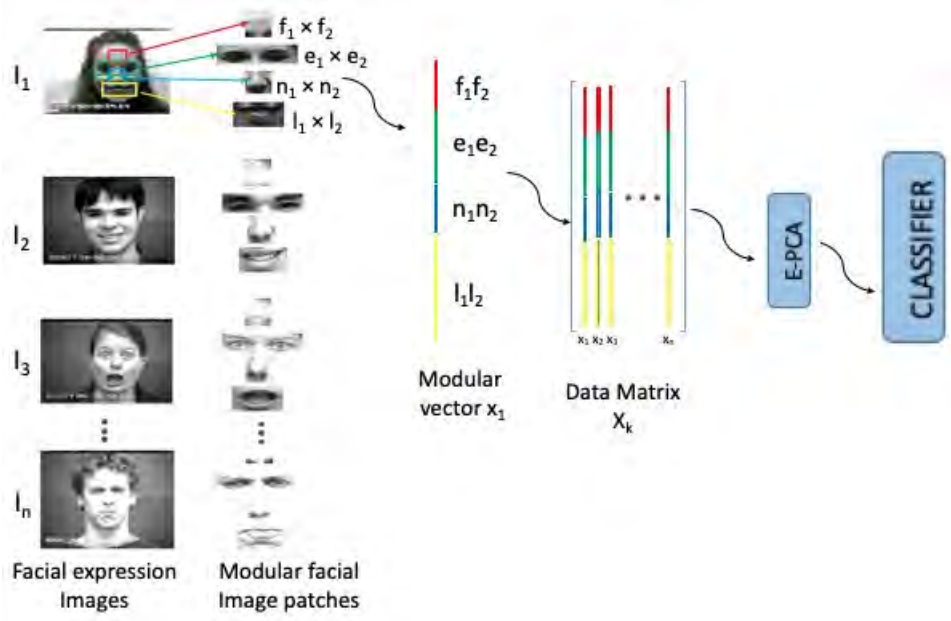

Figure 5.1: Generic pipeline for the proposed facial expression recognition using e-PCA

intensities and the feature space by using the complex number. It retains all the fascinating properties of PCA such as the in-class variations, efficiency and the rotational invariance.

5.2.1 Data preparation for modular expression recognition

Appearance based DR methods consider an $M \times M$ image as a data point in M^2-dimensional space where each image is vectorized either by columns or by rows. Let $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M}$ be the M data points of given training images, so the data matrix $\mathbf{X}$ can be defined as $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M}] \in \mathbf{R}^{M^2 \times M}$. It is proven in [169] that area between eye-brows, eyes, area containing nose and lips plays major role in expressing emotions while the rest of the facial area does not provide any significant information for expression recognition. The modular approach considers only these areas while recognizing expressions, thus the each data vector $\mathbf{x_i}$ corresponding to image i is prepared considering above mentioned four areas from the image. Each significant area is cropped from the whole image and vectorized, these four vectors of facial region are then concatenated to make l-dimensional vector $\mathbf{x_i}$ as shown in Fig. 5.1. Note that the size of these regions across all face

images should be same so that the resulting vector representation be a point in $\mathcal{R}^l$.

5.2.2 Proposed e-PCA

Euler-PCA could be a Kernel PCA that utilizes the robust dissimilarity. It's additionally work on the Euler illustration of of complex numbers. A set of n images $\mathbf{B_k} \in \mathbb{R}^p$, $(k = 1, ..., n)$, of p pixels. Where $p = m \times n$ (each image size). Each image B_k initial remodeled into vector kind, $\mathbf{I_k} \in \mathbb{R}^p$. $(k = 1, ...n)$. Assuming that n images $\mathbf{I} \in \mathbb{R}^{p \times k}$ are provided to the system at the start and every image is in vectorized kind. Then I is normalized to [0,1] vary to get $\mathbf{X} \in \mathbb{R}^{p \times k}$. Then the pixel intensities in $\mathbf{X}$ are mapped on the complex representation, $\mathbf{X} \in \mathbb{C}^{p \times k}$. The specialty of e-PCA utilize the cosine based dissimilarity measure [39] which replaces the l_2 -norm in normal PCA.

$$\mathbf{Z_k} = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{i\alpha\pi X_k(1)} \\ . \\ . \\ e^{i\alpha\pi X_k(p)} \end{bmatrix} = \frac{1}{\sqrt{2}} e^{i\alpha\pi X_k} \tag{5.1}$$

α here may be a real positive issue. It presents the frequency of the cosine function and is optimized to get rid of the values caused by the outliers. It permits the registration of non-rectangular objects. As the estimation of α expands, the massive distance impact caused by the outliers decreases. Compute the Z_k shaped the matrix of transformed data $\mathbf{Z} = [\mathbf{Z_1}, ..., \mathbf{Z_n}] \in \mathbb{C}^{p \times n}$. This can be followed by kernal matrix K, which is computed as

$$\mathbf{K} = Z^H Z \in \mathbb{C}^{n \times n} \tag{5.2}$$

and eigendecomposition of K is computed as

$$\mathbf{K} = U \wedge U^H \tag{5.3}$$

Where U is a complex square matrix of size $n \times n$ whose k^{th} column corresponds to the eigenvectors of K and $\wedge$ is a square matrix whose diagonal elements corresponds to the eigenvalues of K. The eigenvectors area unit positioned during a descendant order in line with their eigenvalues due to the explanation that lowest eigenvalue corresponds to the smallest amount variance [189]. Therefore r reduced set, $\mathbf{U_r} \in \mathbb{C}^{n \times r}$ and $\wedge_{\mathbf{r}} \in \mathbb{R}^{r \times r}$. Then the principle subspace Q is computed as

$$\mathbf{Q} = \mathbf{Z}\mathbf{U_r}\wedge_{\mathbf{r}}^{\frac{-\mathbf{1}}{\mathbf{2}}} \tag{5.4}$$

Fig. 5.2(a) and Fig. 5.2(b) is showing the some samples of the expression images and its eigen faces. For the recognition of facial expression the feature vectors is built as

$$\mathbf{V} = \mathbf{Q}'\mathbf{Z} \tag{5.5}$$

But these feature vectors are in the complex domain. For the further computation we need to go back to the pixel domain. Conversion of complex domain to pixel domain follows as

$$\mathbf{V} = abs(V) \tag{5.6}$$

Where '**abs**' is the absolute value of the complex domain features.

Table 5.1 presents procedure to find feature vector based on e-PCA and mechanism to recognize expression of the test image.

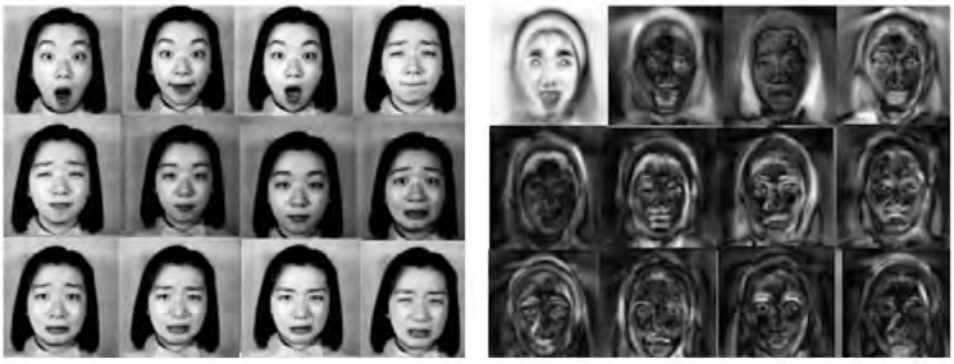

Figure 5.2: (a) Some expression images from the JAFFE database, (b) Eigen faces of these expression images

Table 5.1: Procedure for expression recognition using e-PCA

	Input: A set of n images $I_j, j = 1, ...n$ of p pixels, facial expression images in modular format and number of reduced dimension r and parameter α **Output:** Lower dimension representation $V \in R^{r \times n}$
1:	Represent I_k in the range$[0, 1]$ and obtain X_k by writing I_k in lexicographic ordering
2:	Compute Z_k using eq.(5.1)
3:	Compute kernel matrix K by eq.(5.2) and eigenvalue decomposition by eq.(5.3)
4:	Find the k reduced set, $U_k \in \mathbb{C}^{n \times r}$
5:	Compute principle subspace Q by eq.(5.4)
6:	Compute Embedding on lower dimension by eq.(5.5) and (5.6)
7:	Project test facial expression image represented as modular vector $\mathbf{X_t}$ on learned epca space to get low dimensional representation $\mathbf{V_t}$
8:	Use 1-NN classifier to identify the class label for test image

5.2.3 Experiments and Results

Experiments performed on three well-known facial expression databases: JAFFE, Video, CK+ and Oulu-Casia datasets. As a classifier 1-NN is used because of its simplicity. The purpose of this method is to prove suitability of DR based method for FER, thus sophisticated classifiers are not employed here. Table 6.2 reports average recognition rate for the holistic (where full face considered) and modular (only eyes, nose, lips and forehead considered) approach both by using the proposed e-PCA. Where as Fig. 5.3 [a][b][c][d] compares the results of modular approach by using PCA, KPCA, and proposed e-PCA with different reduced dimensions.

Table 5.2: Accuracy of e-PCA on JAFFE, Video and CK+ with best reduced dimensions (r) [In %]

DataBases	Holistic			Modular		
	PCA	KPCA	Proposed e-PCA	PCA	KPCA	Proposed e-PCA
JAFFE (r=50)	78.25	83.72	85.42	83.20	85.43	89.01
VIDEO (r=700)	87.95	89.43	93.47	87.23	93.25	96.05
CK+ (r=200)	81.23	85.62	88.93	78.32	89.75	91.72
Oulu-Casia (r=300)	89.43	90.12	94.12	86.39	94.09	95.32

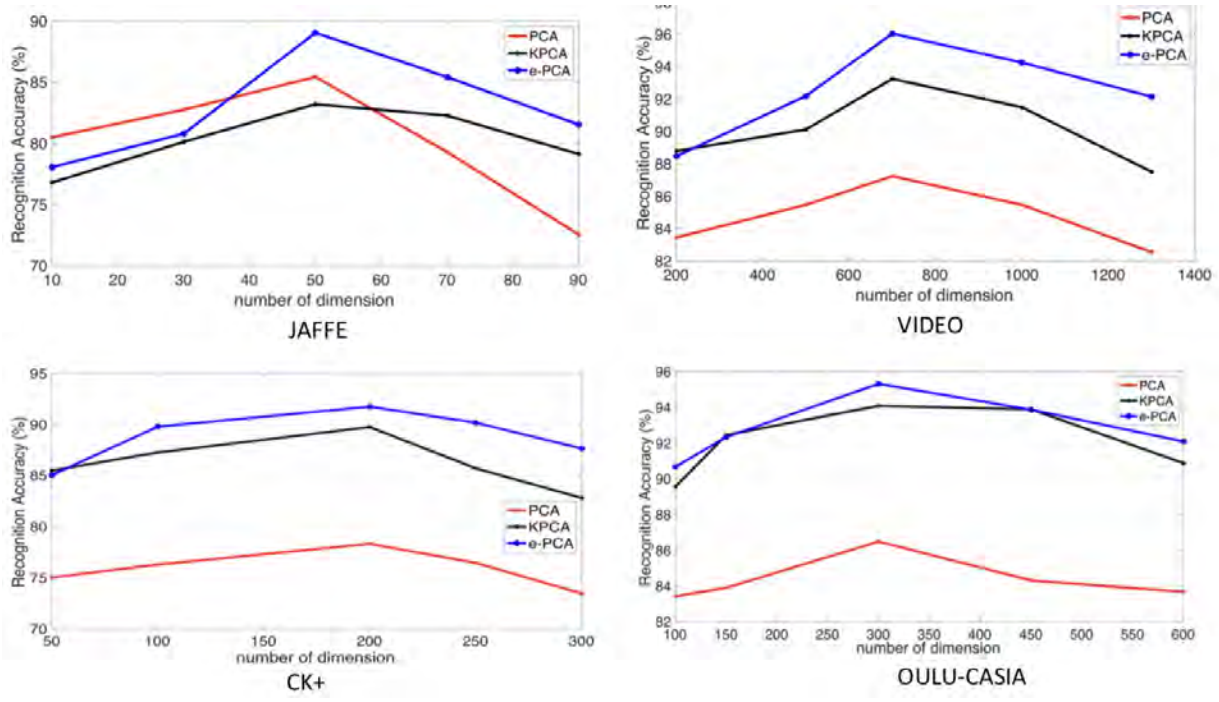

Figure 5.3: Recognition Accuracy (%) with varying number of dimensions (r) for (a) Jaffe (b) Video (c) CK+ (d) Oulu- Casia datasets

Though, Local Features based methods have been successfully applied to Facial Expression Recognition problems. The results we achieved is satisfactory but we want more. So we move some other dimensionality reduction method.

5.3 CS-ONPP: Class Similarity based Orthogonal Neighborhood Preserving Projection

In conventional ONPP [77], the neighbors of data point $\mathbf{x_i}$ are selected based on Euclidean distance (in unsupervised method) or based on class label information

(in supervised method). In [206], authors have proposed a Enhanced Supervised LLE where the Euclidean distance is simply modified by adding a constant for the pairs of data that belongs to different class, keeping others unchanged. Let $\Delta(i,j)$ be the euclidean distance between data points $\mathbf{x_i}$ and $\mathbf{x_j}$. Class similarity based distance denoted by $\Delta'(i,j)$ can be defined by

$$\Delta'(\mathbf{x_i}, \mathbf{x_j}) \quad = \Delta(i,j) + \alpha \max(\Delta)(1 - \mathcal{S}(i,j)) \tag{5.7}$$

where, $\alpha \in [0,1]$ is a tuning parameter. $max(\Delta)$ indicates maximum pair-wise distance or data diameter. $\mathcal{S}(i,j)$ is class similarity between $\mathbf{x_i}$ and $\mathbf{x_j}$, which is defined as,

$$\mathcal{S}(i,j) = \begin{cases} 1; & \mathbf{x_i} = \mathbf{x_j} \\ \mathbf{p}(\mathbf{x_i})^T \mathbf{p}(\mathbf{x_j}); & \mathbf{x_i} \neq \mathbf{x_j} \end{cases} \tag{5.8}$$

We used Logistic Discrimination (LD) to find probability of each data point $\mathbf{x_i}$ belonging to class c_i. Performing LD on high dimensional data causes huge computational burden, thus lower dimensional representation is sought using PCA. Let $\mathbf{z_i}$ be a lower dimensional representation of $\mathbf{x_i}$, to find probability vector $\mathbf{p}(\mathbf{x_i})$. The c^{th} element of $\mathbf{p}(\mathbf{x_i})$ corresponding to class c can be computed by

$$p_c(\mathbf{x_i}) = \frac{\pi(\mathbf{z_i}; \alpha_c, \beta_c)}{\sum_{c=1}^{C} \pi(\mathbf{z_i}; \alpha_c, \beta_c)} \tag{5.9}$$

$$\text{where the function, } \pi(\mathbf{z_i}; \alpha_c, \beta_c) \quad = \quad \frac{\exp(\alpha_c + {\beta_c}^T \mathbf{z_i})}{1 + \exp(\alpha_c + {\beta_c}^T \mathbf{z_i})}$$

The neighbors for each data point $\mathbf{x_i}$ will be chosen based on class similarity based distance given in equation (5.7).

Step 2: Calculating Reconstruction Weight: In this step, the neighborhood $\mathcal{N}_{x_i}$ is expressed as a linear combination of neighbors with reconstruction weight $w_{ij}s$ as $\sum_{j=1}^{k} w_{ij}\mathbf{x_j}$. The weight w_{ij} are computed by minimizing the reconstruction error

i.e. error between $\mathbf{x_i}$ and linear combination of $\mathbf{x_j} \in \mathcal{N}_{x_i}$.

$$\arg\min \mathcal{E}(W) = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \parallel \mathbf{x_i} - \sum_{j=1}^{k} w_{ij}\mathbf{x_j} \parallel^2 \tag{5.10}$$

subject to $\sum_{j=1}^{k} w_{ij} = 1$

For each data point $\mathbf{x_i}$, optimization problem given in (5.10) can be modeled as a least square problem $(\mathbf{X_{N_i}} - \mathbf{x_i e^T})\mathbf{w_i} = \mathbf{0}$ with a constraint $\mathbf{e^T w_i} = \mathbf{1}$. Here, $\mathbf{X_{N_i}}$ is a matrix having $\mathbf{x_j}$ as its columns, where $\mathbf{x_j} \in N_{x_i}$. Note that $\mathbf{X_{N_i}}$ includes $\mathbf{x_i}$ as its own neighbor making it a matrix of dimension $l \times k+1$. Solving the least square problem results in a closed form solution for $\mathbf{w_i}$ given by equation (5.11).

$$\mathbf{w_i} = \frac{\mathbf{G}^{-1}\mathbf{e}}{\mathbf{e}^T\mathbf{G}^{-1}\mathbf{e}} \tag{5.11}$$

Here, $\mathbf{e}$ is a vector of ones having dimension $k \times 1$ same as $\mathbf{w_i}$. $\mathbf{G} \in \mathcal{R}^{k \times k}$ is a Gramiam matrix, each entry of G is given by $\mathbf{g_{pl}} = (\mathbf{x_i} - \mathbf{x_p})^T(\mathbf{x_i} - \mathbf{x_l}),\ for \forall\, \mathbf{x_p}, \mathbf{x_l} \in \mathcal{N}_{x_i}$.

Step 3: Finding Projection Matrix: Last step is dimensionality reduction or finding the projection matrix V that explicitly maps l-dimensional data point $\mathbf{x_i}$ to d-dimensional representation $\mathbf{y_i}$ assuming that the neighborhood relationship among $\mathcal{N}_{\mathbf{x_i}}$ with corresponding weights w_{ij} will be preserved in lower dimensional space, too.

The optimization problem to achieve such mapping can be formed as minimization of the sum of squares of reconstruction errors in lower dimensional space. The cost function is given by

$$\arg\min \mathcal{F}(\mathbf{Y}) = \arg\min_{\mathbf{Y}} \sum_{i=1}^{N} \parallel \mathbf{y_i} - \sum_{j=1}^{k} w_{ij}\mathbf{y_j} \parallel^2 \tag{5.12}$$

subject to orthogonality constraint, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$

Solving the optimization problem results in eigenvalue problem $\mathbf{XMX^T V} = \lambda\mathbf{V}$. Here, columns of $\mathbf{V}$ are eigen-vectors that corresponding to the smallest d eigen-values. The matrix $\mathbf{M} = (\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W^T})$. Note that $\mathbf{XMX}^T$ is symmetric

Table 5.3: Procedure for expression recognition using Class-Similarity based ONPP

	Input: Dataset $\mathbf{X} \in R^{l \times N}$ representing N facial expression images in modular format and number of reduced dimension d **Output:** Lower dimension representation $\mathbf{Y} \in \mathcal{R}^{d \times N}$
1:	Find low dimensional representation $\mathbf{z_i}$ of data by projecting on d_{pca} dimensional space using PCA ($\mathbf{z_i} = V_{pca}^T \mathbf{x_i}$)
2:	Use Logistic Regression on $\mathbf{z_i}$ to find class probability vector $\mathbf{p_i}$
3:	Calculate modified distance for all data point pairs $\Delta'(\mathbf{x_i}, \mathbf{x_j})$ using equation (5.7)
2:	Compute NN $\mathcal{N}_{\mathbf{x_i}}$ with modified distance$\Delta'(\mathbf{x_i}, \mathbf{x_j})$
3:	Compute the weight W for each neighbor data point $\mathbf{x_j} \in \mathcal{N}_{\mathbf{x_i}}$ as given in equation (5.11)
4:	Compute Projection matrix $V \in \mathcal{R}^{l \times d}$ whose column vectors are smallest d eigen-vectors of matrix $\mathbf{XMX^T}$
5:	Compute Embedding on lower dimension by $\mathbf{Y} = \mathbf{V^T X}$
6:	Project test facial expression image represented as modular vector $\mathbf{x_t}$ on learned ONPP space to get low dimensional representation $\mathbf{y_t}$
7:	Use 1-NN classifier to identify the class label for test image

and positive semi-definite. ONPP explicitly maps $\mathbf{X}$ to $\mathbf{Y}$, which is of the form $\mathbf{Y} = \mathbf{V^T X}$, i.e. each test sample $\mathbf{x_t}$ can now be projected to lower dimension by just a matrix-vector product $\mathbf{y_t} = \mathbf{V}^T \mathbf{x_t}$.

Considering the under-sampled size issue where the number of samples N is less than dimension l. In such situation, the matrix $\mathbf{XMX^T} \in \mathbf{R}^{l \times l}$ will have maximum rank $N - c$, where c is number of classes. To ensure that the resulting matrix $\mathbf{M}$ be non-singular, one may utilize an initial PCA projection that reduces the dimensionality of the data vectors to $N - c$. If $\mathbf{V_{PCA}}$ is the projection matrix of PCA, then on performing the ONPP the resulting dimensionality reduction matrix is given by $\mathbf{V} = \mathbf{V_{PCA} V_{ONPP}}$. Note that the PCA projection is most common pre-processing applied in many dimensionality reduction methods and in this article we are using it in our advantage to define new distance measure for local neighborhood. Table 5.3 gives procedure to find Class-similarity based ONPP subspace and mechanism to recognize expression of the test image.

5.3.1 Experiments and Results

Class-similarity based approach applied to ONPP and MONPP, and recognition performance of both the approaches are compared with respective algorithm on some well-known face database and handwritten numerals databases. Table 5.4 reports average and best recognition result of 20 such iterations. Comparison of performance of ONPP [77] and CS-ONPP on facial expression databases in the light of recognition score (in %) with corresponding subspace dimensions are reported in Table 5.5. Where as Table 5.6 reported the comparison between the MONPP [3] and CS-ONPP. CS-ONPP are reported along with tuning parameter Alpha and PCA dimension (d_pca). Table 5.7 repeats the comparisons given in [80] for Local features based Facial Expression Recognition methods along with the dimensions of feature vectors for the given method.

Table 5.4: Best recognition Accuracy (%) achieved with proposed method of three benchmark databases along with related parameters: PCA subspace dimension (dpca), Number of Nearest Neighbors (k), tuning parameter α and ONPP subspace dimensions

Database	Recognition Accuracy (%)	ONPP dimensions	dpca	alpha α	Number of Nearest Neighbors(k)
JAFFE	94.54	100	24	0.25	7
Video	94.76	110	20	0.50	13
CK+	86.76	510	22	0.25	5

Table 5.5: Comparison of performance of ONPP and CS-ONPP on facial expression databases in the light of recognition score (in %) with corresponding subspace dimensions. CS-ONPP are reported along with tuning parameter Alpha and PCA dimension (d_pca).

Databases	**ONPP**		**CS-ONPP**				
	RecAcc	Subspace dim	RecAcc	Subspace dim	d_pca	Alpha	NN_k
JAFFE	93.62	155	94.54	100	24	0.25	7
VIDEO	94.12	175	94.76	110	20	0.5	13
CK+	85.66	705	86.76	510	22	0.25	5
Oulu-casia	91.09	810	92.89	670	26	0.25	10

Table 5.6: Comparison of performance of MONPP and CS-ONPP on facial expression databases in the light of recognition score (in %) with corresponding subspace dimensions. CS-ONPP are reported along with tuning parameter Alpha and PCA dimension (d_pca).

Databases	**MONPP**		**CS-ONPP**				
	RecAcc	Subspace dim	RecAcc	Subspace dim	d_pca	Alpha	NN_k
JAFFE	93.76	145	95.35	85	20	0.25	9
VIDEO	94.4	170	95.66	105	20	0.5	14
CK+	87.14	675	87.88	485	18	0.5	9
Oulu-Casia	93.23	710	94.34	600	20	0.5	9

Table 5.7: Comparison of proposed method with Local feature based methods in the light of Feature vector length for JAFFE dataset for 256×256

Holistic Approaches		
Method	Feature Length	Recognition Accuracy (%)
Local Binary Pattern (LBP)	65536	89.42
Local Gradient Code (LGC)	65536	90.38
Histogram of Gradients (HOG)	20736	85.71
Local Directional Pattern (LDP)	14337	85.20
Modular Approaches		
Method	Feature Length	Recognition Accuracy (%)
Histogram of 2^{nd}Order Gradient (HSOG)	38582	94.15
Proposed ONPP with CS	11541	94.54

5.4 Conclusion

Though, Local Features based methods have been successfully applied to Facial Expression Recognition problems, the resulting feature vector lengths usually are of order 10^5 which slow down classification process. The article proposes a Dimensionality Reduction based method which can be employed in FER. Basically, state-of-the-art DR methods PCA and ONPP are used. Euler -PCA (e-PCA) and a novel approach of neighborhood selection based on class similarity are proposed to suit FER application. Proposed methods is tested on four benchmark databases and proved to be gaining huge margin in terms of feature vector length while maintaining same recognition accuracy.

Till now we worked on feature based approaches, they can do well in certain well-controlled cases. The fundamental issue with hand-crafted features based

arrangement approaches is that they require space learning and not generalize well like in the complex dataset. So we implemented the deep learning models for the FER tasks. Table. 6.3 reported the comparison between dimensionality reduction methods and the previously proposed methods.

Table 5.8: Compare HSOG ,2DTFP , E-PCA and CS-ONPP

DataBase	**2DTFP** [chapter 3]	**HSOG** [chapter 4]	**E-PCA**	**CS-ONPP**
JAFFE	92.78	94.15	89.01	94.54
VIDEO	94.86	95.98	96.05	95.66
CK+	93.78	93.89	91.72	87.88
Oulu-CASIA	96.87	97.00	95.32	93.15

CHAPTER 6

Deep Learning Based Feature Extraction Techniques

A huge volume of existing techniques conducted facial expression recognition that is supported on image/image sequences while not considering temporal data due to the convenience of data handling and the easy accessibility of training and testing material. As mentioned earlier, small FER datasets which directly train the deep neural networks are inclined to overfitting. To moderate this issue, several studies used further task-oriented information to pre-train their networks from fine-tuned or scratch on existing pre-trained models like VGG [160], VGG-face [123], GoogleNet [167] and AlexNet [78]. Kahou et al. [64], [67] demonstrated that the utilization of extra information could get models with high capacity without overfitting, accordingly may improve the FER execution.

The objective of this is more basic, but also more general, namely: can recurring connectivity from associative areas to perceptive areas be useful for classifying expressive events? Our hypothesis is that the deep connectivity of the neural network offers an advantage in recognizing and anticipating more ambiguous expressions. For example, at the beginning of a sequence composed of expressions of neutral with higher intensity. To validate this very general hypothesis using computational models, we compare the simplest and comparable types of deep neural networks to test the importance of recurrent connections, with everything as similar as possible (ie identical learning rate, synaptic weight correction, procedure of training / test, etc.). So we implemented two types methods which are given below:

6.1 DNN based on Fourier transform followed by Gabor filtering

In this segment, we discusses the premise of our technique and proposed framework which improves the efficiency and accuracy of the facial expression recognition. As mentioned earlier in the modular approach, now onward we only consider forehead, eyes, nose, and lips regions. Fig. 6.13 demonstrates the procedure of the proposed framework which is divided into three phases - 1) Preprocessing 2) Feature extraction 3) Classification using SVM and KNN.

6.1.1 Preprocessing

We used the little similar preprocessing as in the EMPATH model given by Dailey et al. [22]. Before the facial recognition, some image preprocessing need to be done first. Our preprocessing starts with the transformation of the input facial image to grayscale. This process minimized the variation of face images. This is a necessary step because CNN depicted later expects 3 channel input facial image, this grayscale facial image is depicted within the 3 channel. Subsequently, we run two procedures, Fourier transforms followed by Gabor filters (improve the speed and encodes the edges) and Data Augmentation (increase number of face images in the database). The subsequent section describes each of those steps in details.

Fast Fourier Transform (FFT) and Gabor filtering

Fast Fourier transform can speed up our procedure very smoothly. Computation of the 1 Dimensional (1D) Fourier transformation of N points specifically requires the order of N^2 addition/multiplication operations. Whereas Fast Fourier Transform (FFT) fulfills the same task in $NlogN$operations. 2D Fourier transform is computed by the given equation-

$$F(p,q) = \frac{1}{MN} \sum_{r=0}^{M-1} \sum_{s=0}^{N-1} f(r,s) e^{-j2\pi(\frac{pr}{M}+\frac{qs}{N})} \tag{6.1}$$

The face images are transformed in the Fourier domain and filtered by 48

Gabor filters (GFs) corresponding to 6 spatial frequencies, with one octave between the focuses of two continuous spatial frequency channels that are $f_i = 5.41; 10.77; 21.60; 43.20; 86.40; 172.8$ cycles per face image and eight exclusive orientations that are $\theta = 0, \frac{\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}$ in radians.

GF can effectively express the characteristics of the texture. It captures the most exceptional visual properties and has very positive results in facial recognition. GF cores that contain the real part and the imaginary part. GF kernels are similar to the profiles of the receptive field in simple cortical cells, characterized by localization, selective orientation and frequency selectivity. An image is processed by the kernel element and, then, to produce its corresponding frequency images, which are further employed to compute to obtain Gabor features for the image.

Different experiments have demonstrated that the use of GFs impacts in a pinnacle estimation of the responsive fields of the primary cells of the imperative visible cortex [55]), given that the applied math analysis of the residual error between the distinction within the response profiles of V1 easy cells and Gabor filters aren't distinguishable from probability [61].

The face images transferred in the Fourier space to boost the speed and ease the

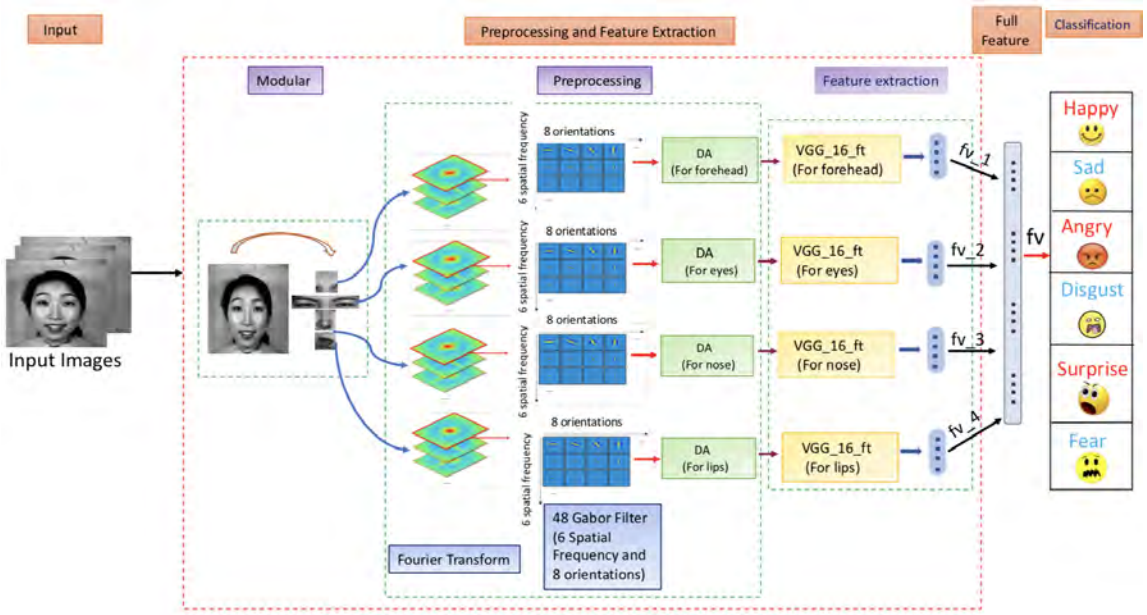

Figure 6.1: Illustration of the proposed Framework

mathematical processes and GFs were applied to every thumbnail by means that of multiplication within the spectral domain (which is resembling a convolution of the Gabor receptive fields within the spatial domain) is:

$$G(p,q) = exp[-(\frac{(u_\theta - f_i)^2}{2\sigma_v^2} + \frac{v_\theta^2}{2\sigma_u^2})] \tag{6.2}$$

where p_θ =p cos θ + q sin θ and q_θ =q cos θ - p sin θ. σ_u and σ_v are standard deviations (SD's) of the Gaussian enfold in the p_θ and q_θ (for example orthogonal to θ). The yields of Gabor channels were the provincial vitality spectra that are multiplied by the kernel of the GF. The GF were applied to the images acquired from the Fourier domain. So now we getting 48 images of each given image from the 6 spatial frequency and 8 Gabor channels.

Data Augmentation

CNN needs massive data so to have the option, to sum up to a given issue. However, publically available FER databases do not have sufficient images to handle the problem. Simard et al. [158] suggested data augmentation (DA) procedure extend the databases through the creation of synthetic face images for every original face image. Inspired by this procedure, the following activities had been utilized as the data augmentation: 1) flipping image vertically and horizontally 2) Rotate each database image, rotate it at right angles if image is square and rotate it as 180^0 if image is rectangular 3) Add the random noise to the landmarks so as to introduce little deformations to faces.

6.1.2 Feature Extraction From Given Facial Images

Our proposed framework utilizes DNN (deep neural networK) for feature extraction for FER is relies on VGG network of Simonyan and Zisserman [160]. They come up with two versions of VGG: VGG-16 and VGG-19 (i.e. sixteen and nineteen layers, respectively). VGG16 is chosen due to the fact of its effective performance in visible detection and speedy convergence. It's concerning 138 million parameters and contains 13 convolutional layers, followed by 3 fully-connected

layers (FCs). The initial two fully connected layers (FCs) have 4,096 outputs and the last layer has 2,622 outputs. Since the VGG framework not designed for the FER tasks so we modified the framework according to our requirements. Fig. 6.9 demonstrates the essential module of the framework. Compared with the original VGG16, our VGG16_ft (where "ft"means fine-tuning) is simplified by doing away with two dense layers.

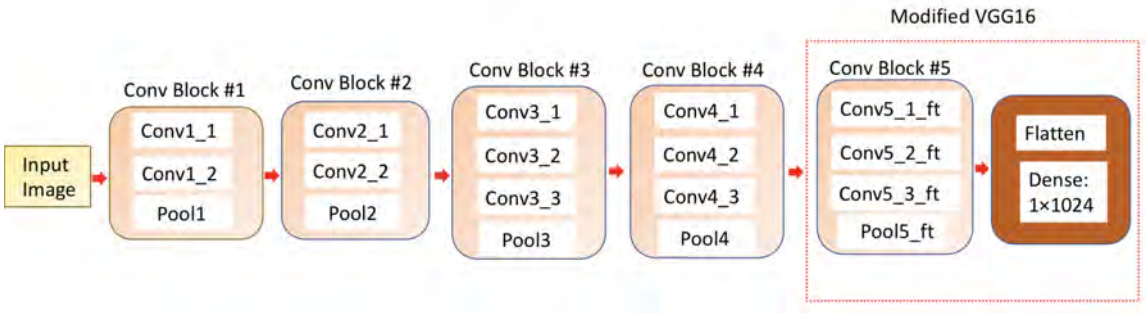

Figure 6.2: Framework for the modified VGG16_ft network, used for extraction of the expression features from the given facial images

Table 6.1: Parameters set for fifth block

	conv5_1 _ft	**conv5_2 _ft**	**conv5_3 _ft**	**Maxpool5 _ft**
Filters	512	512	1024	
size	7×7	5×5	3×3	2×2
stride	1	1	1	2
pad	3	0	0	0

The dimension of the input data for forehead is 54×48, for eyes is 39×117 , for nose is 50×55 and for lips is 48×74. At that point, we fix the structures of the initial four conv (convolution) blocks of the VGG16_ft. But we change the structure of fifth conv block of VGG16_ft and also change the names of each layer just by adding "ft"at the end of the original layer name. So now layer name of fifth conv block is like conv5_1_ft. The parameters whose change the structure of the layer is shown in Table. 6.4. Based on experiments last dense layer preserved and set its dimension to 1×1024. That dimension is actually the extracted feature of input image denoted as feature vector "fv _1" for the forehead, "fv _2" for the eyes, "fv_3" for the nose and "fv_4" for the lips. We decline the learning rates of layers that have a place with the fifth conv block by 10 times (learning rate

for fifth conv block is .001) of other block learning rate (.01 used for other conv blocks) to ensure that that they'll learn more positive information. At last, the initial portion of the system is initialized with the VGG16 model weights which are trained on the Imagenet dataset. ReLu (Rectified Linear Unit) is applied after every convolutional layer.

6.1.3 Concatenation of Different Outputs and Classification

Fig. 6.13 shows our proposed framework. Expression features fv is the concatenation of the feature vector came from forehead (fv_1), eyes (fv_2), nose (fv_3) and lips (fv_4). After getting the feature vector next step to do the classification. In the classification process, the similarity between extracted features of the display set and the probe set is evaluated by the SVM and K nearest-neighbor (K=1,2,3) classifier with various distance measures. Euclidean distance, Chi-square distance, as well as histogram intersection (HI) are utilized in our experiments. As we also did in previous chapters.

For the computation loss, we used the MSE (mean square error) till now it is best for the SVM and KNN classification, Which is defined as

$$Loss = \frac{1}{N}\sum_{1}^{N} \parallel O_i - O_i^{'} \parallel^2 \tag{6.3}$$

Where N is the total numbers of input images, Y and Y' the true and predicted outputs, respectively.

6.1.4 Experimental Results and Analysis

The convergences of the proposed methodology are assessed in four benchmark datasets, and the outcomes are delineated in Figs. 6.3, 6.4, 6.5 and 6.6. Each subfigure demonstrates the trends of accuracy and loss with the rise in iterations. Table 6.2 shows the Comparison between the Holistic and Modular approach in our proposed framework in the light of SVM and KNN as the classifier for all datasets.

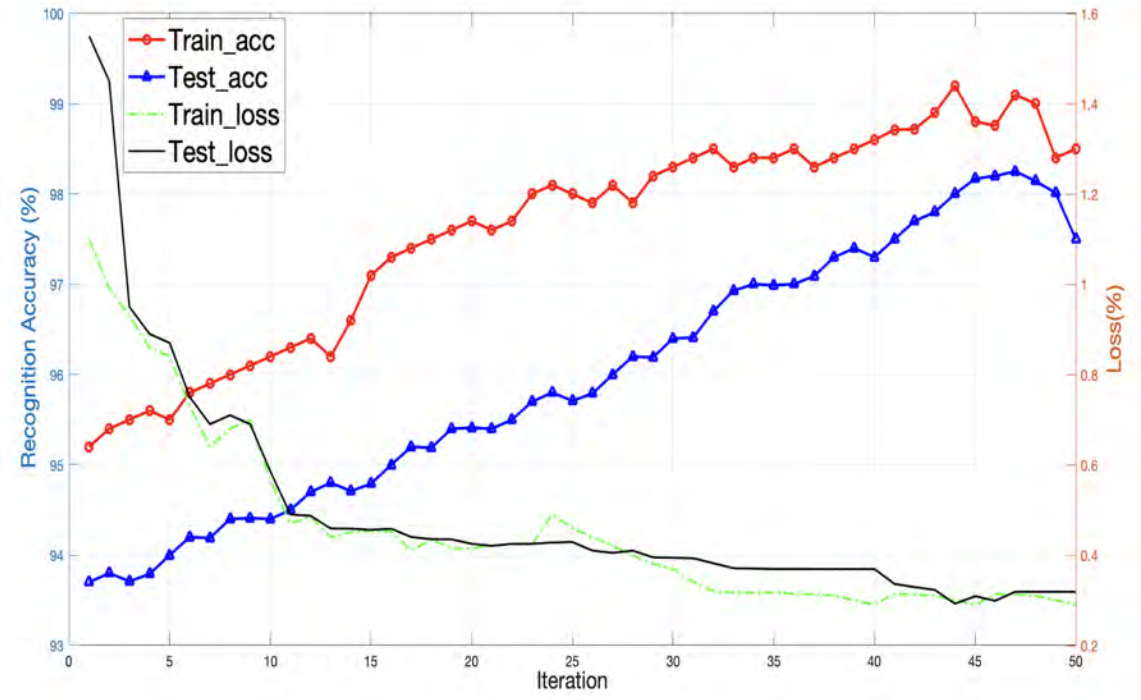

Figure 6.3: Curves of Accuracy and Loss during training and testing phases for JAFFE dataset

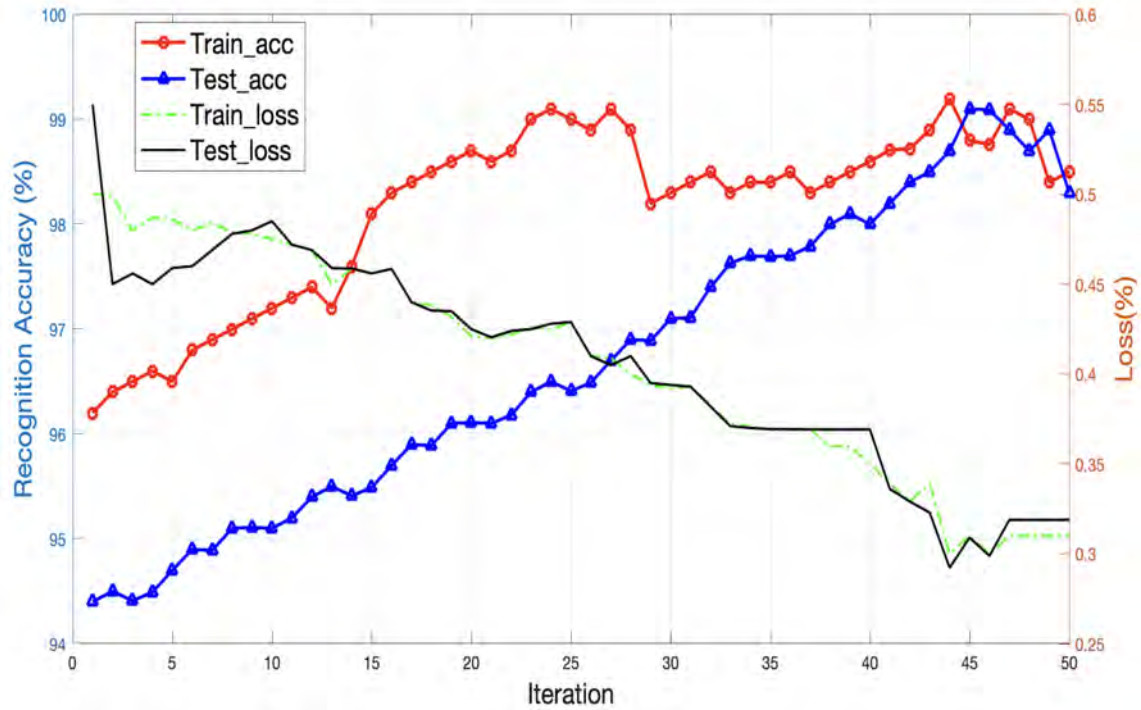

Figure 6.4: Curves of Accuracy and Loss during training and testing phases for VIDEO (DA-IICT) dataset

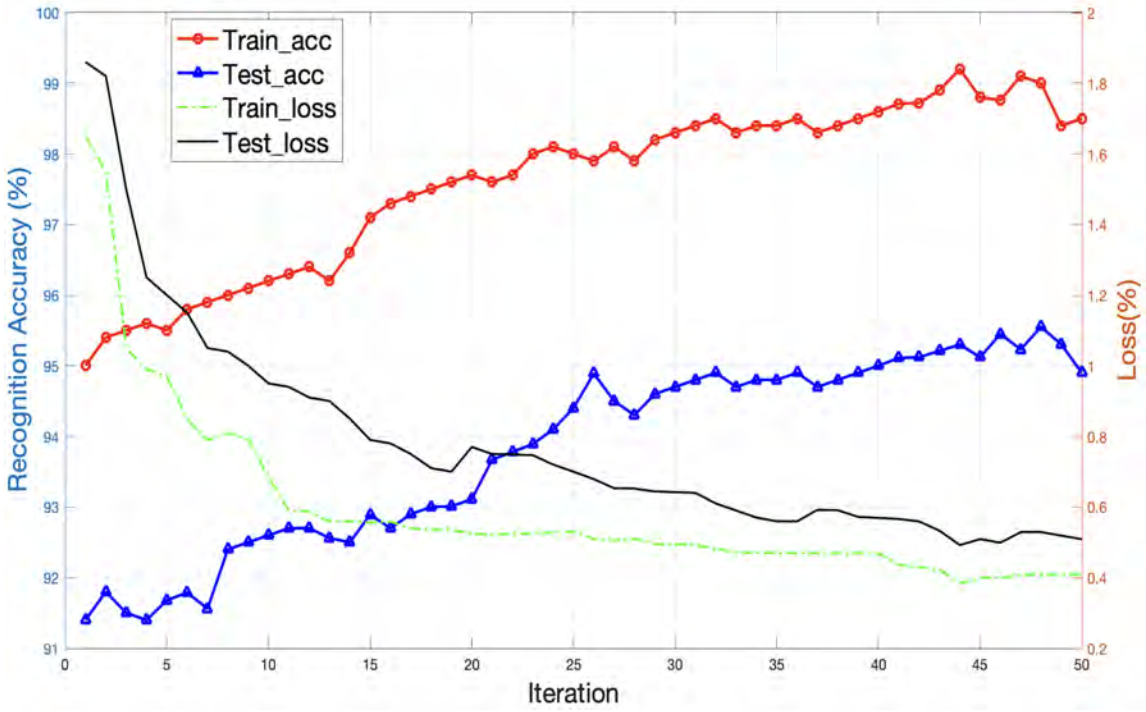

Figure 6.5: Curves of Accuracy and Loss during training and testing phases for CK+ dataset

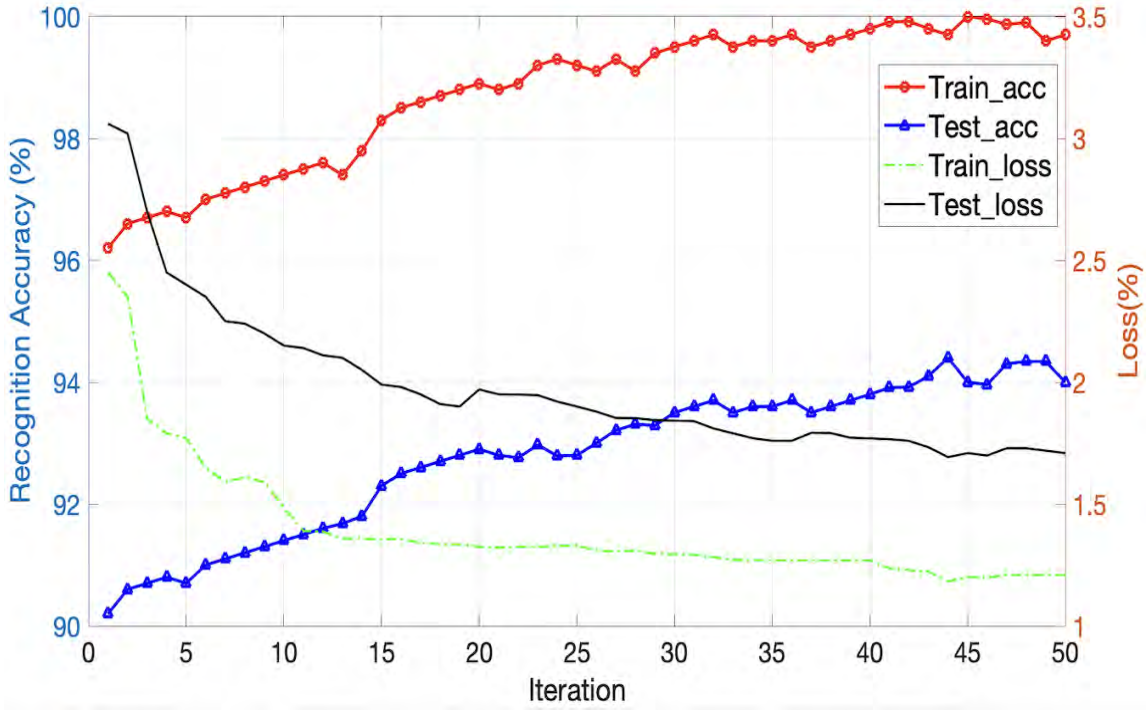

Figure 6.6: Curves of Accuracy and Loss during training and testing phases for OULU-CASIA dataset

Table 6.2: Comparison between the Holistic and Modular approach in our proposed framework in the light of SVM and KNN as the classifier for all datasets (In terms of average accuracy (%) reported for 50 iterations)

Datasets	Holistic				Modular			
	SVM	KNN			SVM	KNN		
		Euclidean	Chi Square	Histogram Intersection		Euclidean	Chi Square	Histogram Intersection
JAFFE	93.02	90.32	82.42	80.02	95.96	93.42	90.32	87.27
VIDEO	92.47	88.23	80.98	78.02	96.67	92.50	89.41	85.49
CK+	91.45	88.71	86.41	83.54	96.78	91.24	87.79	89.64
OULU-CASIA	91.40	87.30	79.89	76.20	96.08	89.56	85.64	83.20

Table 6.3: Compare all proposed methods

DataBase	2DTFP [chapter 3]	HSOG [chapter 4]	E-PCA [chapter 5] [Section 5.1]	CS-ONPP [chapter 5] [Section 5.3]	DNN-FG
JAFFE	92.78	94.15	89.01	94.54	95.96
VIDEO	94.86	95.98	96.05	95.67	96.67
CK+	93.78	93.89	91.72	87.88	96.78
Oulu-CASIA	96.87	97.00	95.32	93.15	96.08

In Table 6.10 we compared all our proposed methods. We have worked on the one channel architecture, which gives us excellent results compared to feature-based methods as we compared in Table. 6.3. But we want to improve the architecture by adding more than one channel in the architecture. In the next section, we implemented double channel-based architecture where we utilized VGGFace [123] instead of VGG16 [160] as we used in this section. VGG16 is trained on millions on images, whereas VGGFace is trained on a large face dataset, it might help increase our FER accuracy.

6.2 Double Channel Based Deep Neural Network

Preprocessing, such as histogram equalization (HE) and data augmentation (DA) are needed for the given facial images. HE is an easy but effective technique in image processing, which might build the distribution of the gray values in numerous images more uniform and decrease interference caused by illumination. CNN

needs massive sets of data to be able to generalize to a given problem. However, publically available FER databases do not have sufficient images to handle the problem. Simard et al. [158] suggested data augmentation procedure to extend the databases through the creation of synthetic face images for every original face image. After the preprocessing, we able to extract facial features from the given facial regions. Appearance and Geometric features extraction techniques are generally used. In earlier research, the region of various facial points are extricated and combined it and made into a feature vector which encodes the geometric information of face like distance, angle, and location [57]. Appearance-based features are utilized to demonstrate the appearance variations of a selected face by way of spatial evaluation [155]. Motion's information features are being used for facial expression recognition in given facial images [177]. Finally, the suitable classifier is used for expression recognition on the extracted features.

Here we primarily depicted the premise of our technique and proposed the framework, which improves the efficiency and accuracy of the facial expression recognition. As mentioned earlier here, we used the modular approach where we only take forehead, eyes, nose, and lips. So from onwards, we worked on these four facial regions. The proposed FER technique utilized in this paper is based on the double channel architecture that can do expression recognition efficiently. Fig. 6.13 demonstrates the procedure of the proposed framework, which is separated into three phases - 1) Preprocessing, 2) Double channel feature extraction technique 3) Classification using SVM and KNN.

6.2.1 Preprocessing

Before the facial recognition, some image preprocessing need to be done first. Our preprocessing begins with the transformation of the input face image to grayscale. This process minimized the variation of face images. This preprocessing is a necessary step because CNN depicted later expects 3 channel input face image and received grayscale face image can be represented within the 3 channel. Subsequently, we run two procedures, which are Histogram Equalization (illumination handling) and Data Augmentation (increase number of the face image in the

database). The subsequent section describes each of those steps in details.

Histogram equalization

In the face images, some issues should also be considered. Due to the different illumination conditions, while taking images, the segments of the face will show in various brightness, which may cause massive interference on facial recognition results. Thus, we tend to conduct histogram equalization (HE) earlier than recognition. The histogram equalization is the distribution of a particular type of data. By equalizing the histogram we can improve the contrast and appearance of an image. The entire pixel spectrum (0-255) will be extended by the histogram equalization. A histogram that covers all possible values which is used by gray scale is determined as a good histogram. A good histogram tends to have good contrast and the details of an image that can be easily observed. After the histogram equalization, the gray value of each image uniformly covers the entire gradation range, the image contrast is improved and the gray distribution of the different images becomes more unified as shown in Fig .6.8. We can conclude that the histogram equalization is effective in reducing the interference caused by different lighting conditions. There are many methods for the illumination normalization like, but due to simplicity and effectiveness of Histogram equalization we used it.

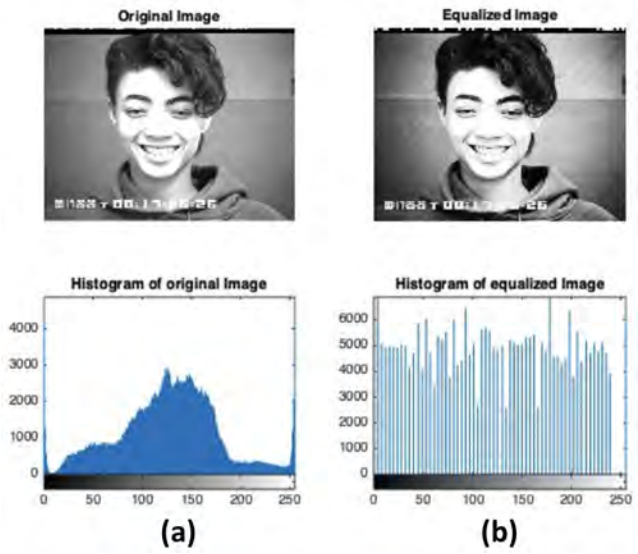

Figure 6.7: (a) Original Image with its histogram (b) Equalized image with its corresponding histogram

Some times no need to have histogram equalization step after TFP. However due to illumination condition the range of TFP images many not have full range (0-255). Hence we suggested the step of histogram equalization after TFP. Note that if TFP images of full range then no use of histogram equalization step. We carried out experiments with and without histogram equalization step. It has been observed that the results with histogram equalization step is slightly improved, not significantly.

Data Augumentation

CNN needs massive sets of data to be able to generalize to a given problem. However, publicly available FER databases do not have sufficient images to handle the problem. Simard et al. [158] suggested data augmentation procedure to extend the databases through the creation of synthetic face images for every original face image. Inspired by this procedure, the following operations had been utilized as data augmentation: 1) flipping image vertically and horizontally 2) Rotate each database image, rotate it at right angles if image is square and rotate it as 180^0 if image is rectangular 3) Add the random noise to the landmarks so as to introduce little deformations to faces as shows in Fig .6.8.

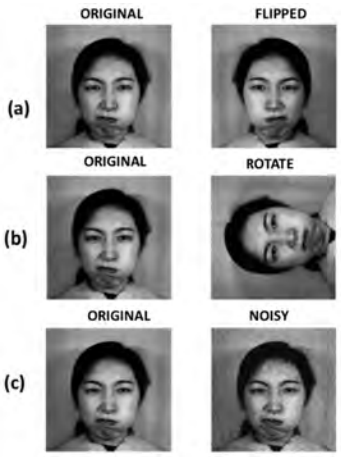

Figure 6.8: Original Image with (a) Flipped Image (b) Rotate Image (c) Noisy Image

So the resulting face image is different from the non-processed face (original face) the ones that used with CNN for the pre-train. These difference between processed and non processed data could affect the results, It may be because of

network learned the features from the original face image, and it may not be able to extract features from the processed face images. So, we additionally provide results with a network trained with original face images.

6.2.2 Feature Extraction From Gray Scale Facial Images

Absence of adequate training samples constrains the execution of CNN-based Facial expression recognition approach.

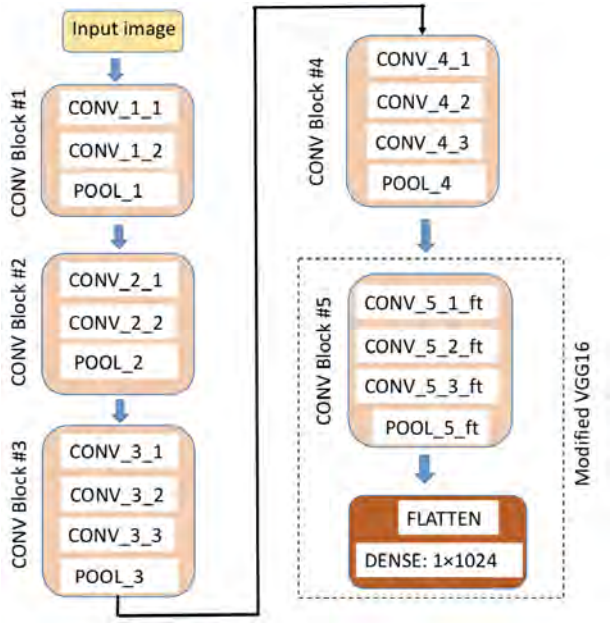

Figure 6.9: Framework for the modified VGGFace_ft network, used for extraction of the expression features from the given face images

Data augmentation can partially handle the issue exposure of over-fitting. Hence, fine-tuning is utilized to the extraction of facial expression related features from the given input face image by way of referring to the deep neural network (DNN) that attained excessive success in similar kind of tasks.

Our proposed framework utilizes DNN for feature extraction for FER, which relies on VGGFace network [123]. VGGFace is chosen as the basic model, which has the same architecture of VGG16 [160] and has been trained on large-scale face dataset. Due to the similarity between face recognition and facial expression recognition, the transfer learning of features is facilitated via fine-tuning VGGFace (VGGFace_ft). In VGGFace, the first few layers capture universal features

like blobs and edges, which are also relevant to other similar tasks. It has 138 million parameters and 13 convolutional layers, followed by 3 fully-connected layers (FCs). The initial two fully connected layers (FCs) have 4,096 outputs, and the last layer has 2,622 outputs. Thus, the parameters of the first four convolution blocks are frozen and the parameters of the fifth convolution block are fine-tuned. This allowed us to make use of the more general features that these layers were already trained to extract while performing fine-tuning on some of the layers that extract more specific features of facial expressions. Fig. 6.9 demonstrates the essential module of the framework. Compared with the original VGGFace, our VGGFace_ft (where "ft"means fine-tuning) is simplified by doing away with two dense layers. The dimension of the input data for forehead is 54×48, for eyes is 39×117 , for nose is 50×55 and for lips is 48×74.

At this point, we fix the structures of the initial four CONV (convolution) blocks of the VGGFace_ft. But we change the structure of fifth CONV block of VGGFace_ft and also change the names of each layer just by adding "ft"at the end of the original layer name. So now layer name of fifth CONV block is like CONV_5_1_ft. The parameters whose change the structure of the layer is shown in Table. 6.4. Based on experiments, the last dense layer preserved and set its dimension to 1×1024. That dimension is actually the extracted feature of input image denoted as feature vector "fv _1" for the forehead, "fv _2" for the eyes, "fv_3" for the nose and "fv_4" for the lips. We decline the learning rates of layers that have a place with the fifth CONV block by 10 times (learning rate for fifth CONV block is .001) of other block learning rate (.01 used for other CONV blocks) to ensure that that they'll learn more positive information. At last, the initial portion of the system is initialized with the VGGFace model weights, which is trained on the large scale face dataset. ReLu (Rectified Linear Unit) is applied after every convolutional layer.

Table 6.4: Parameters set for fifth block

	CONV_5_1_ft	CONV_5_2_ft	CONV_5_3_ft	POOL_5_ft
Filters	256	256	512	
size	7×7	3×3	3×3	2×2
stride	1	1	1	2
pad	3	0	0	0

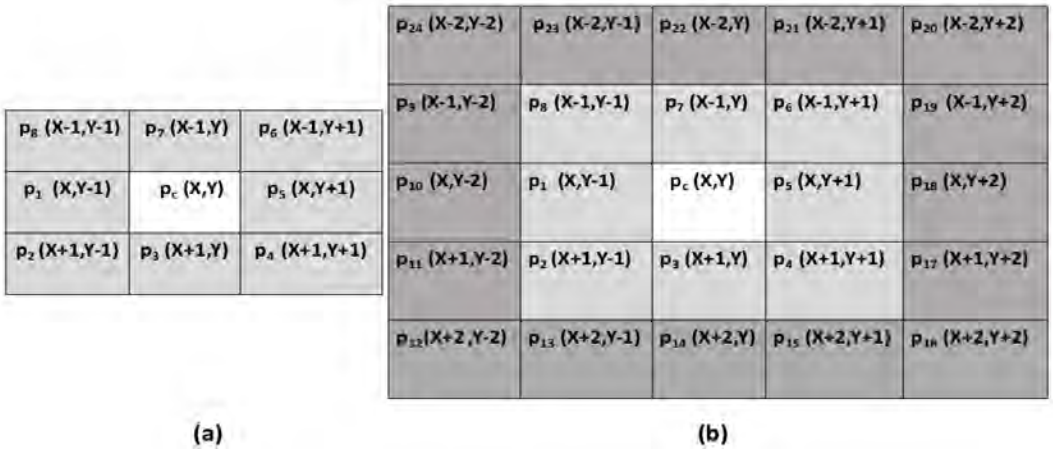

Figure 6.10: (a) 1^{st} order Pixel taylor feature $f_1(p_c)$ (Texture1 (T1) with 3×3 pixels) (b) 2^{nd} order Pixel taylor feature $f_2(p_c)$ (Texture2 (T2) with 5×5 pixels)

6.2.3 Feature Extraction From 2DTFP (Taylor Feature Pattern) Facial Images

To the best of our knowledge, there is no model trained on the TFP images. So here first compute the TFP facial images.

Calculating 2D Taylor Feature Pattern Facial Images

Here, Ding Yuanyuan [30] used one dimensional Taylor expansion for facial expression recognition. Induced by this, we are trying to implement the two dimensional (2D) Taylor expansion as full detail about this is already discussed in Chapter .3. Here we are only discussed only some required steps.

2D Taylor pixel feature extraction Based on the one dimensional (1D)Taylor expansion, we describe the Two Dimensional Taylor expansion for the facial expression recognition. Let $f_1(p_c)$ be the 1^{st} order 2D Taylor pixel feature of the central

pixel p_c . According to the one dimensional Taylor expansion, Two dimensional $f_1(p_c)$ can be approximately defined as:

$$f_1(p_c) \approx f(\phi, \psi) + [(p_c - \phi)\frac{\partial f}{\partial x} + (p_c - \psi)\frac{\partial f}{\partial y}] \tag{6.4}$$

where

$$\phi = \frac{\frac{1}{\sqrt{2}}p_8 + p_7 + \frac{1}{\sqrt{2}}p_6 + \frac{1}{\sqrt{2}}p_4 + p_3 + \frac{1}{\sqrt{2}}p_2}{\frac{4}{\sqrt{2}} + 2}$$

These p_8, p_7, p_6, p_4, p_3, p_2 are the pixels in the first layer (marked in grey) of TU1 in x direction. As shown in Fig. 6.10(a) .

$$\psi = \frac{\frac{1}{\sqrt{2}}p_8 + p_1 + \frac{1}{\sqrt{2}}p_2 + \frac{1}{\sqrt{2}}p_4 + p_5 + \frac{1}{\sqrt{2}}p_6}{\frac{4}{\sqrt{2}} + 2}$$

These p_8, p_1, p_2, p_4, p_5, p_6 are the pixels in the first layer (marked in grey) of TU1 in y direction. As appeared in Fig. 6.10(a).

$$f(\phi, \psi) = \frac{\phi + \psi}{2}$$

$$\frac{\partial f}{\partial x} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \phi \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \phi < 0 \end{cases}$$

$$\frac{\partial f}{\partial y} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \psi \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \psi < 0 \end{cases}$$

2^{nd}order taylor pixel feature $f_2(p_c)$ can be expressed as:

$$
\begin{aligned}
f_2(p_c) \;\approx\; & f(\phi,\psi) + [(p_c-\phi)\frac{\partial f}{\partial x} + (p_c-\psi)\frac{\partial f}{\partial y}] \\
& + \frac{1}{2}[(p_c-\phi)^2\frac{\partial^2 f}{\partial x^2} + 2(p_c-\phi)(p_c-\psi)\frac{\partial^2 f}{\partial x \partial y} \\
& + (p_c-\psi)^2\frac{\partial^2 f}{\partial y^2}]
\end{aligned} \tag{6.5}
$$

In the computation of ϕ and ψ there are some term like ϕ_1, ψ_1, ϕ_2, ψ_2 used. These terms defined as

$$
\phi_1 = \frac{\frac{1}{\sqrt{2}}p_8 + p_7 + \frac{1}{\sqrt{2}}p_6 + \frac{1}{\sqrt{2}}p_4 + p_3 + \frac{1}{\sqrt{2}}p_2}{\frac{4}{\sqrt{2}} + 2}.
$$

These p_8, p_7, p_6, p_4, p_3, p_2 are the pixels in the first layer (marked in grey) of TU2 in x direction. As shown in Fig. 6.10(b).

$$
\psi_1 = \frac{\frac{1}{\sqrt{2}}p_8 + p_1 + \frac{1}{\sqrt{2}}p_2 + \frac{1}{\sqrt{2}}p_4 + p_5 + \frac{1}{\sqrt{2}}p_6}{\frac{4}{\sqrt{2}} + 2}
$$

These p_8, p_1, p_2, p_4, p_5, p_6 are the pixels in the first layer (marked in grey) of TU2 in y direction. As appeared in Fig. 6.10(b).

$$
\phi_2 = \frac{p_{24} + p_{23} + p_{22} + p_{21} + p_{20} + p_{16} + p_{15} + p_{14} + p_{13} + p_{12}}{10}
$$

Here, ϕ_2 is the mean of the gray value in the second layer (pixels are marked in dark grey and furthermore which are in x direction). Pixels $p_{24}, p_{23}, p_{22}, p_{21}, p_{20}, p_{16}, p_{15}, p_{14}, p_{13}, p_{12}$ will came here.

$$
\psi_2 = \frac{p_{24} + p_9 + p_{10} + p_{11} + p_{12} + p_{16} + p_{17} + p_{18} + p_{19} + p_{20}}{10}
$$

Here, ψ_2 is the mean of the gray value in the second layer (pixels are marked in dark grey and also which are in y direction). Pixels $p_{24}, p_9, p_{10}, p_{11}, p_{12}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$ will came here. Finally compute the ϕ and ψ like

$$
\phi = \frac{\phi_1 + \phi_2}{2} \quad and \quad \psi = \frac{\psi_1 + \psi_2}{2}
$$

Then, compute

$$f(\phi, \psi) = \frac{\phi + \psi}{2}$$

Here, evaluate the derivative of x and y like

$$\frac{\partial f}{\partial x} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \phi_1 \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \phi_1 < 0 \end{cases}$$

$$\frac{\partial f}{\partial y} = \begin{cases} \frac{1}{2}; & \text{if } p_c - \psi_1 \geq 0 \\ -\frac{1}{2}; & \text{if } p_c - \psi_1 < 0 \end{cases}$$

$$\frac{\partial^2 f}{\partial x^2} = \begin{cases} \frac{1}{4}; & \text{if } (p_c - \phi_1)\,(p_c - \phi_2) \geq 0 \\ -\frac{1}{4}; & \text{if } (p_c - \phi_1)\,(p_c - \phi_2) < 0 \end{cases}$$

$$\frac{\partial^2 f}{\partial y^2} = \begin{cases} \frac{1}{4}; & \text{if } (p_c - \psi_1)\,(p_c - \psi_2) \geq 0 \\ -\frac{1}{4}; & \text{if } (p_c - \psi_1)\,(p_c - \psi_2) < 0 \end{cases}$$

$$\frac{\partial^2 f}{\partial x \partial y} = \begin{cases} \frac{1}{4}; & \text{if } (p_c - \phi_2)\,(p_c - \psi_2) \geq 0 \\ -\frac{1}{4}; & \text{if } (p_c - \phi_2)\,(p_c - \psi_2) < 0 \end{cases}$$

2D Taylor Feature Pattern (2DTFP) Taylor Feature Pattern of $f_n(p_c)$ is expressed as:

$$2DTFP = \sum_{j=1}^{8} S(f_n(p_c), f_n(p_j).2^{j-1} \tag{6.6}$$

$$S(p_c, p_j) = \begin{cases} 1; & \text{if } p_c \geq p_j \\ 0; & \text{if } p_c < p_j \end{cases}$$

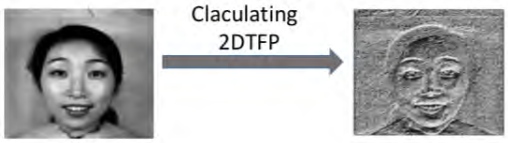

Figure 6.11: Illustration of calculating facial 2DTFP image.

Fig. 6.11 demonstrates a 2DTFP facial image of a given facial image. Expressions associated facial regions, such as eyes, nose, lips are remarkable in 2DTFP images to compare to grayscale images.

According to our knowledge of so far, no existing model is trained on the TFP images. So, we construct two layer CNN model that automatically extracts the features from the 2DTFP facial images.

Fig. 6.12 illustrates the proposed CNN structure, which consists input layer, two convolution layers C_1 and C_2 and two sub-sampling layers S_1 and S_2. 64 filters utilized in the first convolution layer C_1 for the input 2DTFP facial images, which target the exhaustive information of the facial expression. This layer uses a convolution kernel of 7×7 and outputs is 64 images. This layer is accompanied via a sub-sampling layer S_1, which uses max pooling with kernel size 2×2. That sub-sampling layer reduces the image to half its size. Next convolution layer C_2 performs 256 filters with 3×3 kernel to map the preceding layer and is followed through some other S_2 sub-sampling layers with a 2×2 kernel. All parameters utilized in the proposed CNN are listed in Table 6.5.

At this point, the output is given to the fully connected layer (Dense layer) with the 1024 neurons. From here extract the feature vector (fv_2) of size 1×1024. To handle the nonlinear data, add the "Relu" activations after the S_1 and S_2 layers. Data augmentation is employed to increase the number of 2DTFP facial images synthetically. In this manner, over-fitting can be taken care of by utilizing the "dropout" operation [162] sub-sampling layer and fully connected layer.

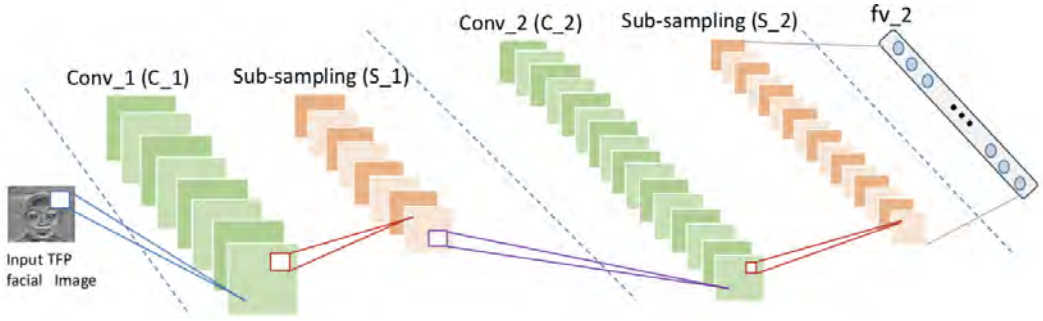

Figure 6.12: Framework for the proposed CNN used for extraction of the expression features from the TFP facial images

Table 6.5: parameter set for the proposed CNN

	C_1	S_1	C_2	S_2
Filters	64		256	
size	7×7	2×2	3×3	2×2
stride	1	2	1	2
pad	3	0	0	0

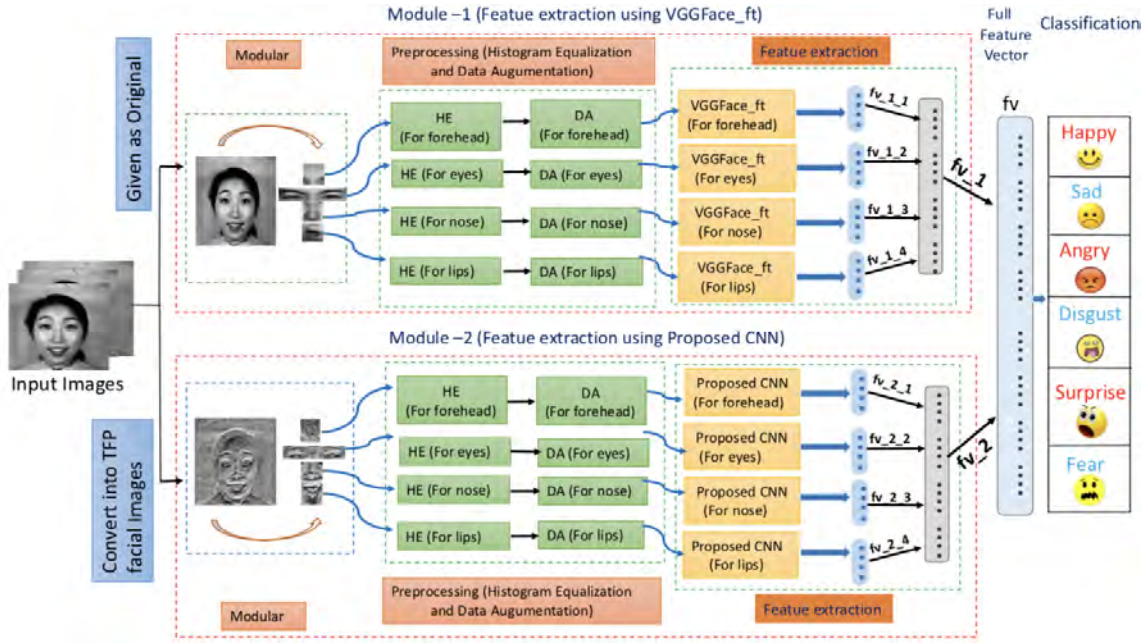

Figure 6.13: Illustration of the proposed Framework

6.2.4 Concatenation of Different Outputs and Classification

Fig. 6.13 shows our proposed framework. Expression features fv_1 is a concatenation of the feature vector coming from the forehead (fv_1_1), eyes (fv_1_2), nose (fv_1_3) and lips (fv_1_4). It is the features that came from the grayscale images using VGGFace_ft with fine-tuning technique. Similarly feature vector features fv_2 is a concatenation of the feature vector coming from the forehead (fv_2_1), eyes (fv_2_2), nose (fv_2_3) and lips (fv_2_4). These fv_2 features coming from the 2DTFP facial images using proposed CNN architecture. Finally, we get full feature vector "fv" that is the combination of the fv_1 and fv_2. Will go in Next step for classification.

In the classification process, the similarity between extracted features of the display setting and the probe set is evaluated by the SVM and K nearest-neighbor

(K=1,2,3) classifier with various distance measures.

The Support Vector Machine (SVM) algorithm is applied to classification. When all local image descriptors are transformed to a fixed length feature vector, distance is computed to measure the similarity between each pair of the feature vectors. Finally, each image for the test is classified into an object class with the maximum SVM output decision value. We tune the parameters of the classifier on the training set, and obtain the recognition accuracy on the test set.

Other classifier is K nearest-neighbor (K=1,2,3) classifier with various distance measures. Euclidean distance, Chi-square distance, as well as histogram intersection (HI) are utilized in our experiments. Which are defined as in Eq. 6.7, 6.8 and 6.9

$$d(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \tag{6.7}$$

$$\chi^2 = \sum_{i,j}\frac{(x_{i,j} - y_{i,j})^2}{(x_{i,j} - y_{i,j})} \tag{6.8}$$

$$D_{HI}(x,y) = -\sum_{i,j} min(x_{i,j}, y_{i,j}) \tag{6.9}$$

SVM classifier predict the class which has highest value directly. On the other hand softmax layer predict the probability of classes. One can predict the class which has highest probability. Softmax is an integral part of DNN used for class prediction. We do not have control over this prediction. In case the input data is very complex in nature the softmax which is part of DNN might not predict the class label accurately. Instead researchers many times prefer a strong classifier to be used in place of softmax. Note that we have also used a simpler classifier such as KNN in place of SVM. Hence another motivation of using sophisticated classifier such as SVM and simpler classifier such as KNN is to show suitability of the features coming out the convolution layer of DNN.

For computation of the loss, we used the MSE (mean square error) till now it

is best for the SVM and KNN classification, Which is defined as

$$Loss = \frac{1}{M}\sum_{1}^{M} \| \ Y_i - Y_i^{'} \ \|^2 \tag{6.10}$$

Where M is the total numbers of input images, Y and Y' the true and predicted outputs, respectively.

6.2.5 Experiment Results and Analysis

To See the efficiency of our method, our FER methodology work on the Keras framework on the macOS Mojave system platform. To do the correct and effective evaluations, 4 benchmarks datasets which we also have been previously used, which are made out of facial images. For the JAFFE dataset, Fig .6.14 the results of one person expression of Jaffe data after the first convolution layer in VGGFace_ft. Whereas Fig .6.15 shows the trends of accuracy and loss during the training and testing with the increase in iterations. Whereas Fig. 6.16 shown the comparison of accuracy with different architectures. Table 6.6 shows reported the average accuracy in the Holistic as well as modular approach both. In the JAFFE dataset average accuracy is 97.16% and the best is reported as 99.35%. Table 6.17 shows the accuracies individual expression for the JAFFE dataset. "Neutral" accuracy is highest as "1.00". Other expressions accuracies is around the ".96" . The recognition is not perfectly done due to the fact that the "JAFFE" dataset are hard to distinguish even by manually.

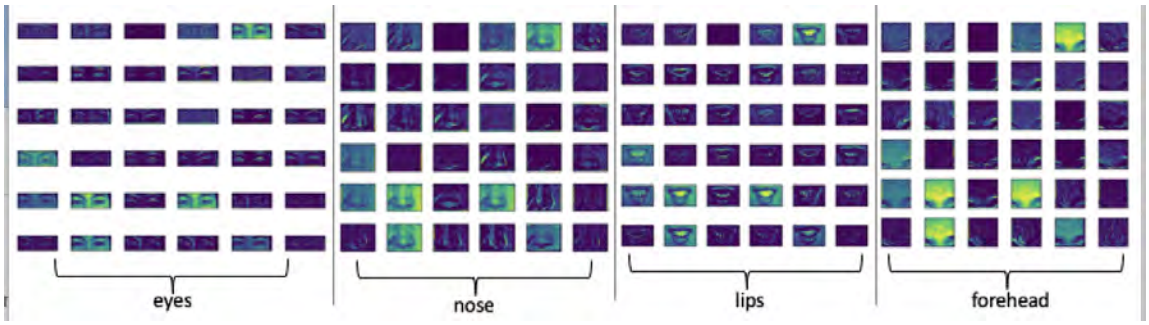

Figure 6.14: Results of 36 filters out of 60 after the 1^{st} Conv2D layer in VGGFace_ft to the given modular input jaffe image

Similarly, for VIDEO dataset Fig. 6.18 shows the trends of accuracy and loss

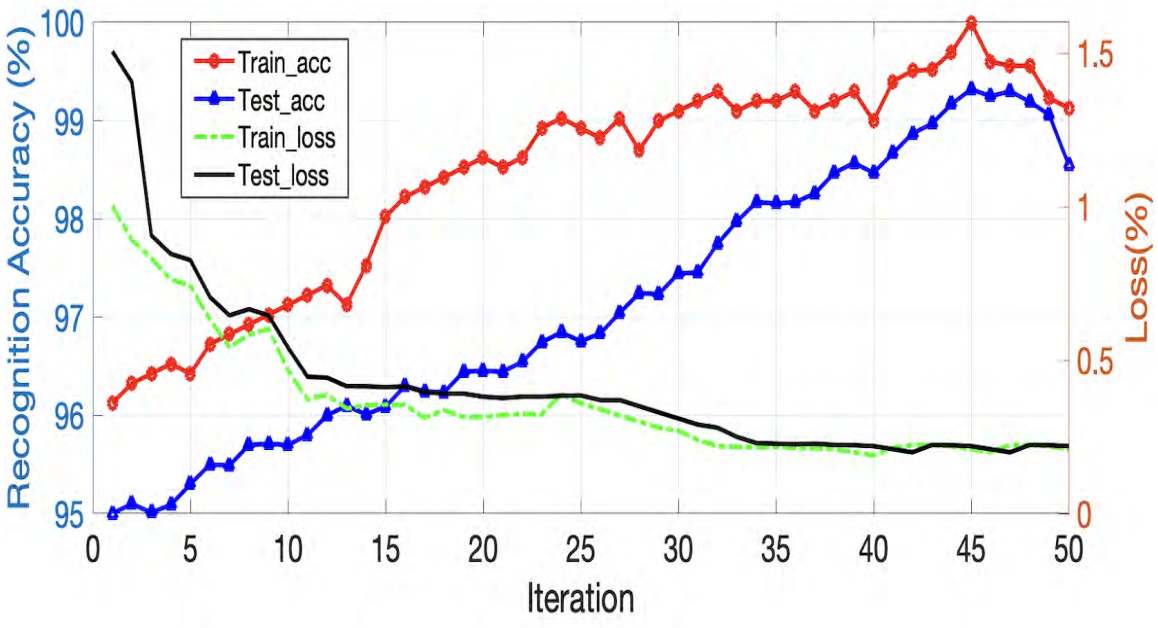

Figure 6.15: Curves of Accuracy and Loss during training and testing phases for jaffe dataset

during the training and testing with the increase in iterations, whereas Fig. 6.19 shown the comparison of accuracy with different architectures. Table 6.7 shows reported the average accuracy in the Holistic as well as modular approach both. In the VIDEO dataset average accuracy is 98.07%, and the best is reported as 99.70%. Confusion matrix for VIDEO dataset shown in Fig .6.20.

Fig. 6.21 shows the trends of accuracy and loss during the training and testing

Table 6.6: Comparison between the Holistic and Modular approach in our proposed framework in the light of the classifiers for JAFFE dataset (In terms of average accuracy (%) reported for 50 iterations)

	JAFFE							
	HOLISTIC				**MODULAR**			
Architectures	**SVM**	**KNN**			**SVM**	**KNN**		
		Eucli-dean	Chi -Square	HI		Euclidean dean	Chi- -Square	HI
VGGFace	66.90	64.38	59.14	58.45	70.79	66.84	65.16	65.14
VGGFace_ft	88.90	86.34	79.23	72.12	93.98	90.25	88.47	79.60
VGGFace_ft+HE+DA (Module 1)	89.01	90.56	81.03	78.34	94.89	93.17	90.36	86.01
ProposedCNN+HE+DA (Module 2)	78.56	91.25	74.24	76.67	92.97	88.48	85.12	79.95
Ours (Module1+Module2)	93.45	91.78	80.35	78.89	97.12	94.27	92.16	88.99

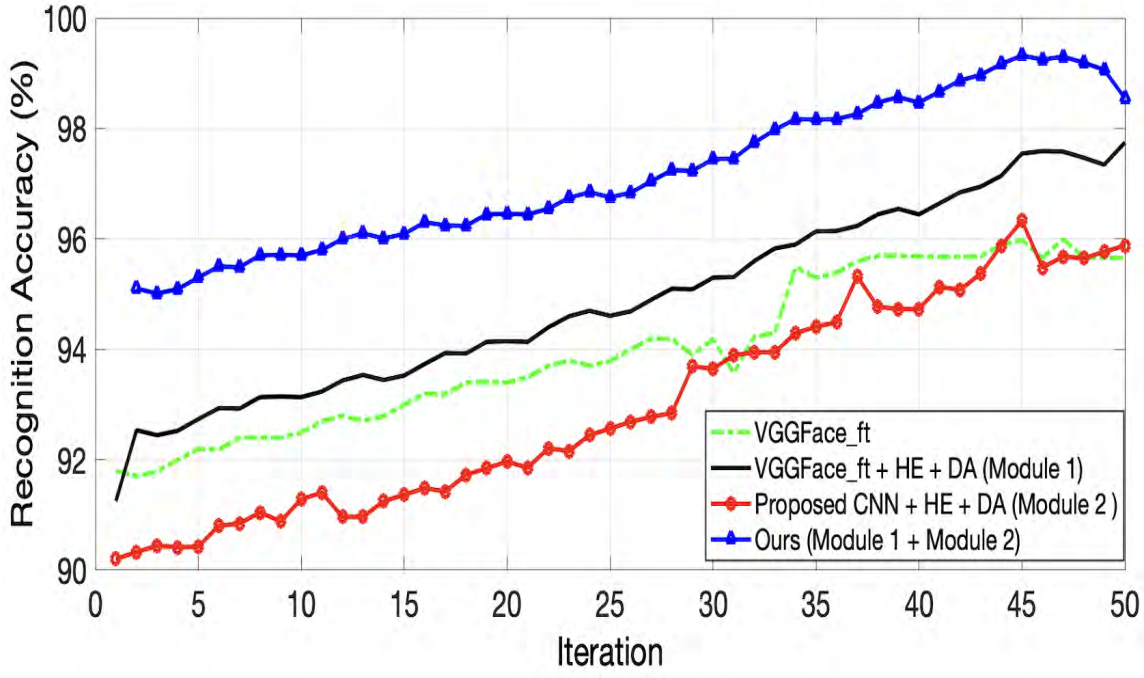

Figure 6.16: compare the ability of each architecture for Jaffe dataset

Label	Happy	Disgust	Angry	Neutral	Fear	Sad	Surprise
Happy	00.97	0.00	0.00	0.01	0.00	0.00	0.02
Disgust	0.01	0.96	0.00	0.00	0.02	0.01	0.00
Angry	0.00	0.00	0.98	0.00	0.02	0.00	0.00
Neutral	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Fear	0.00	0.01	0.01	0.00	0.96	0.02	0.00
Sad	0.00	0.00	0.01	0.00	0.02	0.96	0.01
Surprise	0.01	0.00	0.00	0.011	0.01	0.00	0.9689

Figure 6.17: Expression wise recognition accuracies for the Jaffe dataset

with the increase in iterations. Whereas Fig. 6.22 shown the comparison of accuracy with different architectures. Table 6.8 shows reported the average accuracy in the Holistic as well as modular approach both. In the CK+ dataset, average accuracy is 96.76%, and the best is reported as 98.5%. Confusion matrix for CK+ dataset shown in Fig .6.23.

For the Oulu-Casia dataset, Fig. 6.24 shows the trends of accuracy and loss during the training and testing with the increase in iterations, Fig. 6.26 shown the comparison of accuracy with different architectures. Table 6.9 shows reported the average accuracy in the Holistic as well as modular approach both. In the CK+

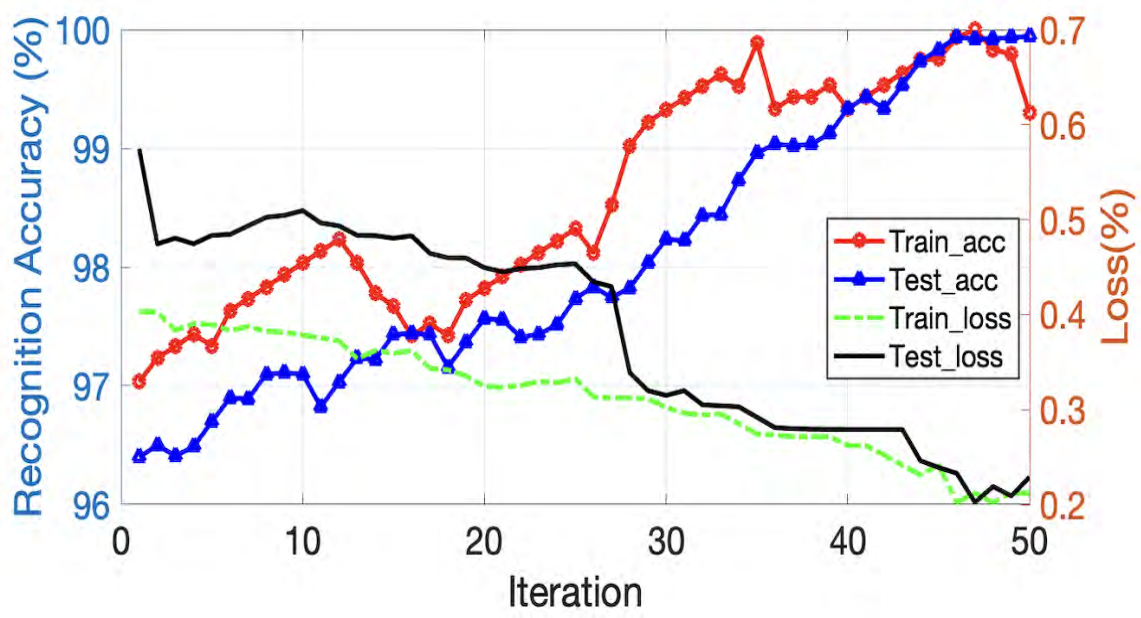

Figure 6.18: Curves of Accuracy and Loss during training and testing phases for VIDEO dataset

Table 6.7: Comparison between the Holistic and Modular approach in our proposed framework in the light of the classifiers for VIDEO dataset (In terms of average accuracy (%) reported for 50 iterations)

	VIDEO							
Architectures	**HOLISTIC**				**MODULAR**			
	SVM	**KNN**			**SVM**	**KNN**		
		Euclidean	Chi-Square	HI		Euclidean	Chi-Square	HI
VGGFace	69.90	66.74	66.28	62.32	75.02	67.67	55.89	52.87
VGGFace_ft	91.23	85.90	83.12	81.43	95.42	91.43	87.54	84.56
VGGFace_ft+HE+DA (Module 1)	91.89	89.12	85.28	83.72	96.30	92.89	90.43	87.38
ProposedCNN_ft+HE+DA (Module 2)	87.36	82.17	79.13	75.26	93.78	88.86	86.49	80.15
Ours (Module1+Module2)	93.87	90.19	85.45	83.47	98.07	94.42	91.39	90.47

dataset, average accuracy is 96.3%, and the best is reported as 97.9%. Confusion matrix for CK+ dataset shown in Fig .6.23.

Expressions at different intensity rate

There are many different models of the nature of expression and how it is characterized in the brain and in the body. The work lies in determining the different

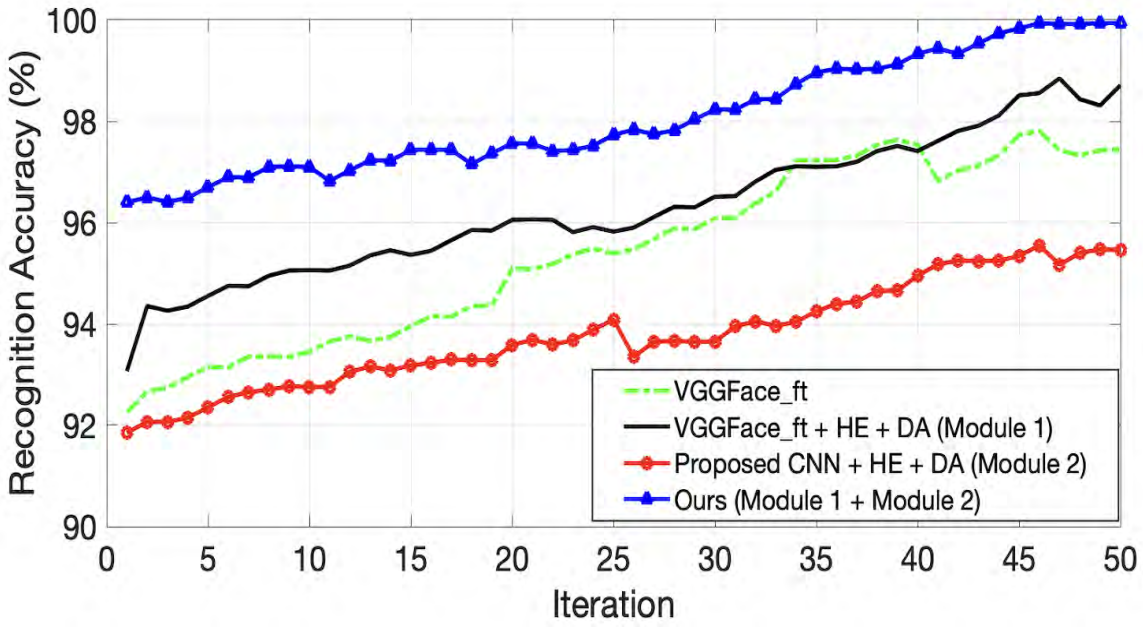

Figure 6.19: Compare the ability of each architecture for VIDEO dataset

Label	Angry	Normal	Smile	Open-Mouth
Angry	0.9626	0.0226	0.00	0.0174
Normal	0.00	0.99	0.01	0.00
Smile	0.0035	0.01	0.9765	0.01
Open-Mouth	0.007	0.00	0.02	0.9730

Figure 6.20: Expression wise recognition accuracies for the VIDEO dataset

Table 6.8: Comparison between the Holistic and Modular approach in our proposed framework in the light of the classifiers for CK+ dataset (In terms of average accuracy (%) reported for 50 iterations)

	CK+							
	HOLISTIC				MODULAR			
Architectures	**SVM**	**KNN**			**SVM**	**KNN**		
		Eucli-dean	Chi -Square	HI		Euclidean	Chi- -Square	HI
VGGFace	80.19	78.16	71.08	68.8	71.07	82.42	80.14	75.67
VGGFace_ft	90.99	87.27	82.16	80.04	93.84	88.17	89.14	84.72
VGGFace_ft+HE+DA (Module 1)	92.74	90.54	84.00	83.14	94.38	92.15	91.19	89.12
ProposedCNN_ft+HE+DA (Module 2)	89.72	85.17	80.14	79.54	92.78	89.21	87.24	81.12
Ours (Module1+Module2)	94.75	91.47	89.24	87.50	96.76	95.12	93.17	91.17

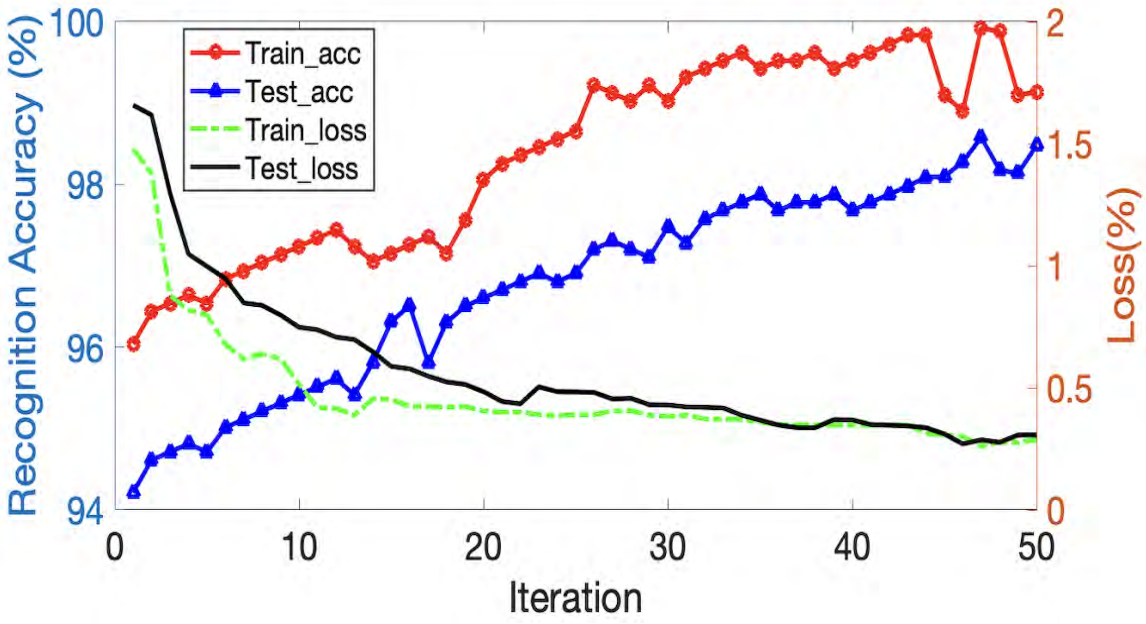

Figure 6.21: Curves of Accuracy and Loss during training and testing phases for CK+ dataset

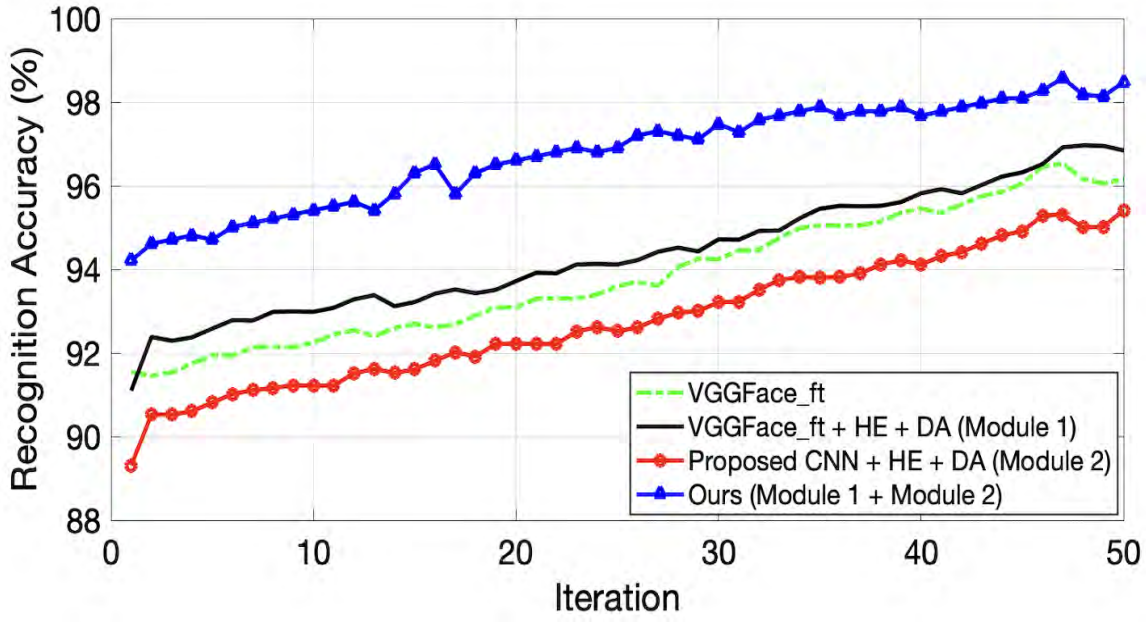

Figure 6.22: Compare the ability of each architecture for CK+ dataset

degree of expressions. Through this proposed method, we are not only finding the dominant expression, but also the percentages of all the expressions presented on the face. Here we analyze the degree of expression by our proposed approach while moving from one stage of expression to the next higher state. There are some other expressions that are affected by changes within a time interval, which

Label	Happy	Disgust	Angry	Fear	Sad	Surprise
Happy	00.9824	0.00	0.00	0.00	0.00	0.0176
Disgust	0.00	0.9476	0.024	0.02	0.0084	0.02
Angry	0.00	0.0298	0.9602	0.01	0.00	0.00
Fear	0.00	0.01	0.01	0.9649	0.0151	0.00
Sad	0.00	0.01	0.01	0.008	0.9620	0.01
Surprise	0.02	0.00	0.0074	0.00	0.00	0.9726

Figure 6.23: Expression wise recognition accuracies for the CK+ dataset

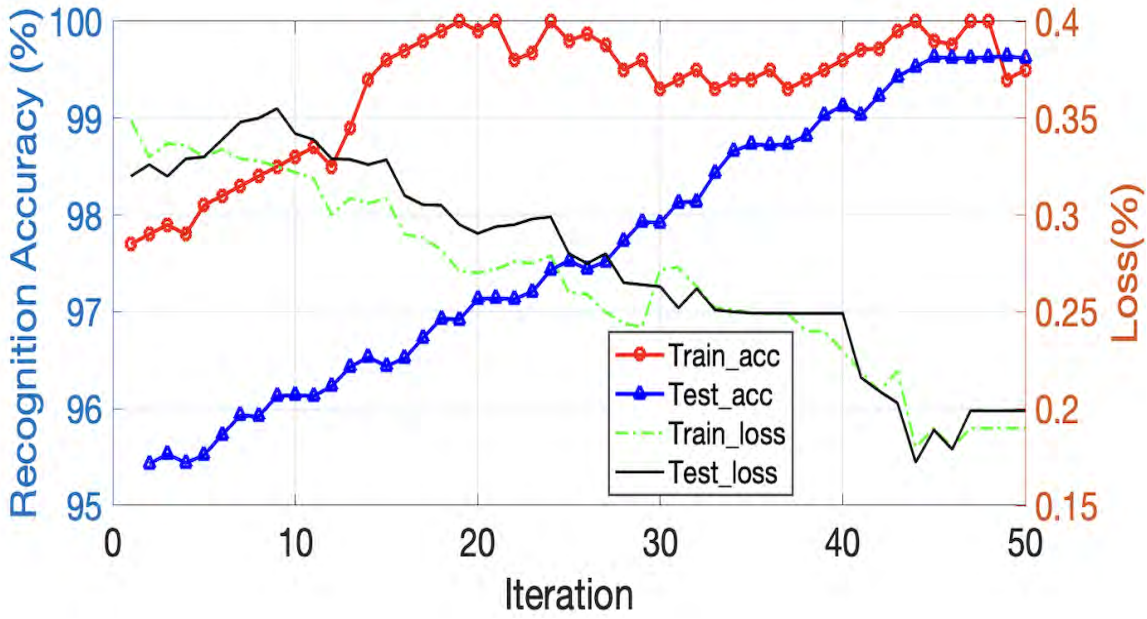

Figure 6.24: Curves of Accuracy and Loss during training and testing phases for Oulu-Casia dataset

Label	Happy	Disgust	Angry	Fear	Sad	Surprise
Happy	00.9887	0.00	0.00	0.00	0.00	0.0113
Disgust	0.00	0.9515	0.014	0.02	0.0045	0.01
Angry	0.00	0.00	0.9947	0.0053	0.00	0.00
Fear	0.00	0.01	0.0030	0.9798	0.0072	0.00
Sad	0.00	0.02	0.0090	0.008	0.9567	0.0063
Surprise	0.01	0.00	0.0011	0.00	0.00	0.9889

Figure 6.25: Expression wise recognition accuracies for the Oulu-Casia dataset

are clearly shown in the graphs below in Fig. 6.27, 6.28, 6.29 and 6.30. For some basic expressions, percentage variations with different time intervals are also represented. Our approach is very useful for exploring micro expressions. In a par-

Table 6.9: Comparison between the Holistic and Modular approach in our proposed framework in the light of the classifiers for Oulu-Casia dataset (In terms of average accuracy (%) reported for 50 iterations)

OULU-CASIA								
Architectures	**HOLISTIC**				**MODULAR**			
	SVM	**KNN**			**SVM**	**KNN**		
		Euclidean	Chi-Square	HI		Euclidean	Chi-Square	HI
VGGFace	79.25	75.13	70.12	68.18	81.10	79.00	76.69	76.10
VGGFace_ft	89.79	88.20	86.17	84.01	95.16	90.79	86.97	89.30
VGGFace_ft+HE+DA (Module 1)	91.30	90.12	88.21	87.27	96.058	91.48	91.90	91.00
ProposedCNN+HE+DA (Module 2)	85.10	86.35	83.19	79.12	94.09	88.00	83.07	80.30
Ours (Module1+Module2)	93.10	92.17	90.92	89.42	97.67	94.72	92.10	90.56

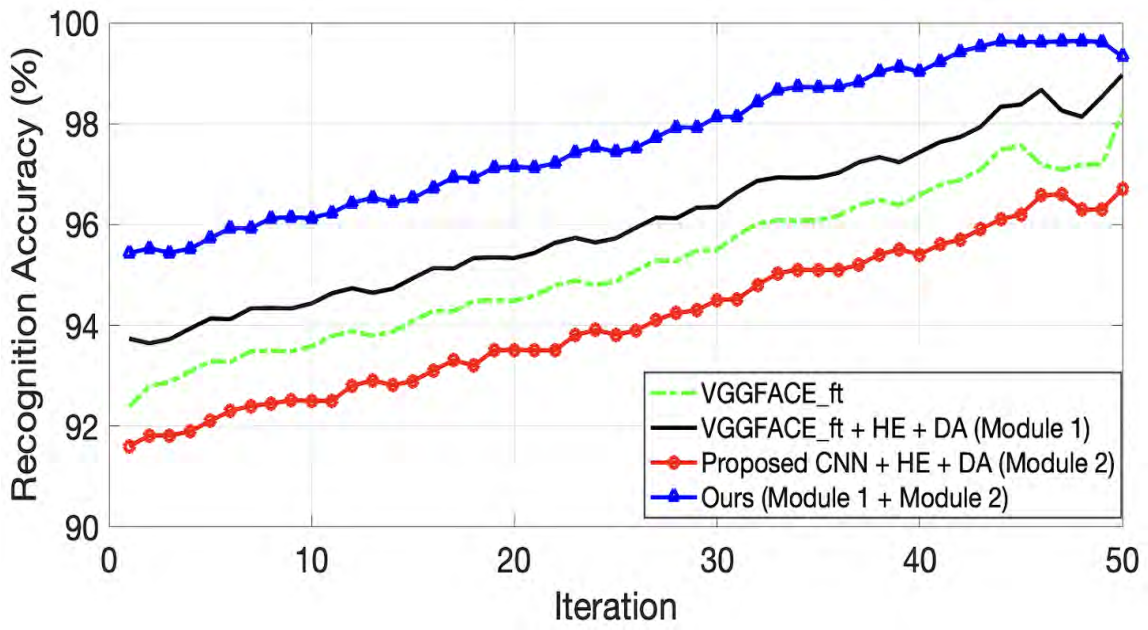

Figure 6.26: Compare the ability of each architecture for Oulu-Casia dataset

ticular face, for the series of the same expression, the highest level of expressions gives us more attention. Therefore, among the series of the same expressions, the highest degree of expression can be considered the ground-truth.

Qualitative Evaluations

To assess the qualitative execution of the proposed framework, facial images have gathered from the Internet for evaluation. In each facial image, the detected four region eyes, nose, lips, and forehead is represented by red, green, orange, and

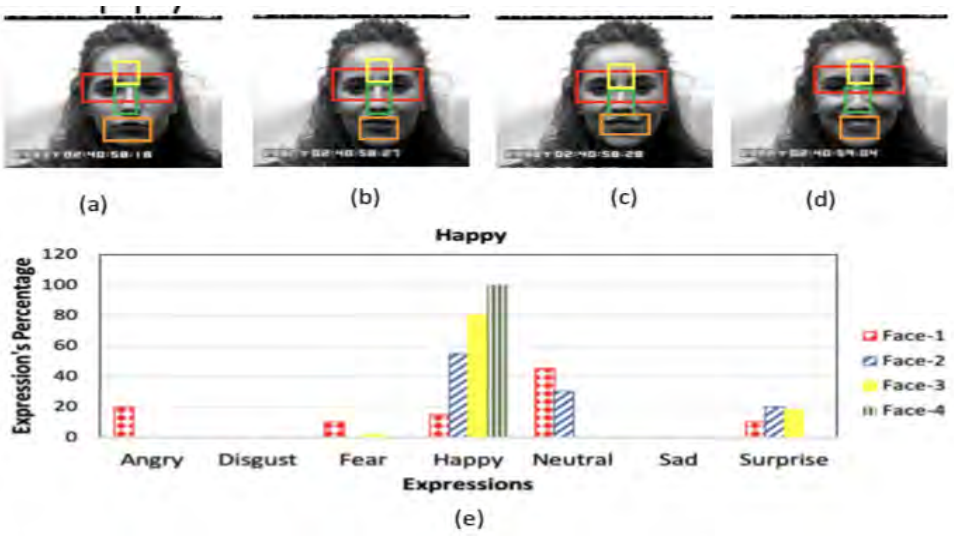

Figure 6.27: (a) (b) (c) (d) Shows the expression of the Happy face from the lower level to the extreme level, and (e)Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

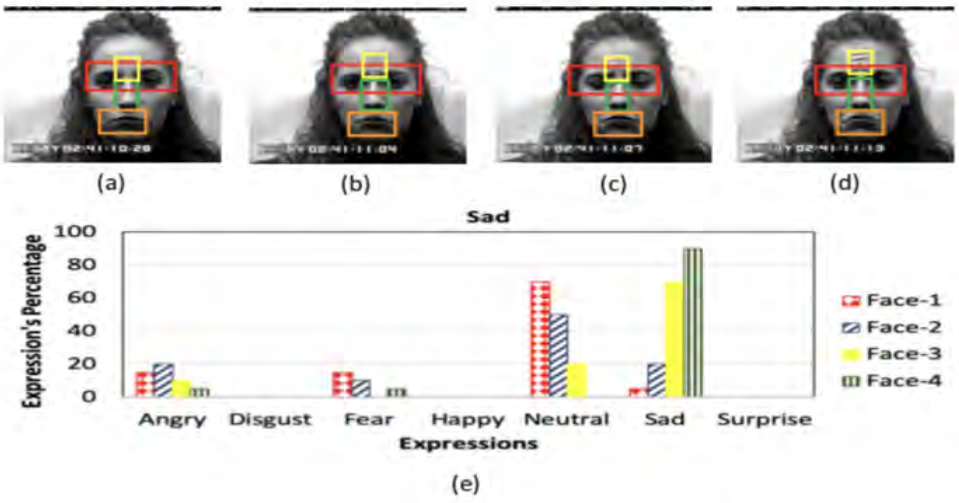

Figure 6.28: (a) (b) (c) (d) Shows the expression of the Sad face from the lower level to the extreme level, and (e)Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

yellow rectangle. The recognition accuracy is shown on the top right of the facial image. Red characters show the ground truth of the given facial image, whereas the green characters show the recognized expression using the proposed framework with recognition accuracy. Fig. 6.31 has interpreted the successful expression recognition for some Internet images. By seeing Fig. 6.31 expressions, such as

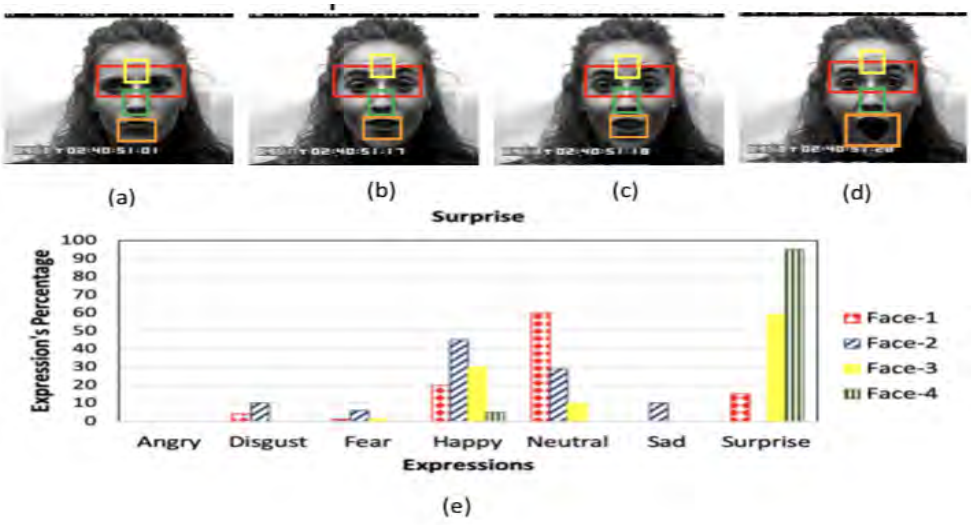

Figure 6.29: (a) (b) (c) (d)Shows the expression of the Surprise face from the lower level to the extreme level, and (e)Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

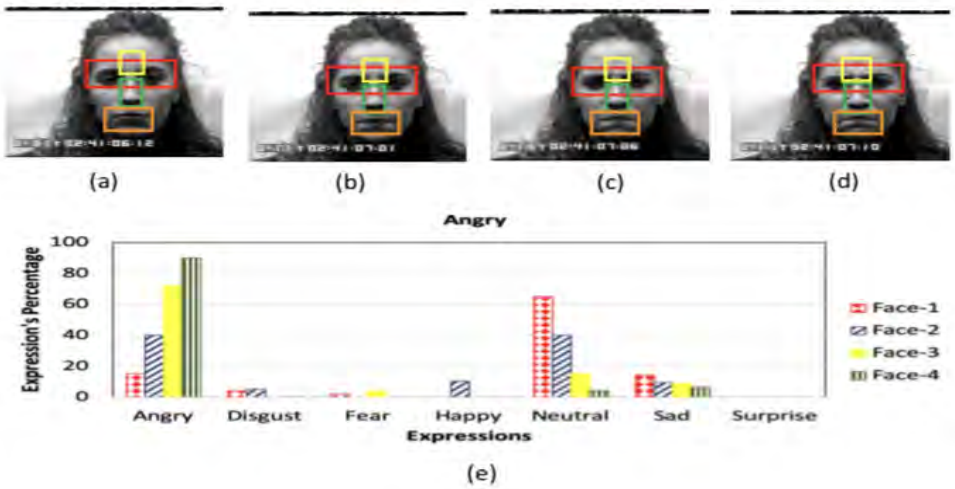

Figure 6.30: (a) (b) (c) (d) Shows the expression of the angry face from the lower level to the extreme level, and (e)Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

"Happy", "Surprise", "Disgust", "Sad ", and "Angry "are smooth to recognize in the frontal face images. This result somehow similar to the result illustrated by the confusion matrix Fig. 6.17, 6.20, 6.23, 6.25. Where as Fig. 6.32 failed recognition of expression. Sometimes, our framework not easily recognize the facial regions due to the poor illuminations, occlusions or some deviation from the frontal face.

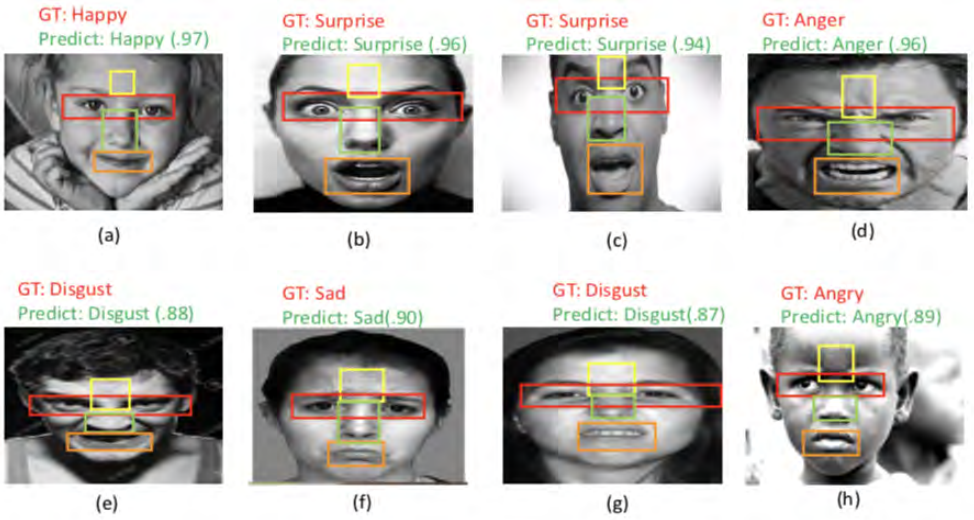

Figure 6.31: Successful recognition of Internet facial expression images

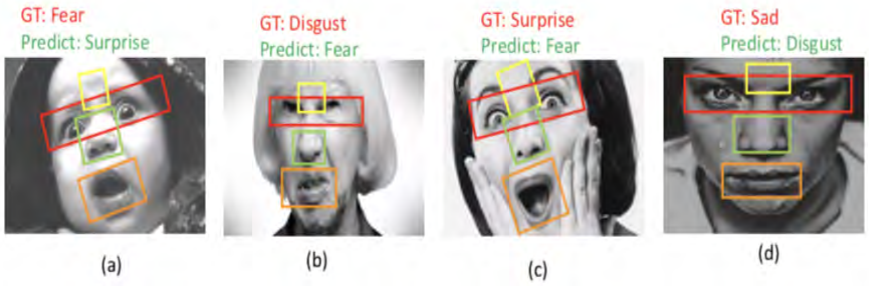

Figure 6.32: Unsuccessful recognition of Internet facial expression images

6.2.6 Conclusion

In this chapter we implement two versions of the Deep Neural Network (DNN) techniques. In the first technique, Deep neural network VGG16_ft, is proposed to automatically extricate features from the given facial images. Fine-tuning is very fruitful to the FER (Facial Expression Recognition) with pre-trained models, if sufficient facial images are not collected. Two preprocessing approaches, Fourier transform followed by Gabor filters and Data Augmentation (DA), are implemented to restrain the regions used for Facial expression recognition (FER). The features from four facial regions are concatenated and classification is done using SVM and KNN (with different distance measure).

Whereas in the second technique, that is primarily based on double channel architecture that processes the grayscale facial image and TFP facial image at the

same time. Both image channels which are utilized are complementary, and it captures local and global information from the given grayscale and TFP facial image. It can enhance the recognition capacity. Concatenation strategy is proposed to completely use the features that are extracted from both image channels (VGGFace_ft and proposed CNN). VGGFace_ft has automatically extracted the features from the given grayscale face images. Fine-tuning is utilized for training the system with the initial parameters got from the Imagenet. A proposed CNN is built to automatically extracts the features from the TFP facial images because of the pre-trained model is not trained on TFP facial images. Furthermore, concatenated features have been classified using SVM and KNN classifier with different distance measures. Capability of our proposed method is to recognize a facial expressions using partial information from the given whole face image. The proposed method is applied to the most informative regions of the face, i.e., forehead, eyes, nose, and lips. It is observed that a combination of these regions is useful enough to distinguish facial expressions of different persons or the same persons in most of the cases. The evaluation did in three datasets (JAFFE, VIDEO and, CK+) to check the effectiveness of our framework by recognizing the basic expressions. This proposed framework is the combination of the two types of deep neural networks, so it is easily utilized the local and global information about the expressions. In Table 6.10 we compared all our proposed methods.

Table 6.10: Compare all proposed methods

DataBase	**2DTFP** [chapter 3]	**HSOG** [chapter 4]	**E-PCA** [chapter 5] [Section 5.1]	**CS-ONPP** [chapter 5] [Section 5.3]	**DNN-FG** [chapter 6] [Section 6.1]	**2DNN**
JAFFE	92.78	94.15	89.01	94.54	95.96	97.12
VIDEO	94.86	95.98	96.05	95.67	96.67	98.07
CK+	93.78	93.89	91.72	87.88	96.78	96.76
Oulu-CASIA	96.87	97.00	95.32	93.15	96.08	98.12

CHAPTER 7

Conclusions and Future Research Directions

In this chapter, we provide the conclusions by summarizing our main contributions and also indicate future research directions.

7.1 Conclusions

For the last three decades, researcher are exploring various algorithms to recognize a face and facial expression recognition. The thesis contributed in the same direction. In particular, the thesis contributed to facial expression recognition using a set of benchmark datasets. Here experiments performed on the four benchmark databases, which are JAFFE, VIDEO, CK+, OULU-CASIA. With the increasing constraints in the sensors accruing face images, the increment in challenges of the face and facial expression recognition is explored. The modular approach presented here mimics the capability of the human to identify a person with a limited facial part. Facial parts like eyes, nose, lips, and forehead contribute more to the expression recognition task. This thesis we have addressed classical feature based approaches to deep learning techniques. This includes following works:

- Two Dimensional (2D) Taylor Expansion for feature extraction
- Histogram of second order gradient (HSOG) for feature extraction
- E-PCA (Euler Principle Component Analysis) for reducing feature length
- CS-ONPP: Class Similarity based Orthogonal Neighborhood Preserving Projection: A new approach to find neighbors to reduce the dimension of features.

- DNN based on Fourier transform followed by Gabor filtering
- Double channel based deep neural network

Starting from the conventional feature-based approach researchers in the current era are mostly relying on deep and convolutional neural network for the mentioned task. In the same line thesis contributed a couple of feature based approaches and the CNN (Convolution Neural Networks) approaches as well for FER. For conventional feature based approach, researchers in the past explored various techniques of feature extraction assuming the input signal is one dimensional. However, image is a two-dimensional signal, hence the thesis proposed a couple of two-dimensional feature extraction approach unlike existing one dimensional approaches for the FER task. In last decade it has been observed that dimensionality reduction for face and FER tasks attain maximum attention from the researchers. Following the same direction, thesis contributed an approach using the recent technique of dimensionality reduction for FER tasks. Each of the proposed method mentioned in the thesis is compared with the state of the art results. It has been observed that result of proposed methods is among the top-performing algorithms for the FER tasks.

In this thesis, we have proposed different frameworks for automatic recognition of facial expressions. We address the shortcomings of the previously proposed framework. Our proposed structures is discussed in general. We achieve results that exceed state-of-the-art methods for expression recognition. Secondly, they are computationally efficient and simple as they process only perceptually salient region(s) of face for feature extraction. By processing only perceptually salient region(s) of the face, reduction in feature vector dimensionality and reduction in computational time for feature extraction is achieved. We are thus making them suitable for real-time applications. Note that the processing time of extracting the salient region, either manually or automatically, is not considered while computing on the processing time of the facial expression recognition algorithms.

To handle the local illumination variation in the image (LL) Laplace-Logarithmic algorithm is used in the two-dimensional Taylor feature pattern (2DTFP). Most procedures just used the arrangement with global illumination varieties. They

thus yielded more unsatisfactory recognition performances within the case of natural illumination variations that are usually uncontrolled within the globe. Hence, to address the brightening variety issue, we at that point presented the (LL) Laplace-Logarithmic area in this article for further improving the exhibition. We applied the proposed 2D Taylor expansion theorem in the facial feature extraction phase and formulated the 2DTFP method.

HSOG is inspired by the human visual system, thus extract features only from perceptual salient regions Capability of Histogram of second-order gradient (HSOG) descriptor to recognize a person using partial information from the whole face image. Work proposed the local image descriptor that extracts the histogram of second order gradients (HSOG), which capture the local curvatures of differential geometry. The shape index is computed from the curvatures, and its different values correspond to different shapes. In case of facial expression recognition using full-face images, if any portion of the face image is distorted, it may reflect on the recognition performance.

Much work has been done in this field where local texture, features have been extracted and used in the classification. Due to the very local nature of this information, the dimension of the feature vector achieved for the full image is very high, posing computational challenges in real-time expression recognition. In recent times, Dimensionality Reduction methods have been successfully used in image recognition tasks. Though being high dimensional data, natural images such as face images lie in low dimensional subspace, and Dimensionality Reduction methods try to learn this underlying subspace to reduce the computational complexity involved in the classification stage of image recognition task. Proposed E-PCA and CS-ONPP performed well and proved to be gaining a huge margin in terms of feature vector length while maintaining the same recognition accuracy.

Classical FER methods do well in certain well-controlled cases. The fundamental issue with hand-crafted features based arrangement approaches is that they require space learning and not generalize well like in the complex dataset. Fortunately, Deep Neural Network (DNN) is giving a satisfactory solution to these issues which were not able to deliver by the hand-crafted techniques. DNNs

should consider many training parameters, such as size (number of layers and number of units per layer), learning rate, and initial weights. They are sweeping through the parameter space for optimal parameters due to the cost in time and the calculation resources. Numerous tricks, such as batch processing (calculating the gradient in several training examples simultaneously instead of individual examples) speed up the calculation. So proposed DNNFG (DNN based on Fourier transform followed by Gabor filtering) which utilized VGG16_ft.and 2DNN (Two-channel based Deep Neural Network) easily utilized the local and global information about the expressions. For this task we utilized VGGFace, which is trained on 2.6M face images from 2.6k different people. VGGFace architecture is the same as the VGG16. To adapt VGGFace to FER problem, the VGGFace fine-tuned (denoted as VGGFace_ft) by freezing four blocks of VGGFace and tuning the parameter of the last block. DNN based methods improved recognition accuracy compared to classical approaches.

Table 7.1: Compare all proposed methods

DataBase	**2DTFP** [chapter 3]	**HSOG** [chapter 4]	**E-PCA** [chapter 5] [Section 5.1]	**CS-ONPP** [chapter 5] [Section 5.3]	**DNN-FG** [chapter 6 [Section 6.1]	**2DNN** [chapter 6 [Section 6.2]
JAFFE	92.78	94.15	89.01	94.54	95.96	97.12
VIDEO	94.86	95.98	96.05	95.67	96.67	98.07
CK+	93.78	93.89	91.72	87.88	96.78	96.76
Oulu-CASIA	96.87	97.00	95.32	93.15	96.08	98.12

Overall conclusion is that the thesis addresses the classical facial expression recognition approaches and its limitations, then moved to deep learning-based approaches to handle these limitations. All the new proposals are tested on benchmark data-sets of facial expression recognition. In all cases, the new proposals outperform the conventional method in terms of recognition accuracy as mentioned in Table. 7.1. All the proposed methods compared witrh state of art methods incorporated in Table. 7.2.

Table 7.2: Comparison with Recognition Accuracy reported in some State-of-the-Art facial expression methods

DataBase	Methods	Network Type	Additional Classifiers	Accuracy
CK+	Ouellet [121]	CNN (AlexNet)	SVM	94.40%
	Li et al. [87]	RBM	-	96.08 %
	Liu et al. [95]	DBN	Adaboost	96.7%
	Liu et al. [91]	CNN, RBM	SVM	92.05%
	Khorrami et al. [71]	zero-bias CNN	-	95.7%
	Ding et al. [28]	CNN with fine-tune	-	96.8%
	Zeng et al. [201]	DAE	-	93.78%
	Cai et al. [9]	CNN+loss layer	-	94.39%
	Meng et al. [110]	CNN		95.37%
	Liu et al. [96]	CNN+loss layer	-	97.1%
	Yang et al. [192]	GAN	-]	97.3%
	shan et al. [151]	CNN	-	80.303 %
	Su et al. [163]	zero-bias CNN	-	95.12%
	Yang et al. [192]	WMDNN	-	97.02%
	Ours	2DNN	SVM	96.76%
JAFFE	Liu et al. [95]	DBN	Adaboost	93.0%
	Hamester et al. [48]	CNN, CAE		95.8%
	Hamester et al. [48]	CNN, CAE		94.1%
	shan et al. [151]	CNN		76.74 %
	Su et al. [163]	zero-bias CNN		97.6%
	Yang et al. [192]	WMDNN		92.21%
	Ours	2DNN	SVM	97.12%
OULU-CASIA	Ding et al. [28]	CNN with fine-tune		87.71%
	Cai et al. [9]	CNN+loss layer		85.58%
	Yang et al. [192]	GAN		88.0 %
	Yang et al. [192]	WMDNN		92.89%
	Ours	2DNN	SVM	98.12%

7.2 Future Research Directions

This thesis has presented novel approaches for Facial Expression Recognition (FER) based on hand-crafted features and deep learning-based techniques. In the process of this work, however, we identified related problems that one may consider worth pursuing. These are briefly described as follows.

- **Towards Facial Expression Recognition in the Wild** Recognizing facial expression in a wild setting has remained a challenging task in computer vision. The World Wide Web is a good source of facial images in which most of them are

captured in uncontrolled conditions. The Internet is a Word Wild Web of facial images with expressions. So in the future, we will focus on applying our proposed methods in the wild dataset.

- **Speed up the Algorithm**

Our future work will focus on simplifying the network used to speed up the Algorithm. Furthermore, we plan to focus on other channels of facial images that can be used to further improve the fusion network.

- **Recognize the emotions using the fusion of the face and the speech**

Till now, we recognized the expressions using the face images only. So in the future, we can recognize the emotions using the fusion of the features came from the face and the speech because both are having an equal contribution to showing the emotion. That may be done by adding more channels in the Neural Network.

- **Context-Aware Emotion Recognition Networks**

Existing techniques for emotion recognition have focused on facial expression analysis only. Thus we can provide limited ability to encode context that comprehensively represents the emotional responses. In the future, we can implement the networks that exploit not only human facial expression but also context information in a joint and boosting manner.

The future direction mentioned are very recent and hence most of them are not included within the scope of the thesis. The thesis only addressed the problem of facial expression recognition and suggested couple of handcrafted feature based methods as well as methods using DNN techniques which is the state of the research in this domain.

References

[1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1609–1618, 2017.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[3] P. AKoringa, G. Shikkenawis, S. K. Mitra, and S. K. Parulkar. Modified orthogonal neighborhood preserving projection for face recognition. In *Pattern Recognition and Machine Intelligence*, pages 225–235. Springer, 2015.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[5] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Spatio-temporal convolutional sparse auto-encoder for sequence classification. In *BMVC*, pages 1–12, 2012.

[6] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436, 2016.

[7] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *Proceedings Eighth*

IEEE International Conference on Computer Vision. ICCV 2001, volume 1, pages 454–461. IEEE, 2001.

[8] S. Biswas and J. Sil. An efficient expression recognition method using contourlet transform. In *Proceedings of the 2nd International Conference on Perception and Machine Intelligence*, pages 167–174, 2015.

[9] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[10] J. Chen, J. Konrad, and P. Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1570–1579, 2018.

[11] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Information Sciences*, 428:49–61, 2018.

[12] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[13] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187, 2003.

[14] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

[15] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level net-

work for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE, 2016.

[16] M. J. Cossetin, J. C. Nievola, and A. L. Koerich. Facial expression recognition using a pairwise feature selection and classification approach. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 5149–5155. IEEE, 2016.

[17] S. F. Cotter. Sparse representation for accurate classification of corrupted and occluded facial expressions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 838–841. IEEE, 2010.

[18] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[19] A. Cruz, B. Bhanu, and S. Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In *International Conference on Affective Computing and Intelligent Interaction*, pages 341–350. Springer, 2011.

[20] M. Dahmane and J. Meunier. Continuous emotion recognition using gabor energy filters. In *International Conference on Affective Computing and Intelligent Interaction*, pages 351–358. Springer, 2011.

[21] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based hog features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 884–888. IEEE, 2011.

[22] M. N. Dailey, G. W. Cottrell, C. Padgett, and R. Adolphs. Empath: A neural network that categorizes facial expressions. *Journal of cognitive neuroscience*, 14(8):1158–1173, 2002.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[24] C. Darwin and P. Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[25] T. Devries, K. Biswaranjan, and G. W. Taylor. Multi-task learning of facial landmarks and expression. In *2014 Canadian Conference on Computer and Robot Vision*, pages 98–103. IEEE, 2014.

[26] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th international conference on multimodal interaction*, pages 461–466, 2014.

[27] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015.

[28] H. Ding, S. K. Zhou, and R. Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[29] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 506–513, 2016.

[30] Y. Ding, Q. Zhao, B. Li, and X. Yuan. Facial expression recognition from image sequence based on lbp and taylor expansion. *IEEE Access*, 5:19409–19419, 2017.

[31] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[32] P. Ekman. Facial action coding system. 1977.

[33] P. Ekman. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. 1994.

[34] P. Ekman. Facial action coding system (facs). *A human face*, 2002.

[35] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[36] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[37] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450, 2016.

[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[39] A. J. Fitch, A. Kadyrov, W. J. Christmas, and J. Kittler. Fast robust correlation. *IEEE Transactions on Image Processing*, 14(8):1063–1073, 2005.

[40] D. Gabor. A performance evaluation of local descriptors. *J. Inst. Electr. Eng*, 93:429–459, 1946.

[41] Y. Gao, M. K. Leung, S. C. Hui, and M. W. Tananda. Facial expression recognition from line-based caricatures. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 33(3):407–412, 2003.

[42] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011.

[43] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[44] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning. Local features based facial expression recognition with face registration errors. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8. IEEE, 2008.

[45] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

[46] M. Guo, X. Hou, Y. Ma, and X. Wu. Facial expression recognition using elbp based on covariance matrix transform in klt. *Multimedia Tools and Applications*, 76(2):2995–3010, 2017.

[47] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao. Deep neural networks with relativity learning for facial expression recognition. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016.

[48] D. Hamester, P. Barros, and S. Wermter. Face expression recognition with a 2-channel convolutional neural network. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2015.

[49] B. Hasani and M. H. Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–40, 2017.

[50] B. Hasani and M. H. Mahoor. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 790–795. IEEE, 2017.

[51] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[52] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

[53] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 553–560, 2017.

[54] D. Huang, C. Zhu, Y. Wang, and L. Chen. Hsog: a novel local image descriptor based on histograms of the second-order gradients. *IEEE Transactions on Image Processing*, 23(11):4680–4695, 2014.

[55] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[56] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[57] S. Jain, C. Hu, and J. K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 1642–1649. IEEE, 2011.

[58] Y. Ji and K. Idrissi. Automatic facial expression recognition based on spatiotemporal descriptors. *Pattern Recognition Letters*, 33(10):1373–1380, 2012.

[59] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE transactions on cybernetics*, 44(2):161–174, 2013.

[60] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Face and Gesture 2011*, pages 314–321. IEEE, 2011.

[61] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258, 1987.

[62] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[63] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[64] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550, 2013.

[65] J.-K. Kamarainen, V. Kyrki, and H. Kalviainen. Invariance properties of gabor filter-based features-overview and applications. *IEEE Transactions on image processing*, 15(5):1088–1099, 2006.

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

[67] T. Kaneko, K. Hiramatsu, and K. Kashino. Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 327–331, 2016.

[68] S. Kankanamge, C. Fookes, and S. Sridharan. Facial analysis in the wild with lstm networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1052–1056. IEEE, 2017.

[69] H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[70] J. M. Keller, M. R. Gray, and J. A. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4):580–585, 1985.

[71] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[72] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–57, 2016.

[73] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 427–434, 2015.

[74] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236, 2017.

[75] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[76] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598*, 2017.

[77] E. Kokiopoulou and Y. Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2143–2156, 2007.

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[79] S. Kumar, M. Bhuyan, and B. Chakraborty. Extraction of informative regions of a face for facial expression recognition. iet comput vis 10: 567–576, 2015.

[80] J. Kumari, R. Rajesh, and K. Pooja. Facial expression recognition: A survey. *Procedia Computer Science*, 58:486 – 491, 2015.

[81] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. Von Der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on computers*, 42(3):300–311, 1993.

[82] Y.-H. Lai and S.-H. Lai. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 263–270. IEEE, 2018.

[83] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006.

[84] Q. V. Le, N. Jaitly, and G. E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.

[85] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. nature 521. 2015.

[86] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 503–510, 2015.

[87] J. Li and E. Y. Lam. Facial expression recognition using deep neural networks. In *Imaging Systems and Techniques (IST), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[88] S. Li and W. Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[89] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[90] G. Littlewort, J. Whitehill, T.-F. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett. The motion in emotion—a cert based approach to the fera emotion challenge. In *Face and Gesture 2011*, pages 897–902. IEEE, 2011.

[91] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[92] M. Liu, S. Li, S. Shan, and X. Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159:126–136, 2015.

[93] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian conference on computer vision*, pages 143–157. Springer, 2014.

[94] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on multimodal interaction*, pages 494–501, 2014.

[95] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1805–1812, 2014.

[96] X. Liu, B. Vijaya Kumar, J. You, and P. Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–29, 2017.

[97] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Euler principal component analysis. *International journal of computer vision*, 101(3):498–518, 2013.

[98] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[99] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[100] J. Lu, V. E. Liong, and J. Zhou. Cost-sensitive local binary feature learning for facial age estimation. *IEEE Transactions on Image Processing*, 24(12):5356–5368, 2015.

[101] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[102] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki. Facial expression recognition with deep age. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 657–662. IEEE, 2017.

[103] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *2014 International Conference on Smart Computing*, pages 303–308. IEEE, 2014.

[104] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998.*

Proceedings. Third IEEE International Conference on, pages 200–205. IEEE, 1998.

[105] H. Mahersia and K. Hamrouni. Using multiple steerable filters and bayesian regularization for facial expression recognition. *Engineering Applications of Artificial Intelligence*, 38:190–202, 2015.

[106] B. Martinez and M. F. Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in face detection and facial image analysis*, pages 63–100. Springer, 2016.

[107] D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992.

[108] V. Mavani, S. Raman, and K. P. Miyapuram. Facial expression recognition using visual saliency and deep learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2783–2788, 2017.

[109] R. K. McConnell. Method of and apparatus for pattern recognition, Jan. 28 1986. US Patent 4,567,610.

[110] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017.

[111] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

[112] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[113] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings*

of the 2015 ACM on international conference on multimodal interaction, pages 443–449, 2015.

[114] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014.

[115] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes. Deep spatio-temporal features for multimodal emotion recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1223. IEEE, 2017.

[116] S. Nigam, R. Singh, and A. Misra. Efficient facial expression recognition using histogram of oriented gradients in wavelet domain. *Multimedia Tools and Applications*, 77(21):28725–28747, 2018.

[117] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas. Subclass discriminant non-negative matrix factorization for facial image analysis. *Pattern Recognition*, 45(12):4080–4091, 2012.

[118] S. Noh, H. Park, Y. Jin, and J.-I. Park. Feature-adaptive motion energy analysis for facial expression recognition. In *International Symposium on Visual Computing*, pages 452–463. Springer, 2007.

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

[120] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer, 2008.

[121] S. Ouellet. Real-time emotion recognition for gaming using deep convolutional network features. *arXiv preprint arXiv:1408.3750*, 2014.

[122] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang. Audio-visual emotion recognition using deep transfer learning

and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 577–582, 2017.

[123] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[124] R. W. Picard. A ective computing, 1997.

[125] S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 536–543, 2017.

[126] G. Pons and D. Masip. Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing*, 9(3):343–350, 2017.

[127] G. Pons and D. Masip. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*, 2018.

[128] R. C. Rahul and M. Cherian. Facial expression recognition using pca and texture-based ldn descriptor. In *Proceedings of the International Conference on Soft Computing Systems*, pages 113–122. Springer, 2016.

[129] R. Rajesh, K. Rajeev, V. Gopakumar, K. Suchithra, and V. Lekhesh. On experimenting with pedestrian classification using neural network. In *2011 3rd International Conference on Electronics Computer Technology*, volume 5, pages 107–111. IEEE, 2011.

[130] R. Rajesh, K. Rajeev, K. Suchithra, V. Lekhesh, V. Gopakumar, and N. Ragesh. Coherence vector of oriented gradients for traffic sign recognition using neural networks. In *The 2011 International Joint Conference on Neural Networks*, pages 907–910. IEEE, 2011.

[131] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE*

International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 17–24. IEEE, 2017.

[132] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439, 2014.

[133] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012.

[134] B. D. Rigling and R. L. Moses. Taylor expansion of the differential range for monostatic sar. *IEEE Transactions on Aerospace and Electronic Systems*, 41(1):60–64, 2005.

[135] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *International conference on artificial intelligence and statistics*, pages 951–959, 2012.

[136] H. Roy and D. Bhattacharjee. Local-gravity-face (lg-face) for illumination-invariant and heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 11(7):1412–1424, 2016.

[137] B. Ryu, A. R. Rivera, J. Kim, and O. Chae. Local directional ternary pattern for facial expression recognition. *IEEE Transactions on Image Processing*, 26(12):6006–6018, 2017.

[138] F. Z. Salmam, A. Madani, and M. Kissi. Facial expression recognition using decision trees. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, pages 125–130. IEEE, 2016.

[139] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[140] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro. Local zernike moment representation for facial affect recognition. In *BMVC*, volume 2, page 3, 2013.

[141] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 485–492, 2012.

[142] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[143] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[144] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[145] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456, 2012.

[146] M. Schuster. Paliwal, and k. kuldip,"bidirectional recurrent neural networks,". *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[147] N. Sebe, M. S. Lew, and T. S. Huang. The state-of-the-art in human-computer interaction. In *International Workshop on Computer Vision in Human-Computer Interaction*, pages 1–6. Springer, 2004.

[148] T. Senechal, V. Rapp, and L. Prevost. Facial feature tracking for emotional dynamic analysis. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 495–506. Springer, 2011.

[149] C. Shan and R. Braspenning. Recognizing facial expressions automatically from video. In *Handbook of ambient intelligence and smart environments*, pages 479–509. Springer, 2010.

[150] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[151] K. Shan, J. Guo, W. You, D. Lu, and R. Bie. Automatic facial expression recognition based on a deep convolutional-neural-network structure. In *Software Engineering Research, Management and Applications (SERA), 2017 IEEE 15th International Conference on*, pages 123–128. IEEE, 2017.

[152] W. Shang, K. Sohn, D. Almeida, and H. Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *international conference on machine learning*, pages 2217–2225, 2016.

[153] G. Shikkenawis and S. K. Mitra. On some variants of locality preserving projection. *Neurocomputing*, 173:196–211, 2016.

[154] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee. Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. *IEEE Transactions on Image Processing*, 24(4):1386–1398, 2015.

[155] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan, and S. Lee. Depth camera-based facial expression recognition system using multilayer scheme. *IETE Technical Review*, 31(4):277–286, 2014.

[156] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*, pages 250–259. Springer, 2012.

[157] P. C. G. d. Silva. *Multimodal emotion recognition*. PhD thesis, 2015.

[158] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE, 2003.

[159] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[160] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[161] M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou. Image ratio features for facial expression recognition application. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(3):779–788, 2009.

[162] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[163] W. Su, L. Chen, M. Wu, M. Zhou, Z. Liu, and W. Cao. Nesterov accelerated gradient descent-based convolution neural network with dropout for facial expression recognition. In *Control Conference (ASCC), 2017 11th Asian*, pages 1063–1068. IEEE, 2017.

[164] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han. Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognition Letters*, 119:49–61, 2019.

[165] J. Susskind, G. Hinton, R. Memisevic, and M. Pollefeys. Modeling the joint density of two images under a variety of transformations. In *CVPR 2011*, pages 2793–2800. IEEE, 2011.

[166] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

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

[168] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

[170] M. R. Teague. Image analysis via the general theory of moments. *JOSA*, 70(8):920–930, 1980.

[171] Y. Tian, T. Kanade, and J. F. Cohn. Facial expression recognition. In *Handbook of face recognition*, pages 487–519. Springer, 2011.

[172] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[173] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[174] B. Tunç, V. Dağlı, and M. Gökmen. Class dependent factor analysis and its application to face recognition. *Pattern recognition*, 45(12):4092–4102, 2012.

[175] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

[176] A. Uçar, Y. Demir, and C. Güzeliş. A new facial expression recognition based on curvelet transform and online sequential extreme learning ma-

chine initialized with spherical clustering. *Neural Computing and Applications*, 27(1):131–142, 2016.

[177] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 1, pages 635–640. IEEE, 2004.

[178] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.

[179] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Face and Gesture 2011*, pages 921–926. IEEE, 2011.

[180] V. Vielzeuf, S. Pateux, and F. Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576, 2017.

[181] S. Voeffray. Emotion-sensitive human-computer interaction (hci): State of the art-seminar paper. *Emotion Recognition*, pages 1–4, 2011.

[182] L. Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.

[183] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[184] L. Wiskott, N. Krüger, N. Kuiger, and C. Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):775–779, 1997.

[185] B.-F. Wu and C.-H. Lin. Adaptive feature mapping for customizing deep learning based facial expression recognition model. *IEEE access*, 6:12451–12461, 2018.

[186] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett, and J. R. Movellan. Action unit recognition transfer across datasets. In *Face and Gesture 2011*, pages 889–896. IEEE, 2011.

[187] X. Xie and K.-M. Lam. Gabor-based kernel pca with doubly nonlinear mapping for face recognition with a single face image. *IEEE Transactions on Image Processing*, 15(9):2481–2492, 2006.

[188] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal. Video emotion recognition with transferred deep feature encodings. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 15–22, 2016.

[189] W. S. Yambor. Analysis of pca-based and fisher discriminant-based image recognition algorithms. Master's thesis, Colorado State University, 2000.

[190] H. Yan. Collaborative discriminative multi-metric learning for facial expression recognition in video. *Pattern Recognition*, 75:33–40, 2018.

[191] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong. Multi-cue fusion for emotion recognition in the wild. *Neurocomputing*, 309:27–35, 2018.

[192] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[193] H. Yang, Z. Zhang, and L. Yin. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 294–301. IEEE, 2018.

[194] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *Face and Gesture 2011*, pages 866–871. IEEE, 2011.

[195] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen. Holonet: towards robust emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 472–478, 2016.

[196] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[197] Z. Yu, G. Liu, Q. Liu, and J. Deng. Spatio-temporal convolutional features with nested lstm for facial expression recognition. *Neurocomputing*, 317:50–57, 2018.

[198] Z. Yu, Q. Liu, and G. Liu. Deeper cascaded peak-piloted network for weak expression recognition. *The Visual Computer*, 34(12):1691–1699, 2018.

[199] S. Zafeiriou and M. Petrou. Sparse representations for facial expressions recognition via l 1 optimization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 32–39. IEEE, 2010.

[200] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen. Hand-crafted feature guided deep learning for facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 423–430. IEEE, 2018.

[201] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643–649, 2018.

[202] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[203] F. Zhang, T. Zhang, Q. Mao, and C. Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018.

[204] K. Zhang, Y. Huang, Y. Du, and L. Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

[205] L. Zhang and D. Tjondronegoro. Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, 2(4):219–229, 2011.

[206] S.-q. Zhang. Enhanced supervised locally linear embedding. *Pattern Recognition Letters*, 30(13):1208–1218, 2009.

[207] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu. Face recognition under varying illumination using gradientfaces. *IEEE Transactions on Image Processing*, 18(11):2599–2606, 2009.

[208] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia*, 18(12):2528–2536, 2016.

[209] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *IEEE Transactions on Multimedia*, 14(4):995–1007, 2012.

[210] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

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

[212] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[213] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.

[214] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016.

[215] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1):38–52, 2010.

[216] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE, 2012.

List of Publications

- **Journal:**

  1. Sujata, and Suman K. Mitra, "2DNN AND MOD-FER: Double channel based deep neural network for the Modular Facial Expression Recognition", manuscript submitted to *International Journal of Pattern Recognition and Artificial Intelligence*
  2. Sujata, and Suman K. Mitra, "MOD-FER - A Modular Approach For Facial Expression Recognition Using 2D Taylor Expansion", manuscript submitted to *SN Computer Science*

- **Conferences:**

  1. Sujata and Suman K. Mitra. "A Modular Approach For Facial Expression Recognition using HSOG," in *Proceedings of the* 9^{th} *International Conference on Advances in Pattern Recognition, ICAPR* . ISI Banglore, India, December, 2017.
  2. Sujata, Maitry Trivedi and Suman K. Mitra. "A Modular Approach For Facial Expression Recognition using Euler Principal Component Analysis (e-PCA)" in *Proceedings of the IEEE Conference on Applied Signal Processing (ASPCON 2018).*
  3. Sujata, Purvi A Koringa and Suman K.Mitra. "CS-ONPP AND MOD-FER: Class Similarity based ONPP for the Modular Facial Expression Recognition", manuscript accepted in *7th National Conference on Com-*

puter Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG 2019).

4. Sujata and Suman K. Mitra. "DNNFG-MFER: DNN based on Fourier transform followed by Gabor filtering for the modular Facial Expression recognition", manuscript submitted in ICPRAM, 2020.
5. Sujata and SumanK.Mitra."ModularFacialExpressionRecognitionUsingDouble Channel DNNs", manuscript accepted in VISAPP, 2021.
6. Prapti Trivedi, Purva Mhasakar, Sujata and Suman K. Mitra. " Multichannel CNN for Facial Expression Recognition." manuscript accepted in 8^{th} *International Conference on Pattern Recognition and Machine Intelligence (PReMI), 2019.*
7. Pranjal Bhatt, Sujata and Suman Mitra. " Kernel Variants Of Extended Locality Preserving Projection." manuscript accepted in 4^{th} *Computer Vision & Image Processing (CVIP), 2019.*
8. Saloni Mundra, Sujata and Suman Mitra. " Facial Expression Recognition on Noisy Data Using Robust PCA." manuscript accepted in 16^{th} *IEEE India Council International Conference (INDICON 2019).*