Extraction of Facial Informative Regions and Discriminative Shared Space for Facial Expression Recognition

A

thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Sunil Kumar

Department of Electronics and Electrical Engineering

Indian Institute of Technology Guwahati

Guwahati - 781 039, India

May, 2017.

To my family members

for their love, support and encouragement

Certificate

This is to certify that the thesis entitled "**Extraction of Facial Informative Regions and Discriminative Shared Space for Facial Expression Recognition**", submitted by **Sunil Kumar** (126102003), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the Institute and in our opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Place: IIT Guwahati

Dr. M.K. Bhuyan
Associate Professor,
Department of Electronics and Electrical Engineering,
Indian Institute of Technology (IIT) Guwahati,
India - 781039

Acknowledgements

I would like to express my sincere thanks to all those people who made this dissertation possible. First and foremost, I would like to express my profound respect and gratitude to my supervisor, Dr. M.K. Bhuyan, who has been guiding force behind this work. I am greatly indebted for his invaluable guidance, constant encouragement, and his valuable comments on my work. I am fortunate enough to have such an advisor who gave me the freedom to think independently and explore new ideas. More importantly, I would like to thank for the patience he has shown in carefully reading and commenting on the manuscripts, and countless revisions of this dissertation. His commitments and dedication to research have been and will continue to be a constant source of inspiration for me. I have no doubts that finishing my degree in proper and timely manner was impossible without his help. I am highly privilege to have got an opportunity to work with such a wonderful person.

I would also like to thank my doctoral committee members Prof. P. K. Bora, Dr. P. Guha and Prof. Diganta Goswami for their invaluable suggestions, encouragements, and moral supports that helped me to improve my research work. I am also thankful to the Head of the Department, other faculty members and staffs for their kind help carried out during my academic studies.

My special thanks to Mr. Sanjib Das for providing various resources useful for the research work. Also, his contribution to my research work is highly acknowledged. My special thanks to my friend cum moral adviser Ms. Malathi. T (Malathi Ma'am), for her guidance and insightful comments during the entire journey of my PhD life. Her constant support, patience, and care made my stay in IITG memorable.

On a personal note, I would like to thank my friends Mr. Venkat K., Mr. R. K. Bhukya and Mr. B. K. Chakraborty for their constant support and being with me in every aspect of my life. My special thanks go to, Mr. Santhosh, Mr. Parveen, Mr. Gaurav, Mr. Abishek, Ms. Ranjani, and Mr. Anirudadh (Abhijeet), for their motivation and support. I had a great time with my

many friends at IIT Guwahati, including (but not limited to) Mr. Lalatendu Behera, Malathi Ma'am, Anand, D. Das, Pradipta, Debajeet, Tilendra and their team, Vikram, and of-course Chaudhauri tea shop. I would like to thanks them for their support and encouragement. An special appreciation goes to Mr. Dillep, Mr. Anirudadh, Mr. Jafin, and Mr. Debajeet for correcting the thesis and giving critical and valuable suggestions.

I am grateful to my parents, sisters, brothers and bhabhi, whose love encouragement, and support made this research work possible. I would especially thank my Bhaiya, Bhabhi and my most heartfelt love to their children: *Kishan*, *Saurav*, and *Swati*, who are the joy of my life.

I am thankful to IIT Guwahati for providing the research scholarship to undertake my PhD research. A kind thanks to all the doctors, nurses, and staffs of IITG who has taken care of my health time to time. Finally, I would like to thank the Almighty God for bestowing me this opportunity and showering his blessings on me to come out successful against all odds.

Sunil Kumar

Abstract

Recognition of human's emotion through facial expressions has many important applications ranging from behaviour recognition, human-computer interaction, security, psychology, and so on. Recognition of facial expressions from non-frontal faces, and recognition from different views are two important research challenges. As different views of a facial expression are just different manifestations of the expression, the information embedded in different views can be effectively utilized for facial expression recognition (FER). Motivated from the above mentioned facts, we proposed to extract facial informative regions and discriminative shared space for facial expression recognition.

Extraction of discriminative features for different facial expressions is a key step in facial expression recognition. However, most discriminative facial features can be extracted from the informative regions of a face. In this view, the importance of different facial sub-regions are investigated, and subsequently the facial sub-regions which have significant contributions in different facial expressions are only considered for feature extraction. Furthermore, a weighted-projection based local binary pattern (WPLBP) feature is proposed. For this, texture features are extracted from informative regions and they are weighted on the basis of their importance. Finally, an efficient face model is derived from the informative regions of a face. The proposed face model has several advantages, and it gives better performance than other existing face models.

Next, we proposed an Uncorrelated Multi-view Discriminant Locality Preserving Projection (UMvDLPP) analysis to recognize expressions from multi-view face images. The proposed UMvDLPP first transforms expressions of different views to a common uncorrelated discriminative subspace, and then classification is performed. One of the major advantages of our proposed scheme is that classifiers need not be learned separately for all the views. Moreover, it can effectively handle multi-modal characteristics of multi-view data than the existing learning-based methods.

Discriminative shared Gaussian process latent variable model (DS-GPLVM) [1] can give better performance in the multi-view FER than the existing multi-view linear and non-linear learning-based methods. Laplacian-based prior used in DS-GPLVM only captures topological structure of data space without the inter-class separability of the data, and hence, the obtained latent space is not optimal. So, an efficient prior is proposed, which not only depends on the topological structure of the intra-class data, but also on the local-between-class-scatter-matrix of the data onto the latent manifold. The proposed approach employs a hierarchical framework, which is termed as multi-level uncorrelated DS-GPLVM (ML-UDSGPLVM). In this framework, expressions are first divided into three sub-categories. Subsequently, each of the sub-categories are further classified to identify the constituent basic expressions. Hence, first level of ML-UDSGPLVM i.e., 1-UDSGPLVM is learned for classification of different categories, and then a separate 2-UDSGPLVM is learned for recognizing constituent expressions of each of the categories. Extensive experiments on a standard dataset show that our proposed method performs better than the existing multi-view FER methods. This improvement is due to the fact that the proposed method enhances the discrimination between the classes more effectively, and classifies expressions category-wise followed by classification of the basic expressions embedded in each of the sub-categories (hierarchical approach).

Contents

List of Figures

List of Tables

List of Acronyms

FER	Facial Expression Recognition
LBP	Local Binary Pattern
SLR	Sign Language Recognition
HCI	Human-computer Interaction
FACS	Facial Action Coding System
FAPs	Facial Animation Parameters
AUs	Action Units
MPEG	Moving Picture Experts Group
MvFER	Multi-view FER
PCA	Principal Component Analysis
FDA/LDA	Fisher/Linear Discriminant Analysis
MLP	Multi-layer Perceptron
ICA	Independent Component Analysis
SVM	Support Vector Machine
LPP	Locality Preserving Projection
CRI	Common Reference Image
IRE	Informative Region Extraction
WPLBP	Weighted-Projection-based LBP
RRs	Recognition Rates
VLBP	Volumetric LBP

LBP^{u2}	Uniform LBP
NI	Neutral Image
NPE	Normalize Projection Error
IREM	Informative Region Extraction Model
MDA	Multiple Discriminant Analysis
KA	Krippendroff Alpha
AA	Average Accuracy
LFDA	Local Fisher Discriminant Analysis
LDP	Local Derivative Pattern
LDPv	Local Directional Pattern Variance
LNP	Local Directional Number Pattern
SIFT	Scale-Invariant-Feature-Transform
HOG	Histogram of Oriented Gradient
AAM	Active Appearance Model
RBF	Radial Basis Function
ADM	Alternative Direction Method
MTSL	Multi-task Sparse Learning
FDM	Feature Disentangling Machine
HMM	Hidden Markov Model
CRF	Conditional Random Field
HCRF	Hidden Conditional Random Field
CCS	Correlated Common Space
UCS	Uncorrelated Common Space
DCT	Discrete Cosine Transform
GPLVM	Gaussian Process Latent Variable Model
D-GPLVM	Discriminative GPLVM

DS-GPLVM	Discriminative Shared GPLVM
ML-DSGPLVM	Multi-level DSGPLVM
UDSGPLVM	Uncorrelated DSGPLVM
ML-UDSGPLVM	Multi-level Uncorrelated DSGPLVM
kNN	k-nearest neighbourhood
LBCSM	Local Between-Class Scatter Matrix
UMvDLPP	Uncorrelated Multi-view Discriminant LPP Analysis
MvDA	Multi-view Discriminant Analysis
GMA	Generalized Multi-view Analysis
CCA	Canonical Correlation Analysis
GMPCA	Generalized Multi-view PCA
GMLDA	Generalized Multi-view LDA
GMCCA	Generalized Multi-view CCA
PW-CCA	Pair-wise CCA
PW-LDA	Pair-wise LDA
H-UMvDLPP	Hierarchical UMvDLPP

List of Symbols

D	Dimension of the observation space
d	Dimension of the reduced subspace
A_k	Similarity matrix of k^{th} view
$\mathbf{I}$	Identity matrix
U	Uniformity
C	Number of classes
λ	Number of facial sub-regions
$\dim(\cdot)$	Dimension of argument
C	Number of classes
$\perp$	Perpendicular symbol
β	Precision parameter
γ^v	Back-projection parameter
$R(\mathbf{A})$	regularization term
I_{bp}	Independent back-projection
S_{bp}	Single back-projection
n_c^v	Number of samples belongs to c^{th}
$\mathbf{S}_{lb}^v$	Local between-class scatter matrix for v^{th} view
$\mathbf{L}_{net}$	Sum of normalize Laplacian matrices for all the views
$\mathbf{S}_b$	Between-class scatter matrix
$\mathbf{S}_w$	Within-class scatter matrix

$\delta_{i,j}$	Kronecker delta function
$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_V\}$	Kernel parameters of the shared observation space
$\mathbf{X}^v$	Observation space for v^{th} view
$\mathbf{Y}_{ccs}$	Reduced correlated common space
$\mathbf{Y}_{ucs}$	Reduced uncorrelated common space
$\mathbf{Q}_{eq}$	Intra and inter-views transformation matrix
$\mathbf{Q}_{kk}$	Intra-view local between-class scatter matrix
$\mathbf{Q}_{kl}$	Inter-view local between-class scatter matrix
$\mathbf{P}_{eq}$	Intra and inter-views LPP transformation matrix
$\mathbf{L}_{kl}$	Inter-view Laplacian matrix
$\mathbf{L}_{ccs}$	Laplacian matrix for CCS
$\mathbf{B}_{ccs}$	Local between-class scatter matrix for CCS

1

Introduction

Facial expression is the most effective form of non-verbal communication and it provides a clue about emotional state, mindset and intention. Automatic recognition of human's facial expressions has become an active research topic in Computer Vision as facial expression recognition has many applications in various aspects of our everyday life. In the computer vision community, the term "facial expression recognition" often refers to the classification of facial features in one of the six so called basic emotions: happiness, sadness, fear, disgust, surprise and anger, as introduced by Ekman et al. [2]. This thesis contributes to the research and development of facial expression recognition systems from two aspects: first, feature extraction from informative regions of a face for facial expression recognition from frontal view images, and second, extension of this method to recognize expressions from multi-view face images. Experimental results on publicly available databases show that the effectiveness of the proposed approaches for the applications of facial expression recognition. This chapter gives an overview of the facial expression recognition problem, applications of facial expression recognition, and the major challenges in this research. Finally, organization of the thesis is presented at the end.

1.1 Facial Expression Recognition

Facial expression recognition (FER) is a Computer Vision problem of recognizing human's emotions with the help of visual appearance of a face. For the last two-three decades, FER is a growing research field [3]. Recognizing facial expressions was also investigated by several psychologists [4, 5], and one of the important contributions in this direction was presented by *et al.* [2]. Ekman *et al.* defined a set of 44 action units (AUs), and they also demonstrated that a large number of facial expressions can be represented in terms of AUs. Facial expressions like *anger*, *disgust*, *fear*, *happy*, *sad*, and *surprise* are called basic expressions. Most of the existing FER methods mainly focus on recognizing these basic expressions [3, 6–8]. This is due to the fact that basic expressions are more frequent during verbal and non-verbal communications, and these expressions can give significant information of human's emotion [9].

Research on automatic FER has been actively started from last two-three decades due to several advancement in the field of Computer Vision, Machine Learning, and Cognitive Science [10]. Facial expression recognition is an important research problem as it has many diverse applications ranging from behaviour recognition, human-computer intelligent interactions, sign language recognition, video surveillance, robot control, and many more [3, 11–14]. This section highlights some of the major applications of facial expression recognition, which motivates us to carry forward our research in the direction of human's expression recognition.

1.1.1 Human-behaviour recognition

Behaviour recognition is a process of studying internal feeling or mood of a person [15, 16]. The behaviour of a person depends on many uncertain entities such as mind, environment, situation, and so on. Advantage of face-based behaviour recognition is that it provides a vision-based platform to study human's mind. Since different facial expressions are resulted due to the movements of different facial muscles, and by recognizing facial expressions the mood or intension (behaviour) of individuals can be understood [17].

1.1.2 Non-verbal communication

Non-verbal communication is basically a speechless communication, where information is conveyed in a well-defined pattern in the form of sign gestures. Manual parameters such as hand shape, hand orientation, hand location, and their movements, and non-manual parameters such as face, head and body parts are the components of a sign language gesture [14, 18]. In general, manual parameters are more important than non-manual parameters in non-verbal communications. However, manual parameters alone are not sufficient to represent all the gestures of a sign language. This is due to the fact that some of the signs are almost similar in manual parametric space but differ in non-manual parametric space. In such cases, information extracted from non-manual parametric space such as facial expressions and head postures can play a crucial role in differentiating one gesture from the rest. Additionally, non-manual components can also change and/or boost the meaning of a sign. Among different non-manual components, facial expression is the most important gesture as it can influence the meaning of a sign. Hence, it is essential to integrate non-manual parameters such as facial expressions to enhance non-verbal communication.

1.1.3 Human-computer interaction (HCI)

The function of a HCI system is to enable a human to communicate with computer more naturally than the traditional input devices such as a mouse, a keyboard, a joystick etc [19]. Accordingly, the primary aim of the HCI research is to build up a platform through which a human can communicate with a computer using "natural" means that humans employ to communicate with one another. In general, HCI is a very broad interdisciplinary field involving computer scientists, psychologists, cognitive scientists, and other disciplines [20, 21]. Interpersonal behavioural pattern such as facial expressions can also be used to interact with computers [22]. To facilitate more natural and more human-like interaction, computer has to accurately recognize different human emotions with the help of facial expressions [20].

1.1.4 Security

As discussed earlier, behaviour of a person can be judged by recognizing facial expressions, and so, human activities can be partially understood by recognizing associated facial expressions. This has several applications in automated video surveillance such as monitoring activities of individuals in shopping malls, banks, prisons, and so on. Facial expressions may also be used as a biometric trait for authentication. Other applications of facial expression recognition include lie detection, activity monitoring, driver's mood detection, and many more [23, 24].

1.1.5 Medical

For psychological treatments [25], it is important to understand behaviour of a mentally ill patient by looking their facial expressions and body activities. In such cases, facial expressions and behaviour of a person are more crucial than that of verbal communications. So, an automated behaviour monitoring system using facial expressions is an important requirement for medical practitioners.

Motivated from all these applications, this thesis aims to solve some of the existing problems related to facial expression recognition problem.

1.2 General Paradigm of FER System

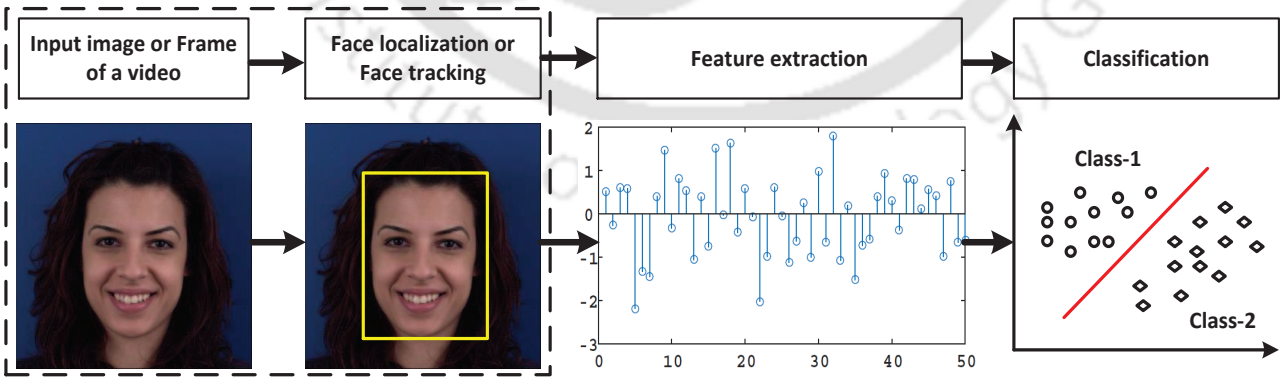


Figure 1.1: General paradigm of facial expression recognition (FER) system.

Figure 1.1 shows a typical structure of facial expression recognition system, and it consists of

mainly three steps – face localization and tracking, feature extraction, and classification. The input to a FER system is either an image or a video, and so, the first step is to localize the face or track the face in a video. This step is quite important as it provides raw data/observations for facial feature extraction. Therefore, face should be accurately localized in order to extract facial features only from the facial regions. Inaccurate localization may give unwanted facial features. The first step of FER *i.e.*, face localization/tracking is shown in the dotted block of Figure 1.1, where first image of the block shows an input face image, and the corresponding localized face is shown in the second image of the block.

Once face is localized, next step is to extract facial features from the localized face. As FER is a pattern recognition problem, feature extraction plays a crucial role in facial expression recognition process. Feature extraction aims to extract distinct features across different facial expressions. More importantly, distinct features can only be extracted from salient regions of a face [26]. This is due to the fact that all the regions of a face do not contribute in different facial expressions. Hence, one important research challenge is to extract features only from the informative regions of a face.

The final step of FER is the classification of different facial expressions. Classification may be supervised or unsupervised. The main difference between supervised and unsupervised classifications is that the supervised classification scheme uses class-label information to train the selected model, whereas unsupervised classification scheme does not require class-labels of training samples. Supervised classification scheme mainly consists of two steps *i.e.*, training and testing. In training step, for a given set of training examples and their associated class-labels, selected classifier learns to find the best parameters which can separate training samples for classification.

1.3 Facial Expression Generation System

Human brain is the main organ of central nervous system as it controls activities of all parts of the body by means of receiving, processing and transferring neuron information [27]. More specifically, the functional unit of a brain which controls activity of muscle deformations of a face is known as *limbic* system [27,28]. Out of several components of limbic system, *hippocampus*

and *amygdala* are mainly responsible for facial expression generation. The pictorial view of a limbic system with their components amygdala (indicated by "a") and hippocampus (indicated by "b") are shown in Figure 1.2 (a). Figure 1.2 (b) shows a block schematic representation of facial expression generation system. This indicates that any changes in the components of a limbic system generate emotions as shown in Figure 1.2 (a). The changes in the limbic system finally changes the appearance of a face by deforming facial muscular regions, and thereby different facial expressions are observed.

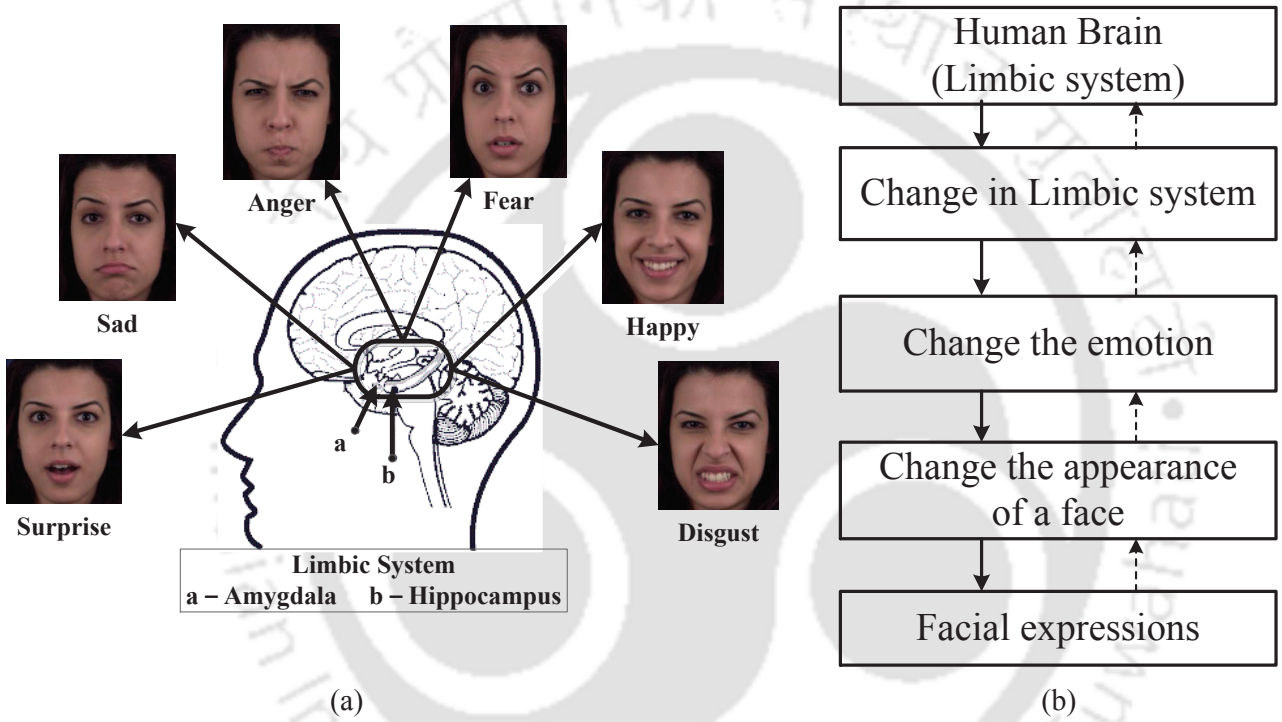


Figure 1.2: Emotion generation process.

Literature shows that ideally infinite number of expressions are feasible, however only few of them are more prominent like *happy*, *disgust*, *fear*, *anger*, *sad*, and *surprise*. These expressions are termed as basic expressions [2]. Intuitively, it is worthless to mention that when all the components of a limbic system are relaxed, then there will not be any deformations on a face. This particular state/expression of a face is known as a neutral expression. On the other hand any deviations of facial regions with respect to a neutral face image can generate different expressions. Hence, analysis of brain activities with the help of facial deformations information may be carried out. Researchers analyzed several characteristics of facial deformations such as

difference between a neutral expression and other expressions to formulate a placement rule to place different expressions in a co-ordinate system [4, 5]. In all their analysis [4, 5], neutral expression is placed at the center of the co-ordinate system.

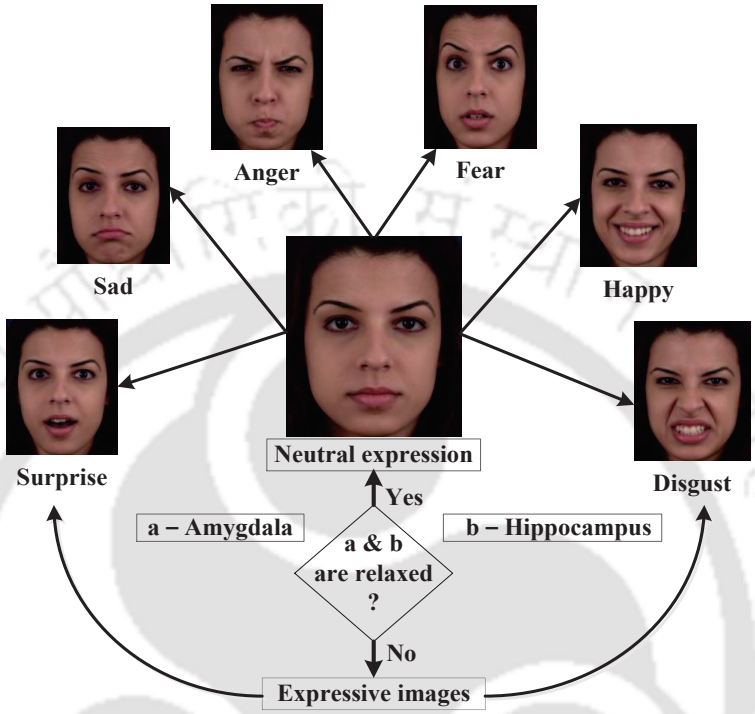


Figure 1.3: Generation of expressive face images from a neutral face image.

The deviations between a neutral expression and other expressions is a cue to recognize basic expressions [29, 30]. Any expressions are generated due to the movements of different facial sub-regions with respect to a neutral expression [2]. The generation process of basic expressions is shown in Figure 1.3.

1.4 Facial Expression Parametrization

Facial expression parameterization is important to describe, analyze, and recognize facial muscle movements of a face, and so, it also helps in feature extraction process. There are mainly two standard facial parametrization models, and they are termed as facial action coding system (FACS) [2] and facial animation parameters (FAPs) [31]. The details of FACS and FAPs models are described as follows:

1.4.1 Facial action coding system

The basis of FACS is action units (AUs), which define muscle movements of different facial regions. Each of the AUs defines the movements of a particular region of a face with respect to their neutral position. For example, AU1 and AU2 represent "inner portion of the brows raised" and "outer portion of the brows raised" respectively. Figure 1.4 shows the deviations of the regions of eyes and eyebrows with respect to their neutral positions, and these are called AU1, AU2, AU3, and AU4 as shown in Figure 1.4. In [2], Ekman et al. defined a set of

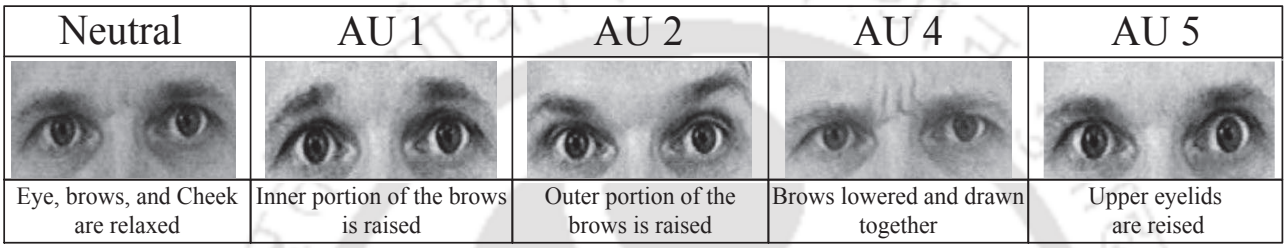


Figure 1.4: Representation of few action units (AUs): AU1, AU2, AU4, and AU5 of a face defined in FACS.

44 AUs, and they showed that more than 7000 expressions can be generated by combining different action units [32]. Two or more action units are additive or non-additive depending upon whether the respective action units are independent or dependent. If the movement of an AU "x" does not affect the movement of AU "y", then AUs "x" and "y" are said to be independent, otherwise they are dependent. It is to be noted that each AU is defined on the basis of the muscle movements of a particular facial sub-region with respect to the state of that particular muscle of a neutral state. Hence, the salient or informative regions of a face can be identified on this basis.

1.4.2 Facial animation parameters

Facial animation parameters (FAPs) [31] is a landmark-based approach to parametrize different facial activities. The FAPs is developed by moving pictures experts group (MPEG) by localizing 84 facial landmark points on a neutral face image. Figure 1.5 shows locations of 84 facial landmark points on a neutral face image for both frontal and profile views. The key idea behind

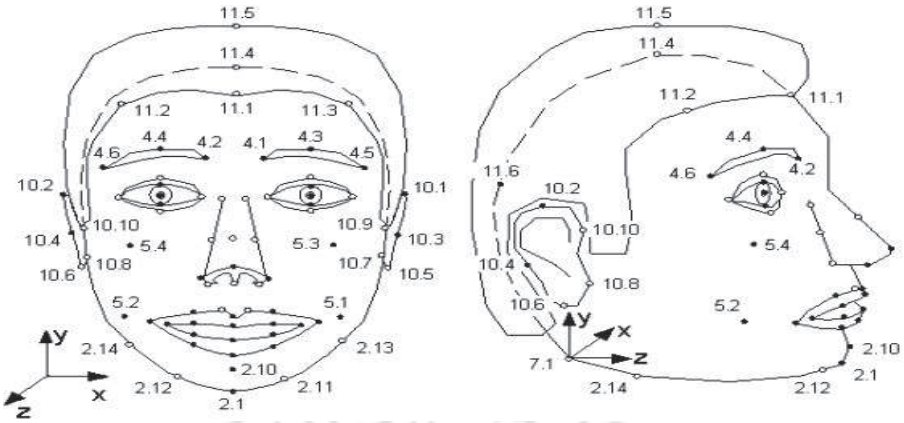


Figure 1.5: Localization of 84 landmark points on a sketched neutral face image defined in FAPs [31].

the FAPs system is to find location differences between a set of facial points of an expressive face image and a neutral face image. These location differences give the information of the movements of different facial sub-regions. The FAPs-based approach describes geometrical deformations of different facial regions, hence this information can also be effectively utilized to recognize facial expressions [1, 33].

In summary, FACS-based parametrization considers changes of appearance of local facial regions. Therefore, texture features are mainly used to recognize different action units. It is also observed that texture features can give better performance as compared to geometric features [34]. However, texture features extracted from all the regions of a face may not be discriminative, as all the regions of a face may not involve in different facial expressions. Hence, most of the existing works use texture features which are only extracted from a set of localized facial points/regions of a face [1, 33, 35, 36].

1.5 Multi-view/View-invariant FER

In many practical situations, captured expressive face images may not be frontal. Facial expression recognition from frontal face images has vary limited applications as compared to multi-view or view-invariant FER as pose-invariant expressions are more natural and realistic.

The state-of-the-art FER methods mainly focus on recognizing expressions either from an

image or a video in a multi-view setup [1, 8, 33]. Multi-view facial expression recognition (MvFER) is a slightly relaxed form of view-invariant FER, where expressions from a set of pre-defined views are recognized. The top row of Figure 1.6 shows "happy" expression for a set of pre-defined views *i.e.*, $-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, and $45°$ views. The bottom row of Figure 1.6 shows images of "happy" expression for arbitrary head-poses. In these cases, the objective is to develop a model to recognize expressions for any arbitrary head-poses.

Figure 1.6: Example of multi-view happy face images from BU3DFE dataset [37] (top) and images of arbitrary-view of happy expressions from SFEW dataset [38] (bottom).

1.6 Major Challenges in Recognizing Facial Expressions

There are many important research issues which are to be addressed to develop an efficient FER system for recognizing expressions from both frontal and non-frontal face images. The following subsections highlight some of the major research challenges in the context of efficient FER.

1.6.1 Face localization/tracking

In general, face localization or tracking is the first step of any facial expression recognition system. Inaccurate localization/tracking can adversely affect recognition process because of improper feature selection and noise. This step is even more challenging in case of multi-view and view-invariant FER, as a face has to be tracked in a real environment. The challenges in face tracking arise due to uneven lighting conditions, occlusions, clutter background, camouflage, and so on. Majority of facial expression recognition methods mainly used Viola-Jones face

tracker [39] to localize a face in a plain background. However, Viola-Jones face tracker fails to localize a face when a face undergoes significant movements. Also, poor lighting conditions adversely affect the performance of Viola-Jones tracker.

1.6.2 Occlusion

A part of a face may not be visible due to obstructions by different objects, and so, recognizing facial expressions only using a part of a face is a challenging research problem [40]. Hiding mouth and regions nearer to eyes significantly reduce the recognition accuracy [41]. Figure 1.7 shows few examples of occluded face due to either hiding of different facial regions by artificial obstacles or occlusion due to movements of a face.

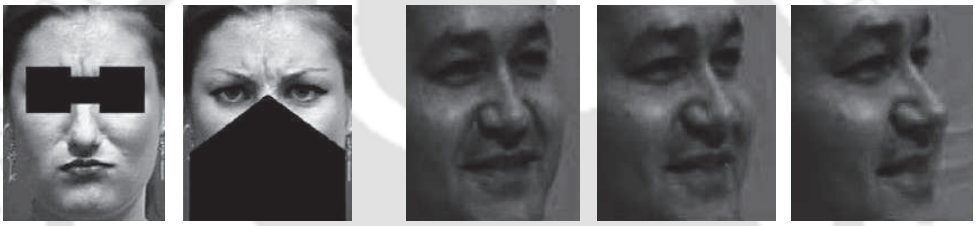
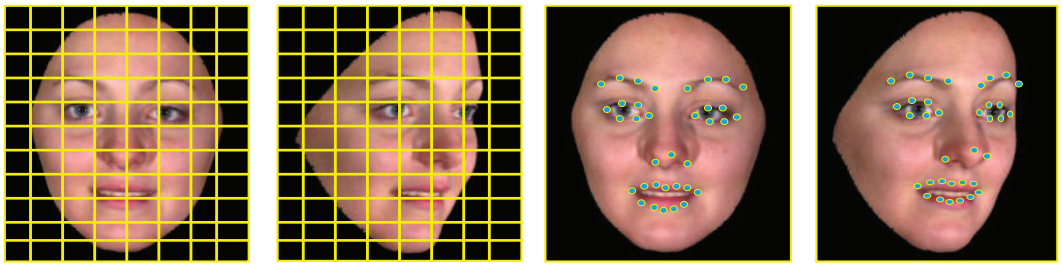


Figure 1.7: (a) Occlusion due to obstacles [41], and (b) Occlusion due to movement of a face [8].

1.6.3 Feature extraction

Extraction of efficient facial features for FER is another important issue. This is even more challenging if the face is not properly localized or a part of the face is occluded or not visible as discussed in subsections 1.6.1 and 1.6.2. Also, as discussed in Section 1.2, another important aspect is to localize informative regions of a face to extract most discriminative facial features.

A few state-of-the-art techniques for extracting facial features are shown in Figure 1.8. Figures 1.8 (a) and (b) show the process of extracting texture features for both frontal and non-frontal views. In this method, entire face is first divided into a number of sub-regions/sub-blocks, and then features are extracted from each of the sub-blocks. Finally, the features are concatenated to get the feature vector. However, features extracted in this fashion add several



(a) (b) (c) (d)

Figure 1.8: State-of-the-art techniques for facial feature extraction.

off-face features and features from several inactive regions of a face. So, the feature vector would be less discriminative, and its dimension will be very high. On the other hand, Figures 1.8 (c) and (d) show the process of extracting features from some selected landmark points (shape-based approach) for both frontal and non-frontal views. In shape-based approach, a set of facial points are first localized in the active regions of a face, and then geometrical features are extracted from the landmark facial points. The advantage of shape-based approach is that both geometric and texture features can be extracted from the landmark facial points. Moreover, shape-based features can more effectively represent a face as compared to texture features for non-frontal face images [1]. In this context, it is to be mentioned that a common feature space needs to be extracted for recognizing multi-view expressions, and this concept is elaborately discussed in the next chapters of the thesis.

1.6.4 Recognition of non-basic expressions

Recognition of any other facial expressions (non-basic expressions) in addition to basic expressions is another important research challenge. As proposed by Ekman *et al.* [2], there may be infinite number of spontaneous expressions, and all these expressions cannot be labelled or annotated for classification. Also, it is quite difficult to represent any spontaneous non-basic expressions in terms of per-defined AUs, as accurate recognition of all 44 AUs is itself a challenging task.

1.7 Organization of the Thesis

This thesis comprises of seven chapters including the present one.

Chapter 2 presents a detailed review on facial expression recognition methods for recognizing from both frontal and non-frontal views. Different feature extraction approaches and classification strategies are elaborately discussed in this chapter. Finally, based on the limitations of existing facial expression algorithms, motivation and objectives of our research work is finalized.

Chapter 3 describes the proposed scheme for extracting features only from the informative regions of a face. For this, informative region extraction model is proposed, which models the importances of different facial regions. Finally, a weighted-projection-based local binary pattern (WPLBP) feature is proposed, which is subsequently employed for recognition of facial expressions from frontal face images.

In **Chapter 4**, an efficient face model based on informative regions of a face is proposed. The proposed face model is used to extract geometrical as well as texture features of a face. The efficacy of proposed face model is analyzed and compared with the existing face models.

A more accurate modelling scheme called uncorrelated multi-view discriminant locality preserving projection (UMvDLPP) analysis is proposed in **Chapter 5** to recognize expressions from multi-view facial images. The proposed method projects images of different views to a common uncorrelated discriminative space, and then classification is performed with the help of a kNN classifier.

A hierarchical framework for multi-view FER is proposed in **Chapter 6**, which is termed as multi-level uncorrelated discriminative shared Gaussian process latent variable model (ML-UDSGPLVM). In the proposed scheme, expressions are first classified into three categories based on first level of our proposed latent variable model. Subsequently, basic expressions embedded in each of the categories are recognized in the second stage of our hierarchical classifier.

Finally, **Chapter 7** will provide the concluding remarks of this dissertation and outline current limitations together with future research directions.

Supplementary details essential to understand this research work are presented in the **Appendix A.1** and **Appendix A.2**.

2

A Review on Methods of Facial Expression Recognition

Facial expression recognition (FER) from visual patterns of a face has attracted significant attentions of many researchers. This chapter gives a brief review on state-of-the-art methods developed for facial expression recognition from frontal face and multi-view face images. More specifically, we reviewed existing state-of-the-art techniques for feature extraction and their modeling to obtain a discriminative space. Face normalization/alignment is an important issue of multi-view facial expression recognition, where the objective is to map each of the views of facial images to a common view (canonical view). Additionally, we also discussed the methods which search a common discriminative space for multiple observations, and the drawbacks of these methods are highlighted. Finally, based on the literature survey, we presented motivation and objectives of the thesis at the end of this Chapter. A brief discussion on the databases used in our experiments is also presented.

2.1 Introduction

Facial expression recognition (FER) is one of the active research topics in the field of Computer Vision [1, 8, 33, 42, 43]. Some of the important applications of FER include human-behavior recognition, neuroscience, psychology, non-verbal communication, human-computer interaction, security, surveillance, and etc. [10, 22–24, 37, 40, 44, 45]. The idea of vision-based FER was introduced in 17^{th} century, however it became an active research topic in last 20^{th} century due to several advancements in algorithms of computer vision, graphics, and machine learning [10]. In the early stage of automatic FER, researchers started facial expression recognition research by recognizing expressions from frontal face images, and then they slowly moved towards recognizing expressions from multi-view and/or view-invariant face images. The current scenario of automatic facial expression recognition mainly focuses on recognizing facial expressions from multi-view/view-invariant face images. More importantly, the features employed for FER from frontal face images have to be modified for FER from non-frontal face images, and this has not been much explored [1, 33]. To highlight current research trend in this direction, we first presented a brief literature on facial feature extraction techniques developed so far, and then we discussed the methods of multi-view/view-invariant FER. Finally on the basis of literature survey, motivation and objectives of this thesis are presented at the end of this chapter.

2.2 Overview of Different Methods of FER

In general, automatic FER refers to the problem of recognizing facial expressions on the basis of some facial characteristics, and so FER is a pattern recognition problem. FER consists of mainly three steps: 1) face localization/tracking, 2) feature extraction, and 3) recognition. A standard representation of FER system is shown in Figure 1.1, where each of the individual blocks of Figure 1.1 is described in Section 1.2. In FER, face localization/tracking is mostly carried out using some face tracking algorithms like Viola-Jones method [33, 39, 46]. Subsequently, face part is extracted in several earlier works [8, 42, 47]. Next, all the face images are normalized to have a common resolution for facial feature extraction.

Next step in FER is feature extraction, and the recognition performance depends on the

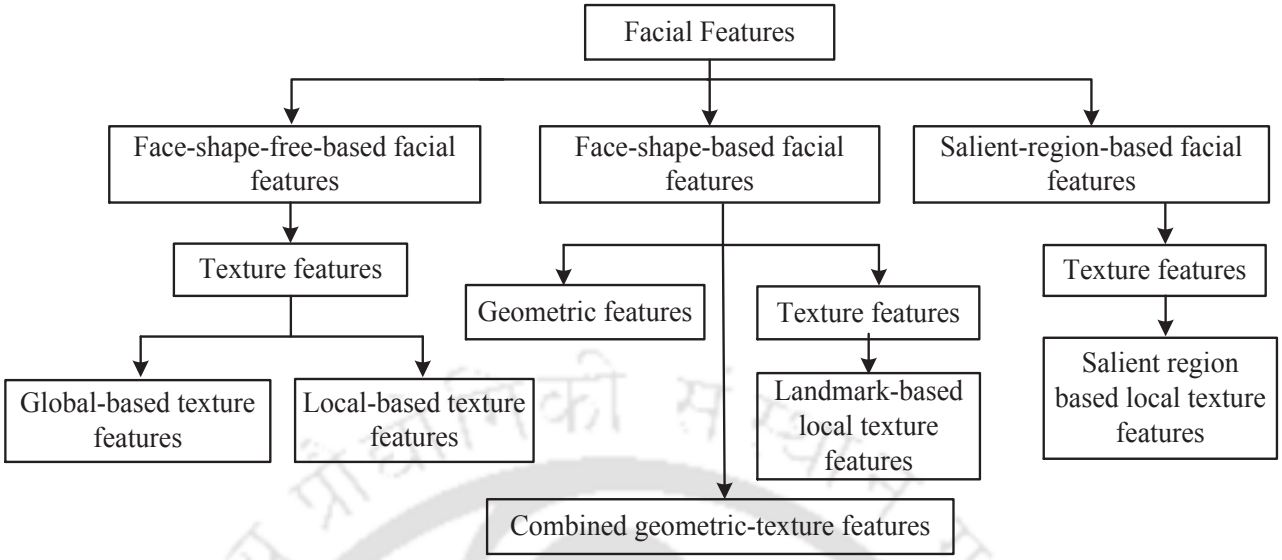


Figure 2.1: Overall classification of existing facial features.

extraction of discriminative features. Minimum intra-class variances (compactness of intra-class samples) and maximum inter-class separation ensure proper discrimination between the features. Therefore, feature extraction aims to extract discriminative features of a face image.

Facial feature extraction algorithms are broadly classified into three categories: 1) face-shape-free-based methods, 2) face-shape-based methods, and 3) salient-region-based methods. Figure 2.1 shows different types of existing facial features. Face-shape-free-based methods extract only texture features. These features are further sub-divided into global and local texture features. However, face-shape-based methods are facial landmark-based approaches, and hence, these methods are mainly used to extract geometrical features like distance between corners of a mouth, distance between eyes and eyebrows, and so on. Most of the facial points are localized regions near to eyes and mouth, as these regions are more informative in recognizing facial expressions. Also, regions around these facial points show significant changes in texture patterns during the execution different facial expressions. Hence, texture features around each of the landmark points are also analyzed for facial expression recognition in the literature. The detailed review on these methods are discussed in Section 2.3.

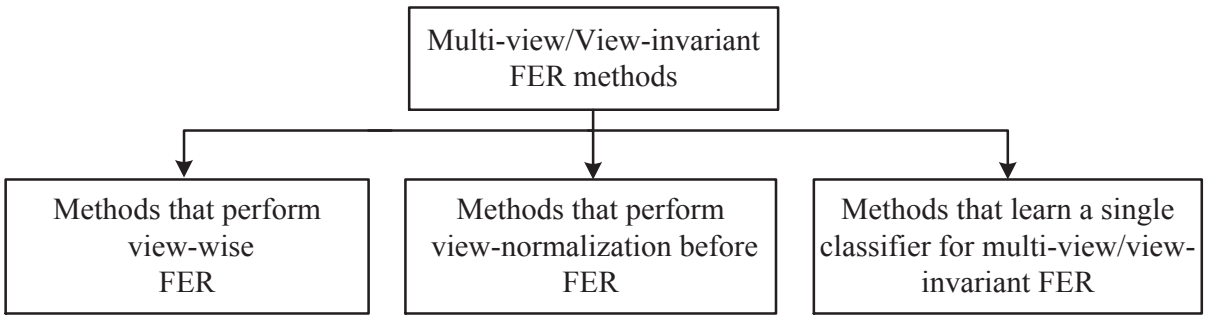


Figure 2.2: Classification of multi-view/view-invariant methods.

2.2.1 Multi-view and/or view-invariant FER

Recognition of facial expressions from multi-view/view-invariant face images is an important research direction of FER. In practical, captured expressive face images are non-frontal due to either head movements or variable camera positions, and so, methods developed for frontal face FER cannot be directly applied for recognizing facial expressions from multi-view and/or view-invariant face images. Methods of multi-view and/or view-invariant FER can be broadly grouped into three categories based on how underline methods are carried out expression recognition. These categories are: 1) methods which perform view-wise *i.e.*, pose-wise FER, 2) methods which perform view-normalization before expression classification, and 3) methods which learn a single classifier for multiple observations of multi-view facial images. This classification strategy is shown in Figure 2.2. The detailed review on the methods of each of these categories are discussed in Section 2.4.

2.3 Facial Features Extraction Methods

As discussed in Section 2.2 that, existing feature extraction approaches can be broadly classified into three groups (Figure 2.1). All these methods are discussed in the following subsections.

2.3.1 Face-shape-free-based methods

In these methods, either the entire face [6, 48–57] or a set of local regions of a face is used for facial feature extraction [3, 7, 41, 58–64]. Further, methods of this category is divided into

two groups based on how these methods extract features from a face. The methods which extract features from an entire face image are termed as global/holistic-based methods. On the other hand, methods which divide a face image into a number of sub-blocks, and then features are extracted from each of the sub-blocks or a set of sub-blocks are termed as local-based methods. Finally, a face image is represented by concatenating features extracted from each of the local regions. A detailed review on global-based and local-based feature extraction methods is presented below.

2.3.1.1 Global-based methods

Global-based methods [6, 48–57, 65] use either intensity values of a face image or their transformed values to form a feature vector. Authors of [49, 50, 56, 65, 66] use an eigenspace-based approach to obtained a low dimensional manifold for FER. The low-dimensional space comprises of a set d eigenvectors corresponding to first d largest eigenvalues of a covariance matrix $\mathbf{\Sigma}$. The matrix $\mathbf{\Sigma}$ is defined as:

$$\mathbf{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' \tag{2.1}$$

where, size of a covariance matrix $\mathbf{\Sigma}$ depends on the dimension of the feature vectors. In Eqn. (2.1), $\mathbf{x}_i$ represents feature vector of i^{th} column of a data/observation matrix, and $\boldsymbol{\mu}$ is the mean of the observation vectors. A separate d dimensional space is learned for each of the expressions. Finally during testing, an input image is first projected onto a low dimensional space using eigenvectors of each class, and subsequently a distance-based classifier is used to identify the class-labels of test samples. FER based on the above approaches is also known as Principal Component Analysis (PCA)-based approach. However, eigenspace-based methods are generally suffered from "curse-of-dimensionality" problem, as dimensions of a feature vector is very large, and limited number of training samples are available [67]. Curse-of-dimensionality problem can be reduced to an extent by selecting a subset of features corresponding to first few principal components. However, principal components obtained by using PCA do not ensure separabilities between the classes. As, PCA maximizes both the between-class and the within-class scatter matrices of the data [3].

Abidin and Harjoko [51] proposed Fisher discriminant analysis (FDA)-based approach to learn a low-dimensional discriminative subspace. FDA is also known as linear discriminant analysis (LDA), which maximizes between-class separability and minimizes within-class compactness of data. For C class problem, LDA learns a $C-1$ dimensional discriminative subspace. Finally, multi-layer perceptron (MLP) is used on LDA-space for training and testing. Long *et al.* applied independent component analysis (ICA) in order to learn a low-dimensional manifold for FER [53]. An enhanced-ICA is applied in [55] to recognize expressions from a video. However, ICA needs higher-order statistics of data such as kurtosis (a forth order moment) in order to learn independent components of a feature. In general, ICA-based approach is a compositionally intensive procedure for facial expression recognition. Also, components of feature vectors obtained by ICA may be correlated as principal directions obtained by ICA may or may not be orthogonal.

Gabor features show better ability to discriminate different texture patterns due to its spatial localization, orientation, and frequency selective properties [68]. Hence, several researchers used Gabor-based features for facial expression recognition [57, 69–71]. A 2D mother Gabor wavelet centered about origin is represented as follows [72]:

$$\psi(x,y) = \frac{1}{2\pi} e^{-\frac{1}{8}\left(4x^2+y^2\right)} \left[e^{i\kappa x} - e^{-\frac{\kappa}{2}}\right] \tag{2.2}$$

where, κ is a constant, and its value depends on bandwidth of the filter. A 2D Gabor kernel with a scale $\sigma_g = \frac{\omega_g}{\kappa}$ and orientation θ, $\psi(x, y, \sigma_g, \theta)$ can be generated by rotating and scaling of the mother Gabor wavelet as given in Eqn. (2.2). Subsequently, $\psi(x, y, \sigma_g, \theta)$ is convolved with a facial image to obtained Gabor texture features at a given scale σ_g and a given orientation θ.

In a holistic-based approach, Gabor kernel is first convolved with a facial image for different scaling and orientation parameters [73]. For a given scale and orientation parameters, the convolved Gabor textured images are of same size as that of original image. Hence, for M number of scales and P number of orientations, there are $M \times P$ number of convolved images. Subsequently, features of each of the convolved facial images are concatenated to represent texture features of a given face image. In general, dimension of Gabor features is very high, and hence it suffers from "curse-of-dimensionality" problem. Also, computational complexity

of Gabor features increases with the number of scaling and orientation parameters. Therefore, feature selection or dimensionality reduction plays a crucial role in the selection of only few discriminative features from overall concatenated Gabor features. The overall performance of Gabor-based FER features is less than 94%.

Kotsia *et al.* [41] addressed facial expression recognition problem by employing features extracted from different parts of a face. Gabor-based texture features along with shape-based features are investigated for this purpose. They found that occlusion in left part of a face or right part of a face has less effect on average accuracy. On the other hand, occlusion in the mouth region of a face affects FER accuracy more adversely than the occlusion in the regions nearer to eyes. This finding shows the relative importance of the regions nearer to mouth as compared to the regions nearer to eyes for FER [41, 61].

2.3.1.2 Local face-shape-free-based methods

These methods initially divide a face image into a number of sub-blocks, and then features from each of the sub-blocks are extracted for FER [7, 58–60, 62–64]. In literature, it is found that local features can give a better representation of a face as compared to global features [74].

Local binary pattern (LBP) is a widely used local feature descriptor for facial expression recognition [60–64, 75]. The popularity of LBP is because of the following reasons: 1) LBP is a non-parametric descriptor, 2) it is easy to compute, and 3) it is robust to an extent of illumination variations. Shan *et al.* [60] used LBP features to investigate the performance of LBP features in FER under low resolution conditions. They also analyzed comparative performances of Gabor and LBP-based features for low resolution images. This study showed that low-dimensional manifold obtained by LBP features is more discriminative as compared to Gabor-based features. For a resolution of 110×150 image, Gabor-based features give a recognition rate of about 89%, whereas LBP features give a recognition rate of 93%. On the other hand, for a very low resolution image *i.e.*, images of size 14×19, LBP gives a recognition rate of about 77%, which is about 2% higher than the recognition rate obtained by Gabor features. They investigated their classification performances using LDA [76] and support vector machine (SVM) classifiers. In [61], Zhang *et al.* extracted LBP features from facial images,

and then dimensionality reduction is performed to obtained a low dimensional subspace. PCA, LDA, locality preserving projection (LPP), and local Fisher discriminant analysis (LFDA) are used to obtain a low-dimension feature space. Overall, they found that LFDA finds a better discriminative space as compared to PCA, LDA, and LPP. In [77], LBP is integrated with kernel extension of linear dimensionality reduction approach. Furthermore, several authors are still modifying LBP features, and applying it with different classifiers to recognize facial expressions [78–80]. One of the extension of LBP is volume-LBP or 3D-LBP proposed in [62]. In [62], Zhao *et al.* showed effectiveness of LBP for modeling dynamic texture variations. In general, dynamic textures can be arbitrarily rotated, and hence, they encoded LBP from three orthogonal planes. In a video (x,y,t), these planes are xy-plane, xt-plane, and yt-plane. This specific case of 3D-LBP is termed as LBP-TOP. However, feature dimensions in this case would be very high when more number of neighbouring pixels are used. Nevertheless, with less neighborhood pixels, it is difficult to capture texture dynamics. On an average, recognition rates of 95.19% (in 2-fold) and 97.37% (in 10-fold) are obtained using LBP-TOP approach.

The above mentioned variants of LBP do not capture directional information. Moreover, LBP is sensitive to non-monotonic illumination variations, and hence, its performance degrades in presence of random noise [81, 82]. Local derivative pattern (LDP) [83, 84], local direction pattern variance (LDPv) [85], local directional number pattern (LNP) [82] are some other descriptors, which also show promising performance. However, all these texture descriptors require more additional operations than the LBP. The major difference between LBP and the rest of the texture descriptors is that LBP features are extracted on the basis of gray scale intensity values of an image pixels, whereas rest of the methods encode directional patterns of the image pixels. Hence, other descriptors need several convolution operations to obtain directional informations. Similar to LBP, a face image is first divided into a number of blocks, and then each of the blocks is convolved with different directional kernels in LDP [84]. So, these convolutions capture directional information, which are encoded to get LDP codes similar to as that in LBP. LDPv [85] is a modified form of LDP, where encoding is done on the basis of variance of outputs of the convolution operation. In LNP [82], encoding is done on the basis of maximum positive responses and minimum negative responses of the convolution operations.

Gabor features obtained by applying Gabor kernels on entire face image are highly correlated [86]. To overcome this drawback, local Gabor-based methods are proposed [7,63,64,71,87]. Deng *et al.* [87] selected a set of Gabor kernels by altering the frequency and the orientation parameters. This approach reduces the overall feature dimension by a factor of half. Subsequently, PCA and/or LDA are/is applied on these reduced feature vectors to obtain few discriminative facial features. They found that overall recognition accuracy obtained by these reduced features is still better than the global Gabor features. In [7], Gu *et al.* proposed a radial-based encoding system to reduce overall response of a Gabor filter image. A multi-layer feature representation based approach is proposed in [64], where each of the sub-blocks of an image is convolved with Gabor kernels, and subsequently convolved images are encoded using local binary pattern. In [63], Huang *et al.* introduced a monogenic signal representation based approach. In this method, convolved local Gabor images are further convolved with Riesz filters, and then magnitude, phase, and orientation informations are extracted. Finally, extracted information is encoded using LBP to obtain a monogenic binary pattern. However, performance of these approaches is still close to LBP-based approach, and they are comparatively more computationally complex than LBP. Hence, these methods are not so popular for FER.

2.3.2 Face-shape-based methods

In local methods, facial features are extracted from a set of local facial regions. As we have mentioned in our earlier discussion, local-based methods outperform global-based feature extraction methods. Local features can give comparatively better discriminative facial features, and they are also less computationally complex. In other words, features extracted only from a few salient regions of a face by discarding some of the non-informative facial regions are more discriminative. Hence, the facial regions which can give more discriminative features are called *Active/Salient or Informative regions*. In FER, one of the important issues is to extract features only from the active regions of a face. In face-shape-based approaches, active regions of a face are localized by a set of facial points *i.e*, landmark points. One of the standard facial landmarking schemes is developed by Moving Pictures Experts Group (MPEG) [31]. Parametrization of

a face using MPEG is called Facial Animation Parameters (FAPs), and it is shown in Figure 2.3.

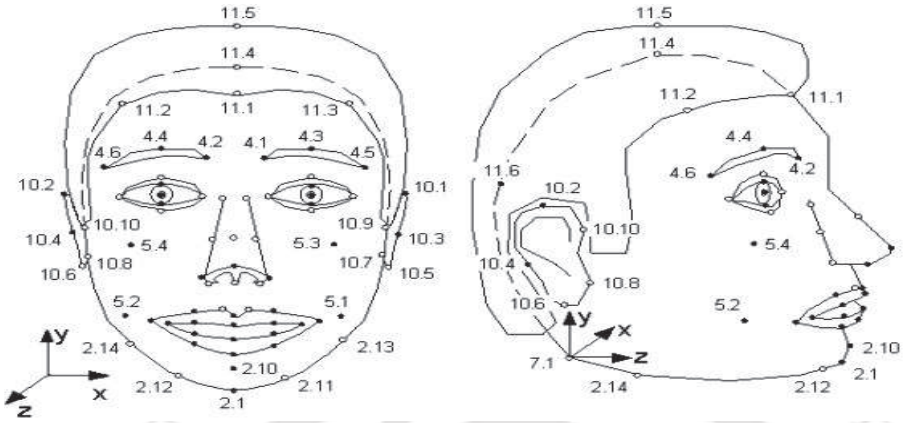


Figure 2.3: Facial Animation Parameters (FAPs) [31].

FAP comprises of 84 facial landmark points localized on a face. However, density of landmark points are comparatively more nearer to the regions of eyes and mouth as compared to other regions of a face such as cheeks and forehead. This is due to the fact that movements of facial regions nearer to eyes and mouth are more as compared to other regions, and therefore these regions conveys more information for vision-based face analysis such as face recognition and facial expression recognition. In face-shape-based methods, features are only extracted from the salient regions of a face, and hence they are more discriminative. However, automatic localization of facial landmark points is a major challenge. Automatic and accurate localization of facial points would be more challenging if more facial points are considered in this estimation. Automatic facial point localization is even more challenging for arbitrary face poses. One of the foremost works in the direction of automatic facial points localization is done by Cootes *et al.* [88, 89]. After locating facial landmark points, next task is to extract geometrical and texture features from these points.

2.3.2.1 Geometric and texture features

In FER, face-shape-based methods [33, 90–95] are mostly used to extract geometrical features. However, geometrical features are very sensitive to accurate localization of facial landmark points. Apart from the geometrical features, shape-based methods can also be used to extract

texture features [1, 92]. Texture features are extracted from small patches around each of the facial landmark points. Features extracted from such active regions or patches are more distinctive. Different state-of-the-art approaches in extracting facial features using shape-based methods are discussed below.

Hu *et al.* [90] manually localized a set of 41 landmark points to approximately represent a face. Subsequently, displacements of landmark points between expressive face images and corresponding neutral face images are used as features for FER. In [33], Rudovic *et al.* used 39 geometric points near to the regions of mouth, eyes, and eyebrows for pose-invariant FER. They recognized expressions from different poses in three steps. In first step, head-pose estimation is done by learning a set of discriminative basis functions using LDA, and then head-pose normalization is performed in the second step. Head-pose normalization is a process of learning a set of mapping functions, which can transform each of the poses to a frontal pose. Finally, pose-invariant FER is performed on the basis of distribution of expressive face images onto the learned common view. Zheng *et al.* [92] considered 83 landmark points to represent a face image, and then scale-invariant-feature-transform (SIFT) features are extracted around each of the landmark points [96]. One of the major shortcomings of this approach is that all the regions marked by 83 landmark points are not active, and hence, the face model representation proposed in [92] is not suitable as some of the non-active facial regions cannot give discriminative texture features. Similar to [92], few other attempts are made in [91, 97], where authors have extracted features like LBP, histogram of orientated gradient (HOG) [98], and SIFT at each of the landmark points. In [97], Hesse *et al.* found that low dimensional feature space obtained by locality preserving projection [99] is better than PCA and LDA-based approaches. Additionally, they used active appearance model (AAM) to localize landmark points on a given face image. A different approach is proposed in [95], where 3D facial points are projected onto a pose-invariant cylindrical surface constructed using positions of the eyes of different subjects. In case of large rotated face, invariant texture image obtained by projecting visible part of a face is used for facial expression recognition.

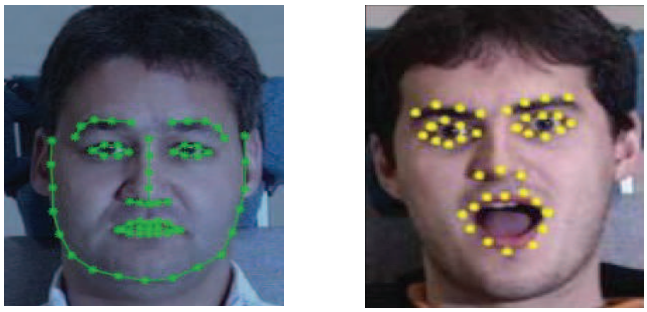
In summary, face-shape-based methods can be used for extracting both geometrical as well as texture features of a face. Also, geometrical features rely on accurate localizations of

facial points, and it does not give information of skin textual deformations such as wrinkles, bulges, and furrows. However, appearance-based features extracted on the basis of texture deformation of a face, and hence it is sensitive to illumination variations and noise. Hence, both the approaches *i.e.*, geometric-based and appearance-based have their own advantages and disadvantages. Hence, one of the alternative ways is to represent a face by both geometric and appearance-based features. In [14], Agris *et al.* showed the efficacy of the combined features (geometric and appearance-based) for sign language recognition. Zhang and Qiang [98] also emphasized the importance of combined features, and they tried to recognized action units and/or basic expressions from a video. One of our contributions in this direction was presented in [100].

2.3.3 Salient-region-based feature extraction methods

So far, we have discussed facial feature extraction methods where features are extracted either from the entire face or a set of pre-defined sub-regions of a face. Also, earlier studies showed that face-shape-based methods can give improved performance as compared to face-shape-free-based methods. This is due to the fact that shape-based methods extract features only from the salient regions of a face. Hence, texture description obtained by shape-based methods are distinctive for different classes of expressions. Some of the salient points about face-shape-based methods are highlighted as follows:

- Existing face models are mainly suitable for extracting geometrical features of a face [90, 91]. Also, shape-based methods are highly rely on accurate localization of facial landmark points.
- Boundary points around the face as shown in Figure 2.4(a) do not play significant role in facial expression recognition [101], and hence the face model shown in Figure 2.4(b) is almost equivalent to the face model shown in Figure 2.4(a) from recognition point of view.
- Also, face model shown in Figure 2.4(b) is not suitable for extracting appearance-based features, as there are no facial points between the regions near to eyebrows and the



(a) (b)

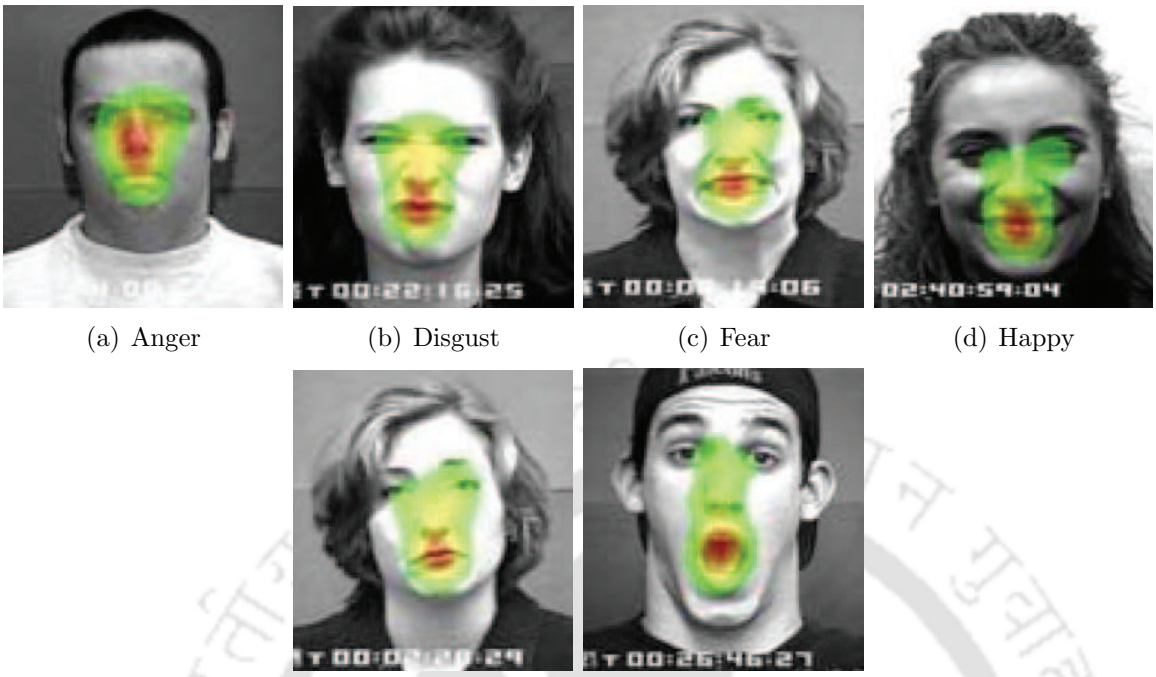
Figure 2.4: (a) Face model with facial boundary points, which have redundant landmark points for expression recognition (b) face model without facial boundary points, which gives equivalent performance as that of the model shown in (a).

regions near to jaw. However, these regions have significance in recognizing expressions like "anger" and "sad".

- Combine features (geometric and texture) give better performance as compared to individual geometric and texture-based features. Hence, there is a need to develop a face model to extract features from all the salient regions. Also, both the geometrical and texture features need to be extracted from the face model.

Few attempts have been made in the direction of finding active regions of a face for facial expression recognition [13, 26, 41, 102, 103]. These methods show a promising result when features are extracted only from the salient regions of a face. One of the early works for localizing active regions of a face is the method proposed by Khan *et al.* [13]. They conducted a psycho-visual experiment with the help of a eye-tracking system to determine the active regions of a face. The facial regions which play significant role during different expressions are shown in Figure 2.5. This experiment also indicates that active regions are different for different expressions. However, psycho-visual experiment is a complex procedure, as it requires a constraint environment with multiple cameras mounted on head and arms of an observer. So, this method has limited applications in finding active regions of a face for FER.

Zhong *et al.* [102] proposed a two stage multi-task sparse learning (MTSL)-based approach to localize a set of salient patches on a face from a set of pre-defined patches (facial subregions) [103]. In the first stage, patches which are informative/salient for all the expressions

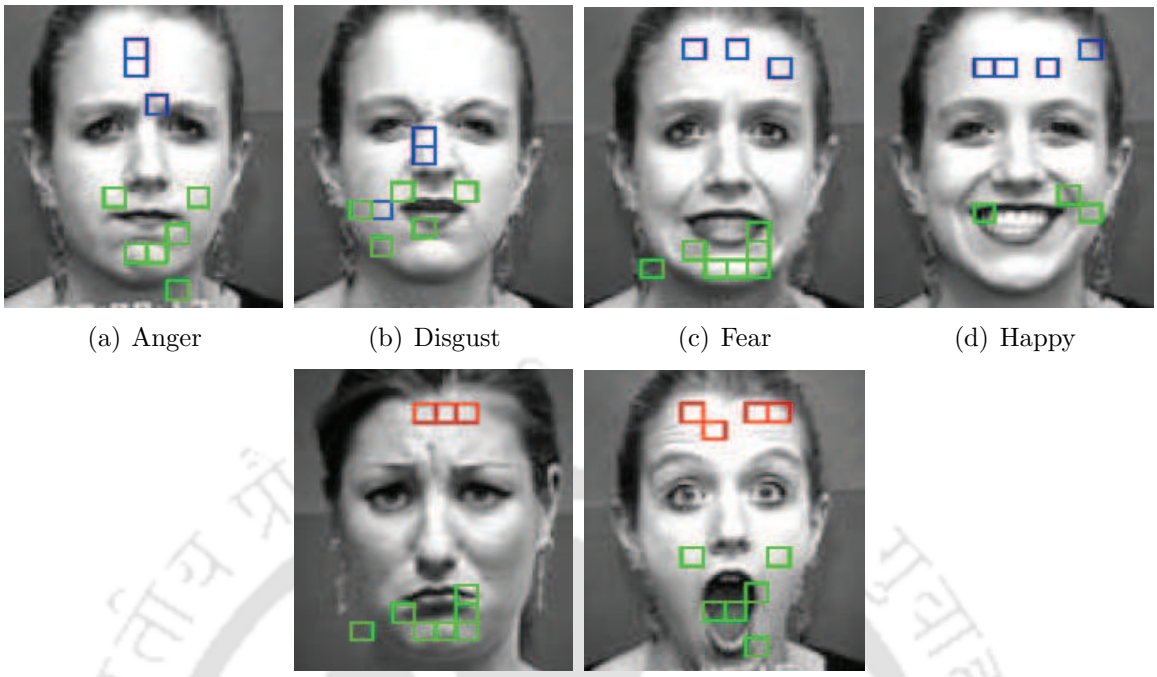


(a) Anger (b) Disgust (c) Fear (d) Happy

(e) Sad (f) Surprise

Figure 2.5: Salient regions highlighted by psycho-visual experiment presented in [13].

are learned by considering it as a binary class classification problem in a MTSL framework. In the second stage, expression specific salient patches are learned by coupling recognition stage with the face verification module into MTSL framework. The green patches shown in Figure 2.6 indicate salient regions of a face, which are common in all the basic expressions. Whereas, red and blue patches (facial sub-regions) are expression specific salient regions. It is observed that most of the common salient regions (green patches) are nearer to mouth, and so facial regions around the mouth capture maximum discriminative information for FER [41]. Liu *et al.* [103] jointly learned the interaction between the common patches and the expression specific patches for feature selection and classification in MTSL framework, which is termed as Feature Disentangling Machine (FDM). They obtained a very promising result as compared to existing feature extraction approaches. However, FDM is formulated as a constraint optimization problem, and hence it needs to learn a number of regularization and step size parameters. It is observed from Figure 2.6 that some of the salient patches are estimated outside the face, which may be due to improper learning of optimization parameters. Also, learning of expression specific patches requires an addition task of face verification, which



(a) Anger (b) Disgust (c) Fear (d) Happy

(e) Sad (f) Surprise

Figure 2.6: Common and expression specific salient regions of a face obtained by using multi-task sparse learning (MTSL)-based approach [102, 103].

increase computational complexity.

Ekman and Frisen [2] defined a set of AUs on the basis of the movements of a set of facial sub-regions with respect to their neutral states. These AUs are identified on the basis of texture pattern variations between an expressive face image and the corresponding neutral face. Our proposed method presented in Chapter 3 is mainly inspired from the basic theory proposed by Ekman *et al.* [47], and we proposed a mathematical model to extract active regions of a face. Our proposed work also supports the recent finding in [26], where active regions are localized based on geometrical analysis of a set of detected facial landmark points.

2.4 Review on Multi-view/View-invariant FER

A number of methods have been proposed to recognize expressions from frontal face images [3, 26,41,60–64,75,103]. Also, several state-of-the-art methods can give promising recognition rates in recognizing basic expressions like anger, disgust, fear, happy, sad, and surprise for frontal

faces [3, 26, 47, 103]. However, above mentioned FER methods in general are not applicable in many practical situations when the faces are not frontal. Hence, there is a need to develop an automated FER system which can recognize expressions from multi-view face images or in general view-invariant face images.

One of the possible ways of solving multi-view FER problem is the direct extension of existing FER methods in a pose-wise/view-wise manner [91, 92, 97]. Recognition accuracies obtained by pose-wise FER methods are not promising, specifically when the movement of a face is large. This may be due to the fact that a significant portion of a face is occluded due to the face movements, and hence, recognition entirely depends on the features extracted only from the visible part of a face. Also, it is observed in [104] that non-rigid movements of different facial regions during facial gesturing and the movement of rigid head part are non-linearly coupled. Hence, one of the important tasks is to decouple the above two components for robust view-invariant facial expression recognition. In this view, few works have been reported to decouple the above two components by first transforming facial images of each of the views to a common view (canonical view), and then recognition is performed. On this basis, existing multi-view/view-invariant FER methods are mainly classified into three groups as illustrated in Figure 2.2. These methods are grouped as: 1) methods which perform view-wise FER, 2) methods which perform view-normalization before FER, and 3) methods which learn a common discriminative space for multi-view/view-invariant FER. The detailed discussion on these methods is presented below.

2.4.1 View-wise/pose-wise multi-view FER

The general paradigm of pose-wise multi-view FER is shown in Figure 2.7. It mainly consists of three steps: 1) pre-processing of an input face image, 2) feature extraction, and 3) recognition. All these steps have to be performed for facial images of each of the views separately, and these steps are shown in Figure 2.7. The methods proposed in [36,91,92,97] deal with pose-wise multi-view FER. Hu *et al.* recognized multi-view FER by applying a set of appearance-based features such as LBP, HOG, and SIFT followed by reducing the dimensions of these features by PCA, LDA, and LPP. Finally, support vector machine is employed for classification. They obtained

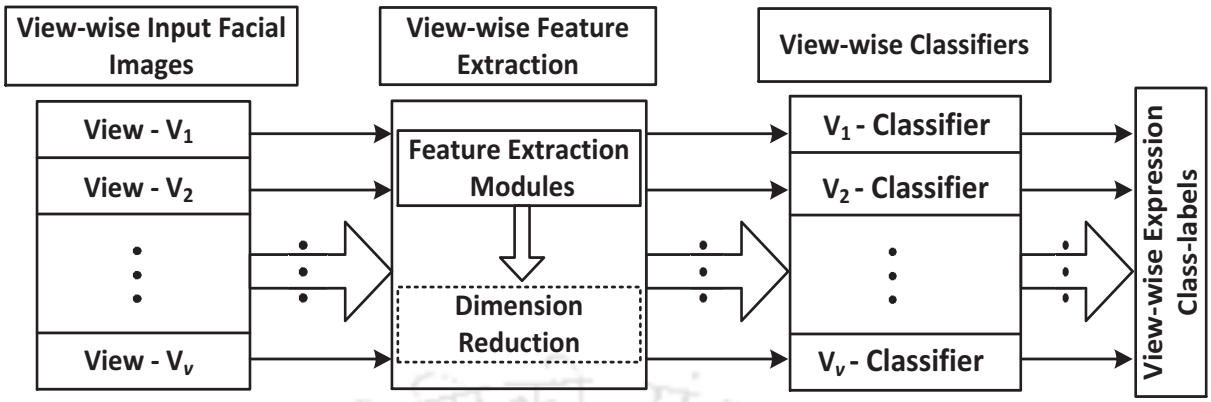


Figure 2.7: General paradigm of pose-wise multi-view FER.

highest classification accuracy with "SIFT + LPP + SVM", and optimal view for which the highest accuracy was obtained is 30° [91]. Zheng *et al.* proposed a Bayesian framework to model multi-view FER [92]. They employed a shape-based approach for feature extraction, where a face model having 83 facial landmark points is used to extract SIFT features. Their experiments showed that the optimal view may be obtained for a pan of 30° to 60°. In [36], Moore *et al.* evaluated performance of multi-view FER using variants of LBP with SVM classifier. They showed that frontal view can give optimal performance. Similar work is proposed in [97], where appearance-based features such as LBP, SIFT, and discrete cosine transform (DCT) [105] are extracted from the local patches around 68 facial landmark points. F-score based approach [106] is used for selection of discriminative features, and it is found that SIFT followed by DCT gives the highest accuracy. The highest recognition rate of 74.1% is obtained for 30° view.

One of the major limitations of view-wise multi-view FER methods is that these methods do not consider correlations which exist among different views of facial expressions [33, 104]. Also, the best average recognition rate obtained by these methods is less than 80%, and hence, there is a scope for improvement.

2.4.2 View-normalization for multi-view FER

The methods which normalize different views for multi-view FER exploit correlations which exist among different views of expressions. So, learning of a canonical view (best among a set of views), is a challenge for optimal multi-view classification. Very few research works have

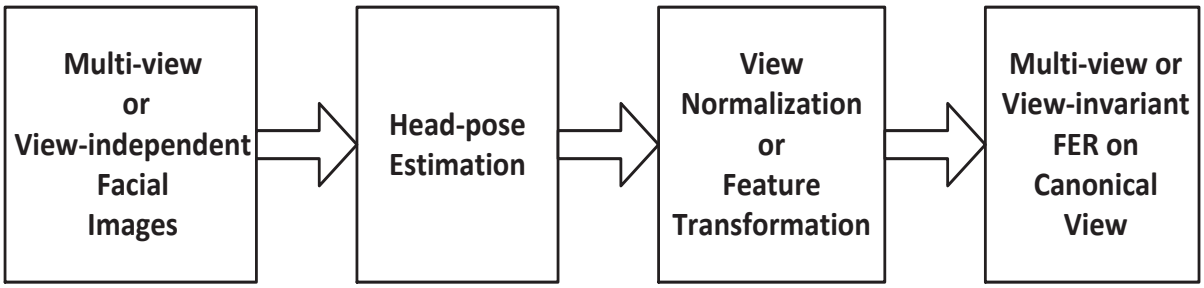


Figure 2.8: General paradigm of methods that perform view-normalization before multi-view FER.

been reported in the literature, and some of the methods are presented in [8, 33, 107, 108]. A general framework of these methods is shown in Figure 2.8, in which head-poses are estimated from different views. Subsequently, feature transformation from different non-frontal poses to a common canonical view (apriori known) is performed by leaning a set of transformation matrices between each of the non-frontal views and a canonical view. Finally, classification of multi-view facial images is performed on the learned canonical view.

The work proposed by Rudovic *et al.* [33] used shape-based geometrical features for head-pose estimation, and view-normalization. Head-pose estimation is implemented on a low-dimensional LDA-manifold, which is obtained by projecting each of the aligned face images onto a LDA space. Then, each of the views is modelled by a Gaussian function. The main contribution of their is the development of a view-normalization process, where a couple Gaussian regression-based framework is proposed. The motivation behind view-normalization in a Gaussian framework is to exploit correlations which exist between a non-frontal view and a frontal view during learning of a transformation matrix. However, it is observed from previous studies (Section 2.4.1) that frontal view may not be the optimal view for expression recognition [91,97]. Therefore, obtained classification accuracies may not be optimal. Wenming Zheng proposed a model based on group sparse reduced rank for multi-view FER [8]. This model learns association between class label vectors and synthesized multi-view facial images. For V views problem, ${}^{V}C_{2}$ number of transformation matrices has to be learned to synthesize facial images of any target view for images of a source view. Block-based texture features are used to learn mapping functions between different views. This consideration ascertain that features from several off-regions, on-regions, and on/off-regions of a face are included. A block of a

facial image is called off-region if it does not include any part of a face, on-region if a block is a part of a face, and on/off-region if it comprises of partially a part of on and off-regions. Though, the author of [8] proposed to assign a weight of 0 for off-region and 1 for on-region, the weight assignment for on/off-region is not defined. Also, allocating weights in this way may directly affect multi-view feature synthesis, as several unwanted features may be added in the observation spaces. Moreover, the major limitation of these approaches is that head-pose normalization and learning of expression classifier are carried out independently, which may affect the classification accuracy.

2.4.3 Learning optimal canonical view for multi-view FER

Figure 2.9: General paradigm of the methods [1] which search a common discriminative space for multi-view and/or view-invariant FER.

In view-normalization-based approaches discussed in subsection 2.4.2, non-frontal face images are transformed to a known apriori view (frontal view). However, several studies showed that frontal view may not be the optimal view for multi-view FER [91]. Optimal performance can even be obtained from a non-frontal face images [91,92,97]. For example, Zheng *et al.* obtained optimal view for 30°, whereas Hesse *et al.* confirmed that optimal view may lie between 30° and 60°. Hence, finding of an optimal view for multi-view FER is a challenging research problem. Recently several researchers have attempted to search an optimal discriminative space for a set of multi-view observation spaces [109–111] , and finally recognition is made on the basis of optimal common space [1]. A general paradigm of methods of this category is shown in Figure 2.9.

One of the pioneering works in finding of a discriminative common space is the work proposed by Stefanos *et al.* [1]. This method is termed as discriminative shared Gaussian process

latent variable model (DS-GPLVM). DS-GPLVM is a non-linear generative approach, which searches an optimal non-linear discriminative space. It assumes that different views of a facial image are just different manifestations of the same facial image. DS-GPLVM is a state-of-the-art multi-view FER approach which gives better performance than existing linear and non-linear learning-based methods [109–111]. The advantage of DS-GPLVM is that it rectifies two major shortcomings of existing multi-view FER methods *i.e.*, it searches an optimal common space instead of classifying images of multi-view on sub-optimal apriori known view, and secondly, the learned optimal space is discriminative. Hence, it does not require a separate classifier such as SVM to learn on the top of DS-GPLVM. Hence, a 1-NN (kNN with $k = 1$) classifier is used on learned shared space for multi-view FER.

However, discriminative nature of a common space learned by DS-GPLVM depends on a prior, which is imposed on Gaussian process [1]. In [1], a more general prior based on a notion graph Laplacian matrix is proposed [112, 113]. The LPP-based prior gives a better discriminative latent space as compared to LDA-based prior [114]. However, DS-GPLVM does not consider between-class separability of data, and hence obtained common space may not be optimally discriminative. Furthermore, samples of obtained discriminative common space may be correlated, which can further affect the classification accuracy [115].

A brief summary of all the FER methods described in the previous sections is presented below to frame the motivation and the objectives of the thesis.

2.5 Summary

In this chapter, several aspects of facial expression recognition problem (recognition from both frontal and multi-view face images) are briefly discussed. Specifically, we discussed different facial feature extraction techniques. Subsequently, methods of multi-view and view-invariant FER approaches are discussed.

Like any other pattern recognition problem, feature extraction is an important step in facial expression recognition. Existing facial feature extraction algorithms can be grouped into three main categories as shown in Figure 2.1. Face-shape-free-based methods can be either global or local. Global-based methods are less effective and more computationally expensive.

Also, dimension of the feature vector obtained globally is comparatively very high, and so "curse-of-dimensionality" is another challenge for these methods. Local feature extraction methods extract features only from a set of local salients facial regions, and hence, features are more discriminative as compared to global-based methods. Face-shape-based methods mainly employed geometrical features. However, a particular shape model is represented on the basis of facial landmark points of salient regions of a face. Also, texture features in general perform better than geometrical features. However, face model-based methods need automatic localization of landmark facial points. Also, a more versatile face model needs to be proposed to extract features from all the informative regions of a face. Literature shows that features extracted only from the salient regions of a face can give better recognition performance.

In early stage of multi-view FER research, view-wise FER methods are investigated. However, these methods have several drawbacks as discussed in Section 2.4. For view-normalization, non-frontal face images are first transformed into a canonical view, and then a classifier is learned on this canonical space to recognize expressions from multi-view face images. Generally, frontal view is considered as a canonical view, which may not be an optimal view for multi-view FER. Furthermore, learning of a canonical view and classification on the basis of canonical view are considered to be independent, which may degrade classification accuracy. DS-GPLVM is a state-of-the-art approach, which sort out two issues of view-normalization by learning a common as well as discriminative latent space. Our proposed method for multi-view is the extension of this framework.

2.6 Motivation of the Thesis

From the brief literature survey presented in this chapter, it is evident that a significant amount of work is needed for efficient recognition of different facial expressions from a frontal view. Additionally, recognizing expressions from different view points is another important issue. To combine these requirements in one algorithm is a major challenge. Accordingly, this thesis looks into several aspects concerned with extraction of most discriminative features only from the informative regions of a face, and aims at developing suitable algorithms that can take care of some of the limitations of the existing methods. The motivation behind this research work

are given below:

(i) Features extracted from the salient facial regions significantly improve the performance of FER than the existing global and local-based feature extraction methods [13, 26, 102, 103]. However, a particular salient region may not be equally important in all the facial expressions. On the other hand, salient regions obtained in [102] and [103] follow an iterative approach, and hence the extracted salient regions are not consistent even for the same facial expressions. Hence, there is a need to develop a model which can extract features only from the informative regions of a face.

(ii) The existing face models [1,33] cannot extract texture features efficiently, as these models are primarily developed for extracting geometric features. However, geometric features are very sensitive to the outliers, and they require accurate localization of facial points. Hence, there is a need to develop a more versatile face model for extraction of both geometric and texture features from a face.

(iii) Multi-view and/or view-invariant FER can be better discriminated in a common discriminative space [1, 109]. Our simple analysis shows that multi-view data inherently demonstrates multi-modal characteristics. Hence, simple LDA-based approach employed for learning a common space may fail to capture an optimal discriminative space [3,116]. So, there is a need to formulate an objective function which can learn a common discriminative space more accurately. Also, the formulation should be robust to multi-modal data.

(iv) Discriminative shared Gaussian process latent variable model (DS-GPLVM) is a non-linear generative-based approach, which can give better performance as compared to linear-based learning approaches in multi-view facial expression recognition [1]. However, DS-GPLVM employs LPP-based prior, which captures only geometric structure of the data. Hence, it neglects between-class separability of the data onto the latent space. A new prior for DS-GPLVM framework has to be formulated to address these issues.

(v) Hierarchical-based approaches can give better recognition accuracies, and so far this approach has been only employed to recognize expressions from frontal face images. So,

there is a scope to extend the hierarchical/multi-level framework to recognize expressions from multi-view face images.

2.7 Objectives

The goal of this research work is to extract facial informative regions and a discriminative common/shared space for facial expression recognition. For this, different regions of a face have to be analyzed for extracting discriminative features from the informative regions of a face. It is also important to develop a more versatile face model with the help of facial informative regions to extract both geometrical and texture features. Finding of an optimal common discriminative space for multi-view FER is another objective of this research. For this, geometric structure of the data and between-class separability of the data onto the latent space have to be simultaneously considered. The thesis also aims at handling multi-modal characteristics of multi-view data. Furthermore, multi-level framework further improves the performance of multi-view facial expression recognition algorithms.

2.8 Standard Databases Used for FER

Experimental results of our proposed methods presented in this thesis are validated on images of four standard databases namely: JAFFE [117], CK+ [118], MUG [119], and BU-3D Facial Expression (BU3DFE) databases [37]. The first three databases are used in our proposed methods presented in Chapter 3 and Chapter 4. The BU3DFE database is used in Chapter 5, and Chapter 6 for evaluating performance of our proposed multi-view FER methods.

3

Extraction of Facial Informative Regions for FER

Facial expression recognition (FER) algorithms aim to extract discriminative features of a face. However, discriminative features can be extracted only from the informative regions of a face. Most of the existing FER methods extract features from all the regions of a face, and subsequently features are stacked. This process generates correlated features among different expressions, and hence the overall accuracy is reduced. This research moves toward addressing these issues. More specifically, our approach entails extracting discriminative features from the informative regions of a face. In this view, we propose an informative region extraction model, which models the importance of facial regions based on the projections of expressive face images onto their neural face images. However in real scenarios, neutral images may not be available, and therefore we proposed to estimate a common reference image. Subsequently, a weighted-projection based local binary pattern (WPLBP) feature is derived from the informative regions of a face and their associated weights. Experimental results on standard datasets like MUG, JAFFE, and Cohn-Kanade (CK+) show the efficacy of the proposed method.

3.1 Introduction

One of the important issues in facial expression recognition is the extraction of discriminative features of a face. Apparently, discriminative features can only be extracted from the informative regions of a face [13, 26, 103]. In this chapter, the importance of different facial sub-regions are analyzed, and subsequently a set of informative regions of a face is judiciously selected on the basis of their importance in facial expressions. Consequently, a weighted-projection local binary pattern (WPLBP) feature is extracted which is subsequently employed for facial expression recognition. The extraction of informative regions of a face leads to an efficient face model, which becomes the focus of Chapter 4. Also, the informative region-based face model is utilized for multi-view facial expression recognition as discussed in Chapter 5 and Chapter 6.

Only few research works for identifying informative regions of a face have been reported in the literature till date. In general, accurate recognition of facial expressions is a tedious Computer Vision problem as a human face can show infinite number of facial expressions. Ekman *et al.* tackled this problem by defining a set of 44 Action Units (AUs), e.g., Inner brow raiser, Outer brow raiser, Upper lid raiser action units [2]. These action units are derived on the basis of change in the facial components [10, 120]. More specifically, an action unit is observed by means of texture difference between an expressive face image with an action unit and the corresponding face image without that particular action unit. This analysis becomes the fundamental basis of our proposed method. The texture difference in a particular region arises due to the movement of a region from their neutral state. In this chapter, we proposed a projection-based approach to measure the texture difference between an expressive face image and a neutral face image. On the basis of texture analysis, different informative regions of a face are extracted to obtain discriminative features. Figure 3.1 illustrates texture difference between few sub-regions of "anger" and "happy" expressions with respect to a neutral face image.

Extensive research has been done towards facial expression recognition [7, 51, 60, 62, 121], where features are either extracted holistically or locally. In holistic feature-based methods, either intensity values or transformed version of intensity values are vectorized to form a feature vector for the entire face image [3, 55]. All these methods generally suffer from the problem

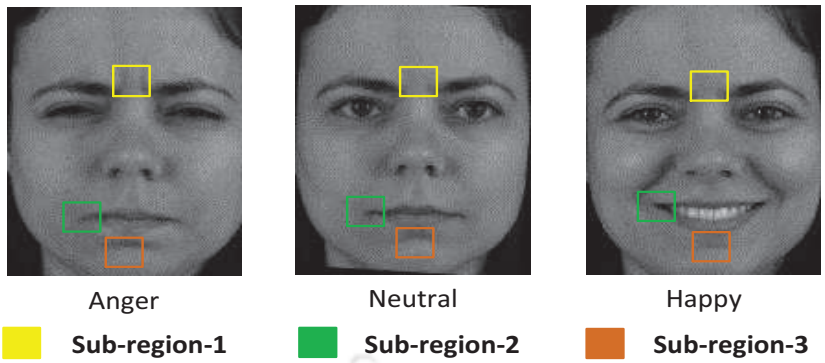


Figure 3.1: An example showing texture difference in expressive facial images with respect to a neutral face image.

of curse of dimensionality. So, another requirement of these methods is to reduce the dimensionality of the extracted feature vector. The popular dimensionality reduction techniques are principle component analysis (PCA), linear discriminant analysis (LDA), local Fisher discriminant analysis (LFDA), independent component analysis (ICA) [55], etc. PCA captures directions along which data has large variability, but it may not be suitable for accurate classification, as it also maximizes within-class covariance of the data [3]. The optimization criteria of LDA resolved the above problem, but it also suffers when the distribution of feature vectors comprises of a number of local maximas. ICA requires higher-order statistics of the data to find optimal independent components, which is computationally more complex than PCA and LDA. Rahulamathavan *et al.* and Zhang *et al.* adopted LFDA to extract holistic features of an image [3,122]. Although it performs better as compared to PCA and LDA, they got 95.25% and 94.37% of recognition rates on MUG and JAFFE datasets, respectively.

Gabor wavelet is another popular texture descriptor used by many of the researchers for facial expressions recognition [69, 70, 123, 124]. In these methods, bank of Gabor filters are created by changing scaling, and orientation parameters. Thus, a set of filtered images can be obtained from a single input image. Concatenating intensity values of these filtered images results a feature vector of very high dimension corresponding to a particular expression, which overlaid to the curse of dimensionality problem to some extent. The shortcomings of Gabor filter are minimized by selecting a subset of discriminative features [125]. The average recognition rates for six class expression classifications obtained by Gabor features are reported between

90%-94%.

In local feature-based methods, the entire face is initially divided into several sub-regions [41, 62–64, 126]. The number of sub-regions depends on the size of the original image and the size of the sub-blocks. Commonly used block sizes are 8×8, 12×12, 16×16, etc. Subsequently, features like LBP are extracted from each of the sub-regions *i.e.*, histogram of LBP codes of each of the sub-regions are obtained, and finally they are concatenated sequentially to obtain the feature vector for the entire face image. Literature shows that local feature-based approaches are superior as compared to the holistic-based approaches. Also, global features are unable to capture the dynamics of local muscular movements of the facial sub-regions. On the other hand, sequential concatenation of local feature vectors of the entire face image leads to inter-correlated features among different classes of expressions. For example, texture pattern of R^{th} region in some expressions (let's say surprise) is highly correlated with the texture of R^{th} region of anger and happy expressions. Due to this, confusion occurs between happy, surprise, and anger expressions, which finally reduces the overall accuracy. In case of video, Zhao *et al.* used volumetric LBP on stacked frames of a video [62]. LBP is applied on three selected orthogonal planes *i.e.,* XY, XT, and YT planes, which is known as LBP-TOP. However, LBP-TOP does not show the importance of the facial regions. The importance of different facial sub-regions is analyzed by Kotsia *et al.* [41]. This was done by manually hiding different parts of a face such as eyes, mouth, and half of the face during the facial expressions recognition. Khan *et al.* conducted a psycho-visual experiment to extract the salient regions of a face [13]. But, these experiments could not provide information of degree of importance of different sub-regions of a face, such as the relative importance of mouth and eyes regions. The methods proposed by Zhong *et al.* and Liu *et al.* employed multi-task sparse learning (MTSL) based approaches, which rely on solving a constraint optimization problem to get active regions of a face from a set of pre-defined sub-regions [102, 103]. However, these methods depend on several tuning parameters such as regularizing parameters, step size, and so on. Also, several parameters have to be manually initialized to get the optimum values of the selected parameters. Moreover, following few important points are noticed which are not sufficiently addressed in the existing facial expression recognition algorithms.

- **Degree of importance of subregions:** Kotsia *et al.* and Khan *et al.* experimentally showed that eyes and mouth regions of a face are more crucial for FER, but they could not give justifications of their claim [13, 41]. Also, they could not extend their theory for all the regions of a face.

- **Sequential concatenation of feature vectors of all the facial sub-regions:** Although this arrangement is quite simple and straightforward, but it is not so effective for classification due to the presence of correlated features among different expression classes. This issue is not much addressed beyond PCA and LDA.

- **Relative analysis of expressions:** Since all the facial expressions are resulted due to the change in the neutral expression and so, the analysis of the expressive images with respect to a neutral image could be an important basis for the classification of different expressions.

To address all the above issues, a simple model-based feature extraction method is proposed in this chapter. In this work, we aim to improve the performance of FER by deriving the discriminative features from the well-known LBP texture features. Projection-based analysis is proposed to determine the informative regions of a face for an efficient FER. Our main contributions of this chapter are highlighted as follows.

- A simple mathematical model is proposed which estimates the importance of different facial sub-regions based on projection analysis. The proposed model is used to select informative regions of a face based on projection errors and so, it is termed as informative region extraction (IRE) model.

- The IRE model is based on neutral images which may not be available in many practical scenarios, and therefore we proposed to estimate a common reference image using Procrustes analysis.

- Subsequently, information gained from proposed IRE model is used to enhance the discriminative property of LBP features, and thereby, mis-classifications among different classes of expressions are reduced. Since the proposed feature is derived based on the projection analysis and the LBP, we termed it as weighted projection-based LBP (WPLBP).

Experimental results on different standard datasets like MUG, JAFFE, and CK+ show that the proposed WPLBP feature significantly gives better performance as compared to the existing feature extraction approaches. The proposed method is described in more detail in the following Sections.

3.2 Proposed Framework

General framework of our proposed IRE model and feature extraction algorithm comprises of the following steps: 1) Extraction of local binary pattern of entire face; 2) Projection analysis of LBP features of expressive face images to their corresponding neutral face images; 3) Modelling of reference image using Procrustes analysis; 4) Feature extraction only from the selected regions of a face obtained from IRE model; 5) Weight allocation to the selected features based on their importance; and 6) Facial expression recognition only by using the features extracted from the selected facial regions.

3.2.1 Proposed informative region extraction (IRE) model

The objective of this step is to select a set of informative regions from a face. In this view, our proposed IRE model first implement first three steps of the proposed framework to model the importance of each of the sub-regions of a face. Subsequently, defining informative regions on the basis of their importance and extracting features only from the informative regions are discussed in Section 3.2.2 [1].

3.2.1.1 Local Binary Pattern

local binary pattern (LBP) was originally proposed in [127], and it is a well known texture descriptor widely used in FER. This is because of the following reasons: 1) LBP is computationally simple and easy to apply, 2) It is a non-parametric descriptor, and 3) It can also handle monotonic illumination variations. In basic LBP, each pixel of an image is compared with their eight neighborhood pixels of a 3×3 block as illustrated in Figure 3.2. For example, let us

[1]This work has been published in *IET Computer Vision 2016* (Refer item 1 in Page 143 for details)

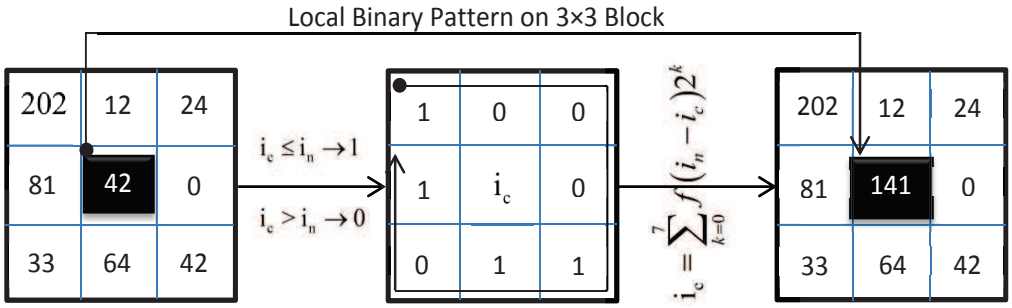


Figure 3.2: Basic LBP operation on a 3×3 image block.

consider $i_n : n = 1, 2, ..., 8$ be the eight neighborhood pixels of the center pixel i_c. In this case, if $i_n \geq i_c$ then $i_n = 1$, else $i_n = 0$, which is an intermediate step in finding the LBP features. Finally, binarization is done by concatenating these i_n from the left-top corner in the clockwise direction *i.e.*, 10001101, and subsequently the corresponding decimal equivalent value *i.e.*, 141 is assigned to the center pixel i_c, which is known as LBP code. This step is repeated for all the pixels of the image, and finally the corresponding histogram of the LBP codes is considered as the local texture features of the original image.

Further several variants of LBP features are proposed such as extended LBP, which is a generalized LBP for any radius R, and also for any number of sampling points P [75, 128]. Uniform LBP is one of the variants of LBP, which accounts a set of LBP codes whose uniformity (U) measure is less than and equal to 2. Uniformity is defined as the number of 0/1 transitions in the LBP code. For example: U value for 10001101 is 4 since there are exactly four 0/1 transitions in the pattern. In the rest of this chapter, uniform LBP features will be used for our analysis. Mathematically, LBP code/decimal representation of the center pixel i_c at (x_c, y_c) is given by:

$$LBP_{P,R}(x_c, y_c) = \sum_{k=0}^{P-1} f(i_k - i_c)2^k \tag{3.1}$$

where, i_k and i_c are gray-level intensity values of the k^{th} neighboring pixel and the central pixel respectively. The function $f(x)$ in Eqn. (3.1) is defined as:

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

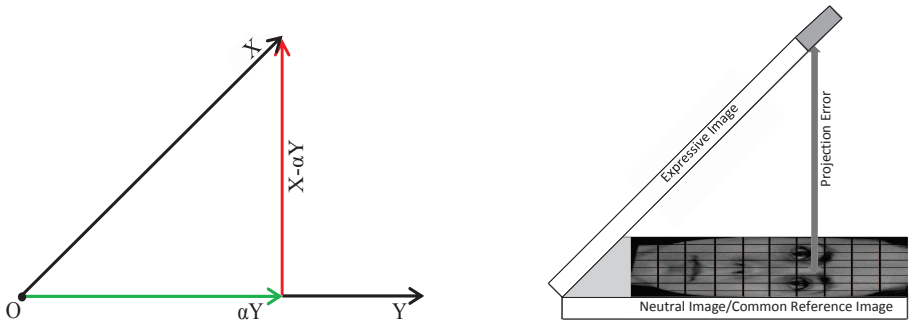


Figure 3.3: Basic operation of projection analysis [Left] and pictorial view of projection analysis in which LBP of i^{th} block is projected on the corresponding LBP of i^{th} block of the reference image [Right].

3.2.1.2 Projection analysis

This is the basis of our proposed approach. Let $\mathbf{X}$ and $\mathbf{Y}$ be the two d-dimensional vectors as shown in Figure 3.3. Let the projection of vector $\mathbf{X}$ on $\mathbf{Y}$ is $\alpha\mathbf{Y}$. Then, the objective is to find the best α which minimizes the projection error given by $\|\mathbf{X}-\alpha\mathbf{Y}\|_2$. The magnitude of error vector is minimum iff the error vector is perpendicular to the data vector *i.e.*, the inner product between the error vector and the data vector must be zero. Mathematically, it can be shown as follows:

$$(\mathbf{X}-\alpha\mathbf{Y})\perp\mathbf{Y} \Leftrightarrow \langle(\mathbf{X}-\alpha\mathbf{Y}),\mathbf{Y}\rangle = 0 \tag{3.3}$$

$$\Rightarrow \alpha=\frac{\langle\mathbf{X},\mathbf{Y}\rangle}{\langle\mathbf{Y},\mathbf{Y}\rangle} \tag{3.4}$$

where, $\langle\cdot,\cdot\rangle$ denotes the inner product. Thus, minimum error magnitude is given by Eqn. (3.5).

$$\text{Minimum error magnitude} = \left\|\mathbf{X}-\frac{\langle\mathbf{X},\mathbf{Y}\rangle}{\langle\mathbf{Y},\mathbf{Y}\rangle}\mathbf{Y}\right\|_2 \tag{3.5}$$

Let $\text{A} = \{\text{Y}_i | i = 1, 2, ..., N\}$ and $\text{B} = \left\{\text{X}_{ij}^k | i = 1, 2, ..., N; j = 1, 2, ..., l; k = 1, 2, ..., C\right\}$ be the set of neutral and expressive facial images respectively, where Y_i is the i^{th} neutral image and X_{ij}^k represents k^{th} class of j^{th} level of expressive image belonging to i^{th} subject in the training dataset. Each image in the training dataset is divided into λ number of subregions. We define

the parameter λ as:

$$\text{Number of sub} - \text{regions}\,(\lambda) = \frac{\text{Size of the image}}{\text{Size of the block}} \tag{3.6}$$

In our case, if the image size is $P \times Q$, then the corresponding block size will be $0.1P \times 0.125Q$. This consideration ascertains that the number of sub-regions always remain 80 regardless of the size of an image.

Let us consider a particular facial expression belonging to class k, a particular subject i, and the corresponding reference image are available. Let $\mathbf{x}_{jm}^{ki}$ be the LBP features of m^{th} sub-region, $m \in \{1, 2, ..., \lambda\}$, of j^{th} level of expressive image of i^{th} subject belonging to k^{th} class. Also, $\mathbf{y}_{im}$ is the LBP features of m^{th} sub-region of i^{th} neutral image. As explained earlier, any expressive images are resulted due to the movements of different facial sub-regions of a neutral face image. Human facial muscles undergo changes for showing different facial expressions, and these changes are emerged as the variations of the texture patterns of different sub-regions of a face. The texture variations of different sub-regions with respect to the texture pattern of a neutral face image indirectly give an indication of corresponding facial expressions. So, there will be a change in distribution of LBP features between the blocks of an expressive face image and the corresponding blocks of a neutral face image.

$$\begin{aligned} e_{jm}^{ki} &= \left\| proj_{\mathbf{y}_{im}} \mathbf{x}_{jm}^{ki} \right\|_2 \\ &= \left\| \mathbf{x}_{jm}^{ki} - \frac{\left\langle \mathbf{x}_{jm}^{ki}, \mathbf{y}_{im} \right\rangle}{\left\langle \mathbf{y}_{im}, \mathbf{y}_{im} \right\rangle} \mathbf{y}_{im} \right\|_2 \end{aligned} \tag{3.7}$$

Thus, m^{th} sub-region of an expressive image and the corresponding m^{th} sub-region of a reference image will have a projection error as shown in the Figure 3.3 [Right], and it is given by Eqn. (3.7). Similarly, Eqn. (3.7) can be used to obtain the projection errors from all other blocks of an expressive image to the corresponding blocks of a reference image by varying $m = 1, 2, ..., \lambda$. So, the projection errors are directly related to the importance of different sub-regions *i.e.*, more is the projection error more is the importance of a facial sub-region, and vice-versa. In other words, a sub-region conveys more information if it gives more projection error. Apparently, this

analysis paves the way to extract the informative facial regions from a number of pre-defined sub-regions.

Let $\mathbf{E}$ be an error matrix whose elements of each row is e_{jm}^{ki}. In this, e_{jm}^{ki} represents the projection error of m^{th} block of an expressive image to the corresponding block of a neutral image belonging to k^{th} expression. Now, suppose that each raw vector $\mathbf{e}$ of error matrix $\mathbf{E}$ is drawn independently from a Gaussian distribution, whose mean is $\mathbf{w}_p$ and covariance matrix is $\mathbf{\Sigma}$. Then, the joint distribution of all the observations of error matrix can be written as the product of the marginal distributions, and hence it can be expressed as:

$$p\left(\mathbf{E}|\mathbf{w}_p, \mathbf{\Sigma}\right) = \prod_{n=1}^{T} p\left(\mathbf{e}_n|\mathbf{w}_p, \mathbf{\Sigma}\right) \tag{3.8}$$

where, T represents the total number of samples used in this modeling, which can be obtained by multiplying C (number of class of expressions), l (number of levels per expression), and N (total number of subjects). Here, our main objective is to estimate the mean of the Gaussian distribution, which can be obtained by maximizing the log likelihood of Eqn. (3.8) with respect to mean vector $\mathbf{w}_p$. The optimal likelihood solution for the mean vector is given as follows:

$$\left(\mathbf{w}_p\right)_{opt} = \frac{1}{T}\sum_{n=1}^{T} \mathbf{e}_n \tag{3.9}$$

Let $\left(\mathbf{w}_p\right)_{opt} = \begin{bmatrix} w_1 & w_2 & \cdots & w_\lambda \end{bmatrix}$ be the mean row vector, where each w_m; $\forall m \in [1, \lambda]$ represents the average importance of m^{th} block in all the facial expressions. Average mean projection error is higher for a block which conveys more information for all the considered facial expressions. Thus, $\left(\mathbf{w}_p\right)_{opt}$ gives the distribution of importance of all the sub-regions. As stated earlier, different local muscular regions of a face undergo gradual changes for showing a facial expression. An expression generally starts from a neutral state, and after successive affine transformations of some facial sub-regions caused by muscular deformations, final peak level of an expression is obtained. So, all the different levels of expressions stating from mid to peak contribute to error distribution of the proposed model, *i.e.*, all the levels of an expression have an influence on the error distribution. That is why, the modeling of informative regions

of a face only with the help of peak level expressions may not convey the actual information of the intermediate deformation of facial subregions. Hence, it is important to consider different levels of expressions in our proposed model.

3.2.1.3 Extraction of a common reference image

Our proposed projection-based analysis as discussed on Section 3.2.1.2 rely on neutral images, which may not be readily available. Hence, we propose to generate a common reference image from frontal view of neutral face images. Common reference image (CRI) is derived by considering shape and texture variations of neutral face images. These variations in the neutral images arise due to change in shape and texture of different subjects. The popular approach to reduce these variations is to align all the neutral images onto a common frame of reference. The procrustes analysis framework is proposed here to solve this particular optimization problem [129]. The proposed procrustes-based alignment approach uses block-based features to align different neutral images, and hence, it does not require manual marking of landmark points on each of the considered images. Additionally, our method is free from a training step, which is an essential step in landmark-based alignment approaches [88, 89, 130] [2].

Let $\{A_i | i = 1, 2, ..., N\}$ be a set of N neutral images, where each image is partitioned into λ number of sub-regions. Let us consider for a particular A_i^{th} neutral image, the uniform LBP features for each of the sub-regions are extracted, and hence the image A_i is represented as:

$$A_i = (\mathbf{y'}_{i1}, \mathbf{y'}_{i2}, ..., \mathbf{y'}_{im}, ..., \mathbf{y'}_{i\lambda}) \, , \; i = 1, 2, ..., N \tag{3.10}$$

where, $\mathbf{y'}_{i\lambda}$ indicates LBP feature vector of λ^{th} sub-region. Then, the objective is to find the best reference image A_u (common reference image) which minimizes the mean square error between the transformed neutral image and the common reference image. Mathematically, it can be written as:

$$J(\theta) = \|T_\theta(A_i) - A_u\|^2 \tag{3.11}$$

where, $T_\theta(\cdot)$ indicates affine transformation matrix with θ as a parameter. The Eqn. (3.11)

[2] This work has been published in *IEEE Conference on Recent Advances in Intelligent Computational System 2016* (Refer item 3 in Page 143 for details)

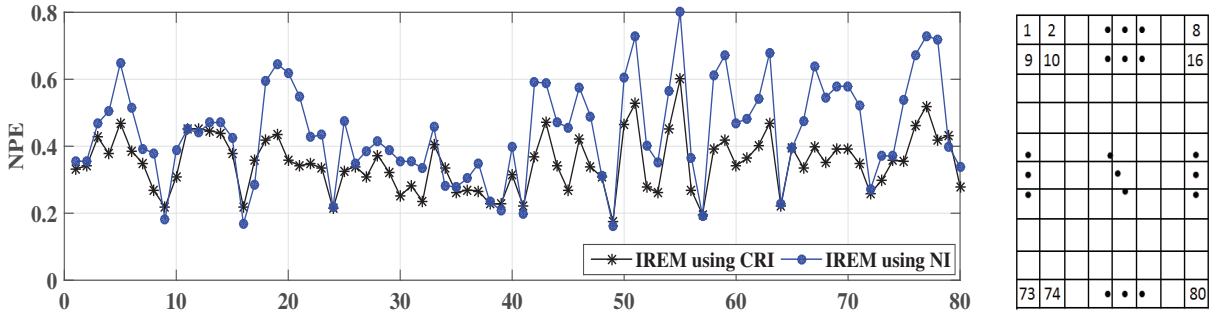


Figure 3.4: Graphical representation of the proposed IRE model for determining the importance of different sub-regions using neutral images (NI) and a common reference image (CRI) respectively. The horizontal axis of left graph shows block indices as shown in the right figure, and the vertical axis shows normalize projection errors (NPE) of the respective blocks.

can be solved by using an iterative approach, which is described as follows:

(i) Translate each neutral image A_i such that the mean is at the origin.

(ii) Find the mean $m = \frac{1}{N}\sum_{i=1}^{N} A_i$, and evaluate $S_i = A_i - m,\ \forall i$.

(iii) Initialize $A_u = S_r$, where r is any positive integer lies between 1 and N.

(iv) Align each S_i with A_u.

(v) Re-estimate mean A_u^{new} from the aligned one.

(vi) If $\|A_u^{new} - A_u\| < \in$, $A_u = A_u^{new}$ and stop, else go to step 3.

To enhance our claim, we also analyze the importance of each of the sub-regions based on projection analysis of expressive images with respect to the proposed reference image A_u. The pictorial representation of projection errors for all the considered blocks is shown in the left of Figure 3.4, where the corresponding block indices are shown in the right of Figure 3.4. Projection errors are relatively high for informative blocks as compared to the non-informative blocks. Thus, distribution of projection errors over the blocks can give important information for selecting the informative regions of a face for a particular facial expression.

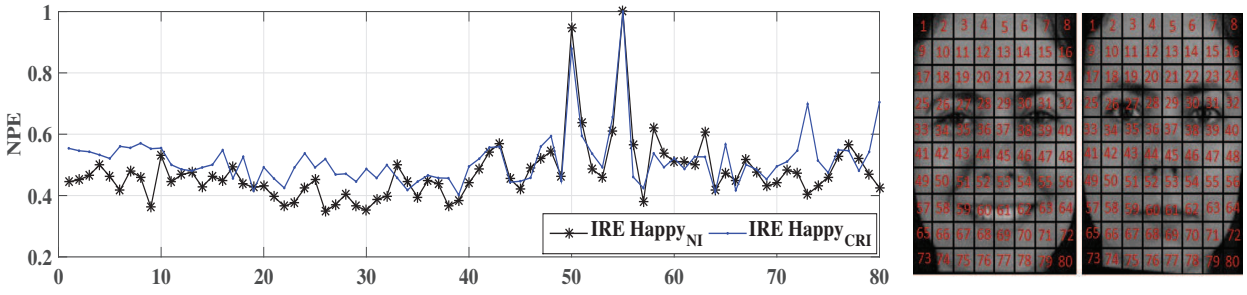


Figure 3.5: Left figure shows the distribution of normalize projection error (NPE) for different blocks in happy expression, whereas right two images show a sample pair of happy and neutral images.

3.2.2 Weighted projection-based LBP (WPLBP)

The proposed weighted projection-based LBP (WPLBP) is derived from LBP features extracted only from the informative regions of a face. Also, the extracted features are weighted on the basis of their importance (derived from projection analysis). The proposed feature is more logical as a particular sub-region has different importances in different expressions. This is analyzed in Section 3.2.2.1, and subsequently we proposed our feature, which is subsequently used for facial expression recognition.

3.2.2.1 Expression specific sub-region analysis

As we model the importance of facial regions with the mean of a Gaussian distribution, which might not reveal the correct sequence of importance of sub-regions for all the expressions. This is due to the fact that the regions which are informative in one expression might not be that much of informative in other expressions. One way to solve the above problem is to model each individual expression separately.

Let us consider, the modified projection error matrix $\mathbf{E}^k$ for a particular class k, where $k \in [1, C]$. The corresponding joint probability and the optimum likelihood solution for the mean vector that maximizes the log likelihood of the joint density function are given by Eqn. (3.12) and Eqn. (3.13) respectively.

$$p\left(\mathbf{E}^k|\mathbf{w}_p^k, \mathbf{\Sigma}_k\right) = \prod_{n=1}^{l.N} p\left(\mathbf{e}_n^k|\mathbf{w}_p^k, \mathbf{\Sigma}_k\right) \tag{3.12}$$

$$\left(\mathbf{w}_p^k\right)_{opt} = \frac{1}{l.N}\sum_{n=1}^{l.N}\mathbf{e}_n^k \tag{3.13}$$

where, $(\mathbf{w}_p^k)_{opt}$ for $k = 1, 2, \cdots, 6$ represents models for different facial expressions (happy, surprise, fear, anger, disgust, and sad). For example, region of importance graph for happy expression is shown in Figure 3.5, where vertical axis represents the normalize projection error for happy IRE model. It is clear that only few sub-regions nearby mouth are more informative in happy expressions. Similar analysis is done for all other expressions, and it is observed that each of the informative sub-regions have different importance in different expressions. The importance of a few selected sub-regions marked in Figure 3.6 [Left] are shown in the right part of the same figure. For example, sub-region ($R = 51$) (indicated in Figure 3.6 [Left]) is more important in happy and disgust, whereas it is comparatively less importance in fear, anger, and sad expressions.

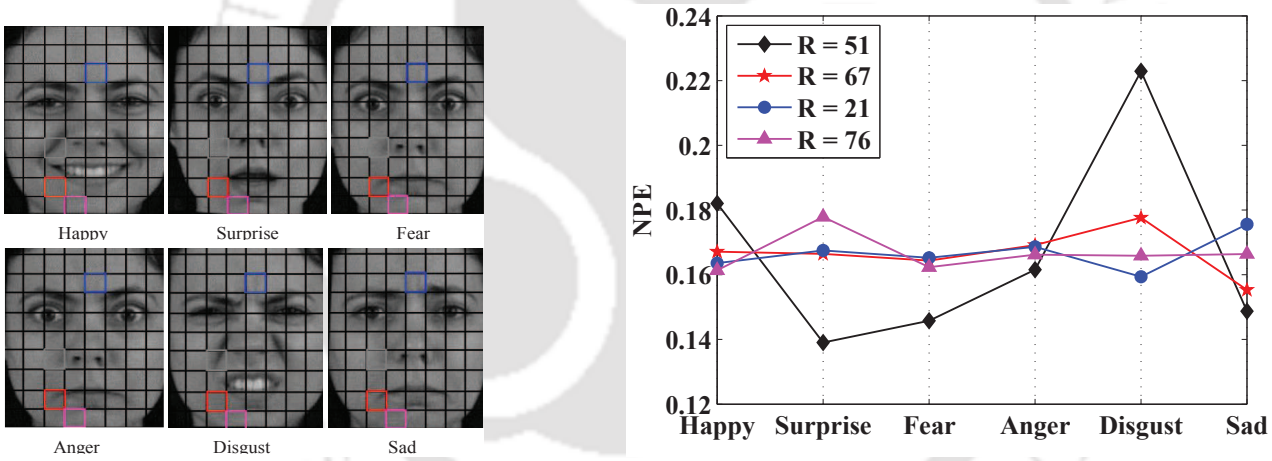


Figure 3.6: Figures show a few marked sub-regions and their importance in different expressions. The [Left] figure indicates marked sub-regions for R = 51, 67, 21, and 76 and their corresponding importance in different expressions are shown in [Right] graph.

3.2.2.2 Feature selection and weight allocation

The aim of this step is to select most discriminative features from the original feature vector. As explained earlier, discriminative features can be extracted only from the informative regions

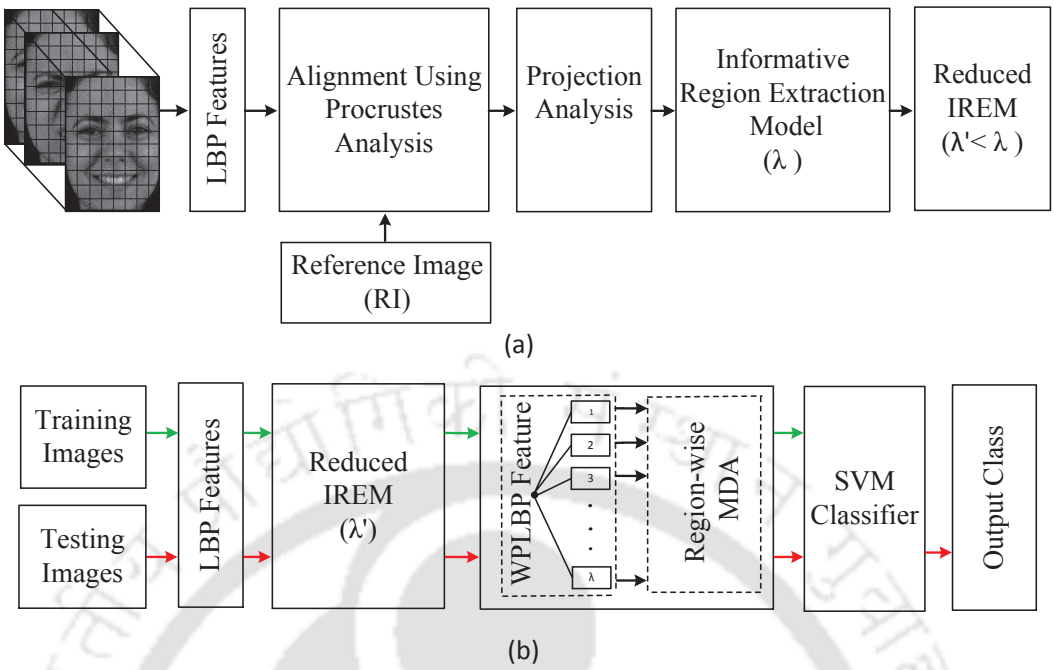


Figure 3.7: (a) Block diagram of our proposed IRE model for determining the importance of facial subregions, (b) block diagram of the proposed WPLBP approach for facial expressions classification, in which dimensionality reduction is done for the selected regions separately.

of a face. In our method, a set of facial sub-regions which corresponds to informative blocks are judiciously selected based on their projection errors (Eqn. (3.9)), and we termed it as a reduced-IRE model. Analysis of WPLBP shows that error distribution for each class of expression is different, which indicates that each of the sub-regions has different importance in different expressions. Hence, in our proposed WPLBP, feature vector of the selected sub-regions of each class is weighted by their own model parameters given by Eqn. (3.13). For each training and testing images, these weights are estimated by projecting the features of the selected sub-regions of expressive images to the corresponding sub-regions of the proposed reference image A_u. Thus, a particular block has different importances for different expressions.

Let X_i^k be the reduced feature vector for i^{th} subject belonging to k^{th} class which is obtained by the reduced projection model given in Eqn. (3.14). Let X_i^k be projected on i^{th} reference image and the corresponding error vector is given by Eqn. (3.15), where $\mathbf{x}'_{i1}$ and w_1^k are the LBP features and projection error corresponding to the most informative region. Similarly, $\mathbf{x}'_{i\lambda'}$

and $w^k_{\lambda'}$ are the LBP features and weight corresponding to the least informative region.

$$\mathrm{X}^k_i = \begin{bmatrix} \mathbf{x}'_{i1} & \mathbf{x}'_{i2} & \cdots & \mathbf{x}'_{i\lambda'} \end{bmatrix}' \tag{3.14}$$

$$\mathbf{w}^k_p = \begin{bmatrix} w^k_{i1} & w^k_{i2} & \cdots & w^k_{i\lambda'} \end{bmatrix} \tag{3.15}$$

$$\mathrm{X}^k_{i,weighted} = \begin{bmatrix} w^k_{i1}\mathbf{x}'_{i1} & w^k_{i2}\mathbf{x}'_{i2} & \cdots & w^k_{i\lambda'}\mathbf{x}'_{i\lambda'} \end{bmatrix}' \tag{3.16}$$

Thus, the proposed WPLBP features can be obtained by using Eqn. (3.16), which is logically more realistic. Experimental results show the efficacy of the proposed WPLBP features.

3.2.2.3 Facial expression classification with reduced number of features

The overall dimension of WPLBP feature vector is $\lambda \times \dim(LBP)$, where $\dim(\cdot)$ indicates the dimension of a LBP feature vector, where λ is the number of selected sub-regions. Multiple discriminant analysis (MDA) is performed for each of the sub-regions separately to further reduce the dimensionality of selected feature vector [76]. For a particular sub-region λ, feature vectors of all the samples are projected from their original higher dimensional space to a lower dimensional feature space, so that the inter-class separability is maximized and the intra-class separability is minimized. In general, MDA gives $C-1$ eigenvectors for a C-class problem, and thus the number of features for each of the sub-regions is reduced to $C-1$. Therefore, for λ number of selected sub-regions, the original dimension $\lambda \times \dim(LBP)$ of WPLBP feature reduces to $\lambda \times (C-1)$. Advantage of our proposed WPLBP approach is that it can capture most of the discriminative information of a facial expression. The block semantic representation of our proposed IRE model and classification scheme using WPLBP are illustrated in Figure 3.7(a) and Figure 3.7(b) respectively. Finally, the extracted features are used to train a supervised SVM classifier followed by testing of unlabeled facial images.

3.3 Performance Evaluation

The proposed facial expression recognition method is tested on three comprehensive benchmark facial expressions databases: MUG [119], JAFFE [117] and CK+ [118]. In our proposed IRE

model, neutral images are used as reference images, and so these images are not included for classification. The remaining six prototypic facial expressions *i.e.*, images of happy, surprise, fear, anger, disgust, and sad expressions are collected from the above databases. We used altogether 2236 (2184 expressive + 52 neutral) images of MUG database, where each subject has $5 \sim 8$ expression levels per class. The JAFFE database has 213 image sequences posed by 10 Japanese women. The database has $2 \sim 4$ expression levels per subject. All the images of this database are considered for our analysis. Experiments are also performed on CK+ database, which is a widely used benchmark database. In our analysis, 236, 184, 143, 130, 161, and 224 images from CK+ database for happy, surprise, fear, anger, disgust, and sad expressions are used respectively. The performance of our proposed method is validated for the following parameters:

- Separability analysis of datasets for different number of selected informative sub-regions.
- Selection of number of facial regions for expression recognition, and
- Performance analysis for different image resolutions.

Table 3.1: Specifications of different databases used in our experiments

Database	Subjects	Images per Class	Images pspc ⋆	Resolutions			
MUG	52	$350 \sim 370$	$5 \sim 8$	160×128	120×96	80×64	60×48
JAFFE	10	$28 \sim 32$	$2 \sim 4$	160×128	120×96	80×64	60×48
CK+	52	$200 \sim 260$	$5 \sim 7$	160×128	120×96	80×64	60×48

⋆ pspc: per sample per class.

For modelling informative regions, samples from all the datasets are used. These samples consist of neutral images and their respective $3 \sim 5$ levels of expressive images. A set of 50 subjects are selected, in which 25, 5, and 20 are taken from MUG, JAFFE, and CK+ databases respectively. Thus, a total of $220 \times 6 + 50 = 1370$ samples are used, which comprises of $25 \times 5 + 5 \times 3 + 20 \times 4 = 220$ samples per class and 50 reference images. The face part of all the samples are cropped manually and normalize them to become the same size. In addition, procrustes analysis is used to align expressive images with their neutral images so that miss-matching error is minimum. Then, LBP features of each expressive image is projected onto

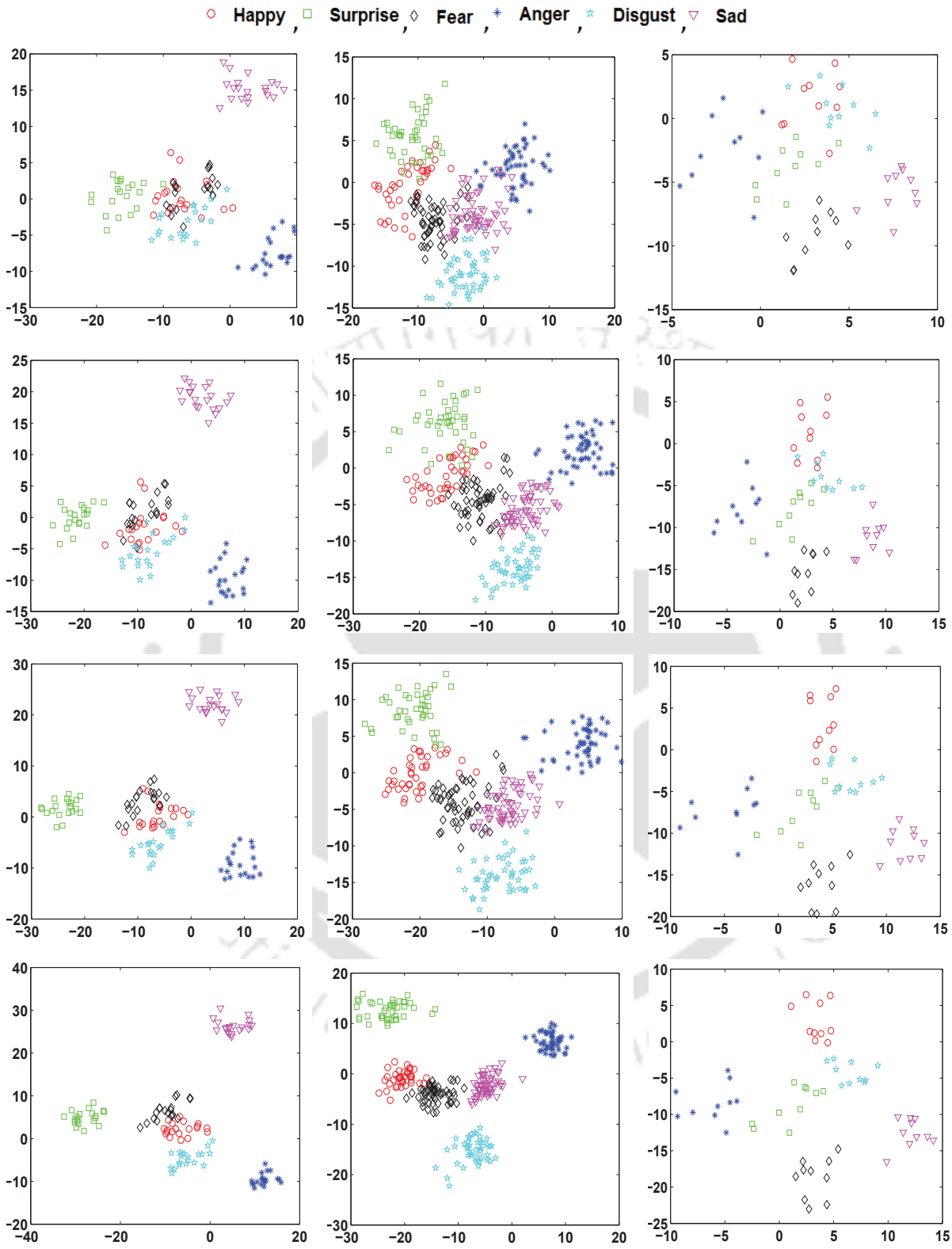


Figure 3.8: Distribution showing class separability of D/N/P training samples with different numbers of informative regions. In our representation of D/N/P, D indicates the selected database, N is the number of samples per expression, and P is the number of sub-regions. For first, second, and third columns, D/N/P represents MUG/20/P, CK+/40-50/P, and JAFFE/10/P respectively. This evaluation is done for P = 10, 20, 30, and 45, which are sequentially shown from the top row to the bottom row.

their respective reference image. The magnitude of projection error for a particular image block gives the information of the deviation of texture patterns between a neutral state face and a expressive face. Large deviation indicates that a particular facial sub-region plays a significant role in a particular facial expression compared to the sub-regions having less projection errors. This analysis is done for all image resolutions as mentioned in Table 3.1. The block size of 16×16, 12×12, 8×8, and 6×6 are judiciously chosen to maintain the number of sub-regions same irrespective of the different image resolutions. The performance of our proposed method is also analyzed with respect to the number of selected regions. As per our assumption, 57 regions ($\lambda' = 57$) are selected from the total of ($\lambda = 80$) regions for our experimentation. For this, 57 informative regions are selected on the basis of magnitude of projection errors. The regions having error magnitude less than 3.6 are not selected as informative regions, as their contributions in a facial expression is marginal. For a particular selected region, experiments are repeated for 10 times, and then average accuracy is calculated from the confusion matrix. If c_{ij} is the element of a confusion matrix, where $(i, j) \in \{1, 2, ..., C\} \times \{1, 2, ..., C\}$, then the accuracy is given by Eqn. (3.17). Also, Krippendroff's alpha (KA) coefficient is used to measure the inter-rater reliability of the proposed model [131]. The separability analysis of our proposed WPLBP feature is shown in Figure 3.8. It is evident from this analysis that discriminative nature of the proposed feature increases with the number of selected regions. Experimental results show that our proposed method outperforms the state-of-the-art approaches [3, 60, 123].

$$Accuracy = \frac{\sum_{i=1}^{C} c_{ii}}{\sum_{i,j=1}^{C} c_{ij}} \times 100\% \tag{3.17}$$

3.3.1 Experiments on MUG database

MUG database is one of the popular datasets used for facial expression recognition. In our experiment, images from 10-15 subjects out of 52 subjects are used for training. For each class, 5 levels of expressions are considered for each of the subjects. For testing, remaining images of 37 subjects which include $37 \times 7 \times 6 = 1554$ images are considered. The proposed algorithm is

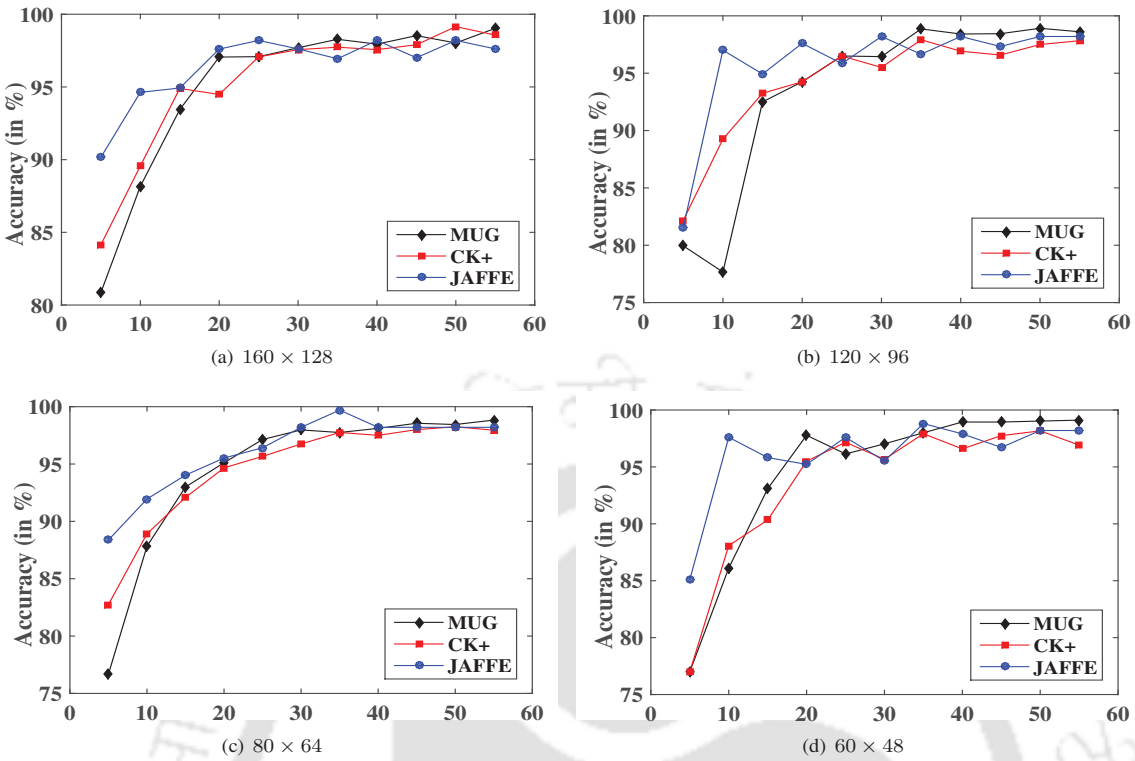


Figure 3.9: Performance of the proposed method for different image resolutions. The horizontal axis represents the number of selected informative sub-regions.

tested for different parameters as listed in Table 3.1. Figure 3.9 shows the detailed analysis of our proposed work on MUG, JAFFE, and CK+ databases. The horizontal axis represents the number of informative regions considered, and the corresponding recognition rates are shown in the vertical axis. From this illustration, it is clear that if the selected regions is 35 or more, the average accuracy of the proposed algorithm is more than 96%. The graphs also indicate that the recognition rates do not degrade significantly up to the image size of 60 × 48. Confusion matrices for different image resolutions of MUG dataset obtained using our proposed WPLBP and SVM classifier are shown in Tables 3.2, 3.3, 3.4, and 3.5.

3.3.2 Experiments on JAFFE and CK+ databases

Similar experiments are performed on JAFFE and CK+ databases. The size of the CK+ dataset is larger than that of the JAFFE dataset, but the expressions of CK+ database are more artificial as compared to the JAFFE and MUG databases. The images of CK+ dataset

Table 3.2: Confusion matrix for 160 × 128 MUG dataset images

	Happy (%)	Surprise (%)	Fear (%)	Anger (%)	Disgust (%)	Sad (%)
Happy (%)	98.33	0	0	0	1.67	0
Surprise (%)	0	95.67	4.33	0	0	0
Fear (%)	0	0.67	96.0	3.33	0	0
Anger (%)	0	0	0	100	0	0
Disgust (%)	0	0	0	0	100	0
Sad (%)	0	0	0	0.33	0	99.67
	Average accuracy = 98.03%					

Table 3.3: Confusion matrix for 120 × 96 MUG dataset images

	Happy (%)	Surprise (%)	Fear (%)	Anger (%)	Disgust (%)	Sad (%)
Happy (%)	99	1	0	0	0	0
Surprise (%)	0	97	3	0	0	0
Fear (%)	0	0	96.67	3.33	0	0
Anger (%)	0	0	0	98.67	0	1.33
Disgust (%)	0	0	0.67	0	99.33	0
Sad (%)	0	0	0	1	0	99
	Average accuracy = 98.44%					

Table 3.4: Confusion matrix for 80 × 64 MUG dataset images

	Happy (%)	Surprise (%)	Fear (%)	Anger (%)	Disgust (%)	Sad (%)
Happy (%)	97.33	0	0	1	1.67	0
Surprise (%)	0	98.33	1.67	0	0	0
Fear (%)	0	2.33	94.67	2	0	1
Anger (%)	0	0	1	95.67	0	3.33
Disgust (%)	0	0	0.67	0	99.33	0
Sad (%)	0	0	0	0.67	0	99.33
	Average accuracy = 97.14%					

Table 3.5: Confusion matrix for 60 × 48 MUG dataset images

	Happy (%)	Surprise (%)	Fear (%)	Anger (%)	Disgust (%)	Sad (%)
Happy (%)	98	0	0	0	1	1
Surprise (%)	0	97.67	2.33	0	0	0
Fear (%)	0	2.33	97.33	0	0	0.33
Anger (%)	0	0	0	96.67	0	3.33
Disgust (%)	1.33	0	0	0	98.67	0
Sad (%)	0	0	1	2	0	97
	Average accuracy = 97.02%					

starts from a neutral image, and gradually the levels of an expression increases. Finally, that particular expression reaches its peak level. In our experiment, different levels of expressions are categorized into three groups similar to the method proposed in [132]. These levels are basically ground level, mid-level, and peak/apex level of an expression. In our method, 3 images from mid-level, and $3 \sim 4$ images from apex level are selected. Also, the ground level is used as a neutral image in our experiments. For JAFFE database, all the images are used in our experiments, whereas a part of CK+ dataset is used as mentioned in Table 3.1. The average accuracy and Krippendorff's alpha for all the combinations of training and testing datasets with 50 selected sub-regions are shown in Table 3.6.

The accuracy of state-of-the-art approaches are shown in Table 3.7, and the comparison of these results with our proposed method as listed in Table 3.6 shows that our proposed method outperforms the existing works. For classification, support vector machine with linear kernel is used [133]. As SVM is originally developed for two class classification, one-vs-one approach is adopted for multi-class classification. In One-vs-One approach, 15 hyperplanes have to be estimated for classification of six classes. The classification decision is taken on the basis of voting.

Table 3.6: Performance analysis of our proposed method on MUG, JAFFE, and CK+ databases, where AA and KA represent average accuracy (in %) and Krippendorff's alpha respectively

		Image sizes →	160 × 128		120 × 96		80 × 64		60 × 48	
Training Dataset	Testing Dataset	Training spC	AA	KA	AA	KA	AA	KA	AA	KA
MUG	MUG	70 ∼ 75	98.03	0.97	**98.44**	**0.98**	97.14	0.96	97.02	0.96
MUG	CK+	- do -	97.32	0.90	**97.32**	**0.93**	94.44	0.84	91.02	0.74
MUG	JAFFE	- do -	**98.05**	**0.94**	98.00	0.94	97.00	0.91	89.50	0.84
JAFFE	JAFFE	10 ∼ 15	**98.51**	**0.98**	98.51	0.96	98.00	0.96	97.66	0.94
JAFFE	MUG	- do -	**94.04**	**0.88**	93.50	0.72	86.00	0.58	73.00	0.52
JAFFE	CK+	- do -	**86.41**	**0.74**	84.91	0.74	83.00	0.65	61.83	0.65
CK+	CK+	70 ∼ 75	97.50	0.97	**97.91**	**0.97**	95.50	0.95	93.25	0.81
CK+	MUG	- do -	96.41	0.93	**97.58**	**0.96**	92.54	0.84	79.70	0.79
CK+	JAFFE	- do -	**97.02**	**0.95**	94.64	0.84	91.66	0.73	88.00	0.73

3.4 Conclusion

The main focus of the proposed method presented in this chapter is to extract discriminative features from the informative regions of a face for efficient facial expression recognition. We proposed an informative region extraction (IRE) model that determines the importance of different facial sub-regions using projection analysis of expressive images. Procrustes-based approach is also proposed to estimate a common reference image (CRI). Subsequently, a weighted-PLBP (WPLBP) feature extraction approach is proposed, which extracts LBP features from the informative regions and then features are weighted on the basis of importance of respective regions. The region of importance is estimated by projecting an image onto the reference image. Our proposed modeling of informative regions is inspired by the theory proposed by Ekman *et al.* [2], which says that any facial expression is resulted from the movements of a set of facial regions with respect to the their neutral state. Because of the movements of different facial regions, the texture patterns of underline regions differ from their original texture patterns of the corresponding regions of a neutral face. These deviations in terms of the projection errors indirectly give information of the informative regions of a face for a particular facial expression. In view of this, IRE model and WPLBP are proposed to estimate the informative regions of a face, and subsequently discriminative features are extracted for expression recognition. The importance of a region is judged on the basis of the analysis of projection errors. Higher projection er-

Table 3.7: Performance of the state-of-the-art methods

Existing works	Features Selection	Classifier	Database	Accuracy (%)
Rahulamathavan *et al.* [3]	Local fisher discriminant	LFDA	MUG	95.24
	- do -	- do -	JAFFE	94.37
Shan *et al.* [60]	Local binary pattern	linear SVM	CK+	94.60
	- do -	- do -	JAFFE	79.80
	LBP Boost	- do -	CK+	95.00 ± 3.2
Oshidari *et al.* [69]	Adaptive Gabor wavelet	SVM	JAFFE	90.00
Dongcheng *et al.* [123]	Gabor wavelet phase	k-NN	JAFFE	92.37
	Gabor amplitude + phase	- do -	- do -	93.48
Kotsia *et al.* [41]	Gabor wavelet	SVM	JAFFE	88.10
	- do -	- do -	CK	91.60
Zhong *et al.* [102]	LBP	MTSL	CK+	89.89
Liu *et al.* [103]	LBP	MTSL + SVM	CK+	**97.70**

ror signifies that the particular region undergoes significant deviation from their corresponding neutral state for showing the expression. As a result, that particular region conveys significant information of the considered facial expression. So, the features are only selected from the informative regions as they convey significant information of a facial expression. Hence, the feature vector derived only from the informative regions is more discriminative as compared to the feature vector obtained from the entire face image.

Performance of our proposed approach is evaluated on three well-known databases namely MUG, JAFFE, and CK+. Our proposed WPLBP approach consistently gives better performance as compared to other existing approaches. The proposed method is validated for different parameters, such as separability analysis and performance for different image resolutions. The Krippendorff's alpha coefficients for different training-testing pairs show the reliability of the

proposed work. However, one of the shortcomings of the proposed method is that projection-based approach cannot be directly extended to non-frontal face images to determine the importance of different facial regions. Moreover, the informative regions of a non-frontal face image is a subset of informative regions of a frontal face image, *i.e.*, although some of the informative regions may be invisible in a non-frontal face images, still recognition can be made with the help of rest of the visible informative regions. Hence, the method proposed in this chapter can be extended to utilize informative regions of a face for multi-view or view-invariant face images.

4

An Informative Region-based Face Model for FER

The accuracy of facial expression recognition algorithms depends on the kind of face model used for recognition. Existing face models mainly extract geometrical features from some pre-defined facial points. Also, the face models available in the literature are not suitable for extracting features from all the informative regions of a face. The drawbacks of existing face models are highlighted in this chapter, and subsequently informative regions extracted in the previous chapter are deployed for generating an efficient face model. The advantage of our proposed face model is that both geometrical and texture features can be extracted from the facial points which are marked on the basis of informative regions. The efficacy of our proposed face model is validated by recognizing gestures only with the help of facial expressions.

4.1 Introduction

The motivation of introducing this chapter is to highlight structural difference of face models [14, 92, 96, 108], and to show the shortcomings of these models. A brief overview of different face models has been discussed in Section 2.3.2. Most of the existing face models extract geometrical facial features from a number of pre-defined facial points. In the literature, it has been established that texture features extracted from landmark points improve recognition accuracy [96, 97]. This is due to the fact that texture features add some of the additional informative regions/facial points which were not previously taken into consideration by the geometrical features. Skin textures around landmark points changes significant during facial expressions, and hence, all the regions showing texture pattern variations should be considered for the face model. However, existing face models are mainly intended to extract geometrical features, and so, existing face models are not optimal for texture-based FER. Also, there are different types of face models, and hence, it is difficult to select the best one for facial expression recognition. No standard analysis is available in the literature related to the selection of a suitable face model for FER.

In this view, few existing face models are analyzed, and the shortcomings of these models are highlighted in this chapter. Finally, we proposed a more efficient face model with the help of informative regions of a face. The performance of our proposed face model is validated by recognizing facial expressions from both images and videos.

4.2 Face model-based FER

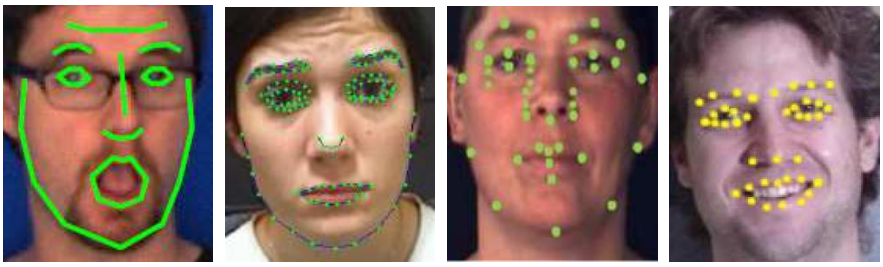
Existing facial expression recognition methods can be broadly classified into two categories: 1) face-shape-free-based methods [8, 36, 134], and 2) face-shape-based methods [33, 93–95]. The main difference between the above two categories occurs in the facial features extraction process. Shape-free-based methods can further be sub-divided into local and global methods. Global-based methods are principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), and so on. In shape-free-based methods, local feature-based methods are more generic and efficient as compared to global feature-based approaches

[61–64]. The detail about these methods can be found in Sections 2.3.1 and 2.3.2. On the other hand, face-shape-based methods make use of 2D or 3D face-shape models. FER based on shape-based methods give better performance than shape-free-based methods [33, 95]. In face-shape-based methods, either shape itself or derived features from a shape are used for facial expressions recognition. However, there exists a number of face-shape models which are used to encode a face [12,33]. These shape-models differ due to the placement of different landmark points on a face. The common regions which are often used to localize landmarks are the regions nearer to eyes, eyebrows, and mouth. This is due to the fact that these regions undergo significant movements from their neutral state during facial expressions. In other words, these regions are more informative as compared to the other regions of a face in the context of facial expression recognition.

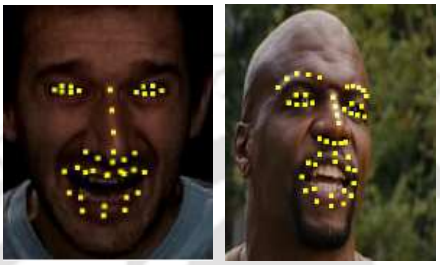
Face models are also widely used in tracking-based approaches to recognize facial expressions from videos. Agris *et al.* proposed a face model with 50 landmark points to show the significance of facial expressions in non-verbal communication [14,88,89]. Ari *et al.* proposed 116 landmark points-based face model [135]. Walecki *et al.* modelled intensity levels of expressions using a probabilistic approach [12]. As discussed in the literature, each of the face models has some inherent advantages and disadvantages. So, it is quite difficult to select a particular face model for efficient FER. Moreover, no convincing analyzing techniques are available in the literature to select the best possible face model. In view of this, we propose to extract a more generic face model with the help of informative regions of a face. Our face model is more versatile in the sense that both geometric and texture features can be extracted from the landmark points of a face. The proposed framework is elaborately discussed in the following section.

4.3 Analysis of Existing Face Models

Few existing face models which are widely used to recognize facial expressions from frontal view and/or non-frontal views are analyzed. Different face models are selected on the basis of the pattern of different landmark patterns points on a face. A set of selected face models for our analysis is shown in Figure 4.1. The drawbacks of existing face models are now highlighted to develop an efficient face model.



(a) Agris [14] (b) Ismail [135] (c) Yang [136] (d) Rudovic [33]



(e) Walecki [12] (f) Walecki [12]

Figure 4.1: Face models widely used in the context of facial expression recognition showing different geometrical pattern of the face models.

- **Boundary points:** Since facial expressions are independent of shape of a face [135], and hence, landmark points located at the boundary of a face do not play any role in facial expression recognition. Therefore, these points may not be included in face models. Hence for FER, models which are free from boundary points are more cost effective as compared to the face models having the boundary points. Thus, an efficient face model may not contain boundary points (Figure 4.1(d), (e) and (f)).

- **Texture feature:** Existing face models are mainly intended to extract geometrical features from a face. So, these models are not suitable for extracting texture features from a face. However, texture features are more informative as compared to geometric features for facial expression recognition [34]. Also, combined use of geometrical and texture features improves the performance of FER. Thus, a good face model should provide both geometric as well as texture features of a face.

- **Localization of facial points:** All the existing face models localize facial points nearby eyes, eyebrows, and mouth regions, as these regions have significant movements with respect to their neutral state. Moreover, there are some other regions of a face which are

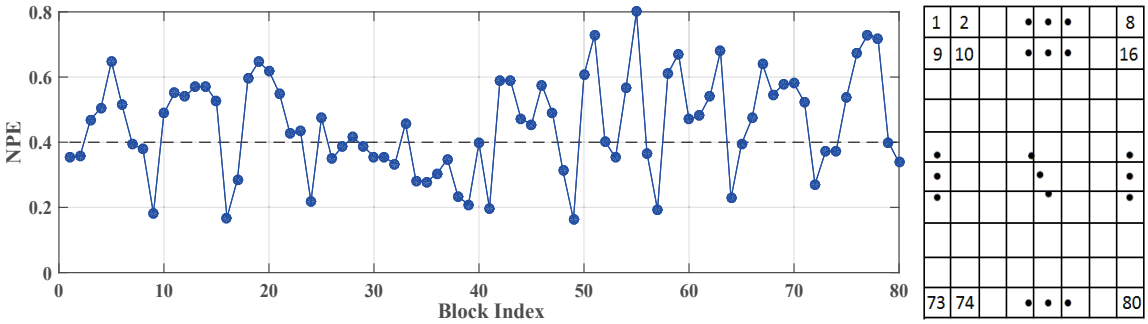


Figure 4.2: Distribution of average projection errors of different sub-regions [Left]. The horizontal axis of left figure shows block indices corresponding to different face regions. The region indices of a face are chosen as shown in [Right] figure.

also important for FER. For example, regions between the eyebrows and regions nearer to jaw are not considered in the existing face models. However, these regions have some significance in "anger" and "sad" expressions. Thus, existing face models need some more additional landmark points to extract relevant texture features from all the informative regions of a face.

4.3.1 Proposed face model

As discussed in the previous chapter, the distribution of informative regions of a face can be estimated from the projection errors, and this distribution is illustrated in Figure 4.2. The proposed face model is derived from the distribution of informative regions [3]. The importance of different informative regions is judged by comparing the magnitude of projection errors with a threshold. In our case, the threshold value is fixed at 0.4, which is manually decided based on large number of observations. The regions having projection errors more than the threshold value are considered to be more informative and vice versa. This distribution clearly shows that regions nearer to mouth and eyes have more deviations as compared to other facial regions. Also, it shows that regions like "between eyebrow", "outer mouth regions", and "jaw part" are also important. These regions play important roles in discriminating expressions such as "anger and disgust", "happy, surprise, and disgust", and "sad and surprise" respectively. Our

[3]This work has been published in *National Conference on Communication (NCC) 2016* (Refer item 2 in Page 143 for details)

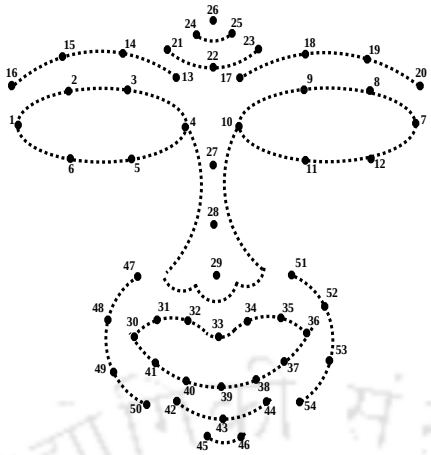
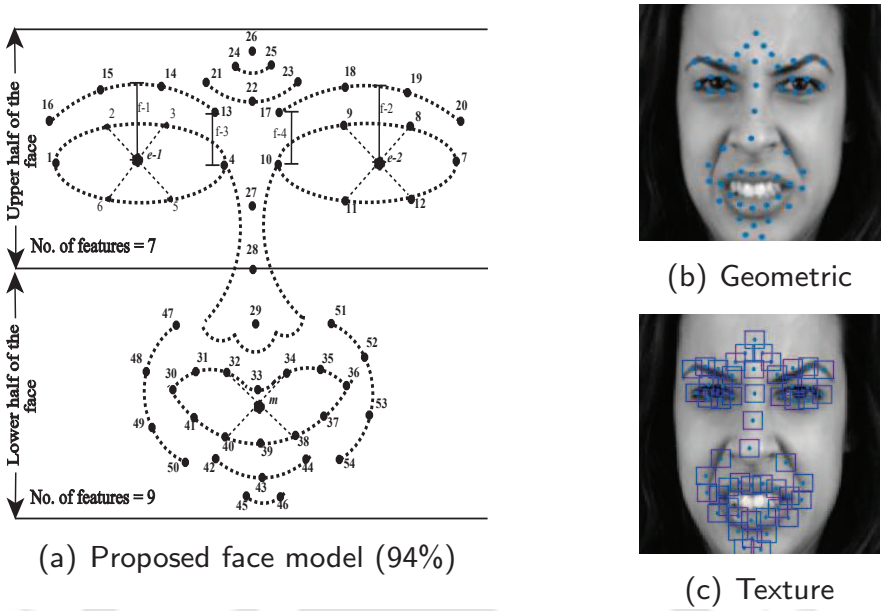


Figure 4.3: Proposed face model.

proposed face model approximately selects all the 49 informative regions based on the values of the projection errors. In addition to that, our model has five non-informative regions which are important for evaluating the geometrical features. Thus, in total our proposed face model has 54 facial points, and it is shown in Figure 4.3. The proposed face model is not only efficient for extracting texture features, but it is also equally important for extracting geometrical features from a face.

4.4 Feature Extraction

Finally, the proposed face model is employed to extract geometrical as well as texture features of a face image. For our proposed face model, the entire face is divided into two parts *i.e.*, upper part of a face and lower part of a face. The upper part of a face mainly includes eyes and forehead, whereas lower part of a face covers mouth and their neighbouring regions. Subsequently, Euclidean distance-based geometric features are extracted from each of the parts. Also, texture features are extracted from each of the sub-regions around the localized facial points.



(a) Proposed face model (94%)

(b) Geometric

(c) Texture

Figure 4.4: Extraction of geometrical and texture features from the proposed face model.

4.4.1 Geometrical features

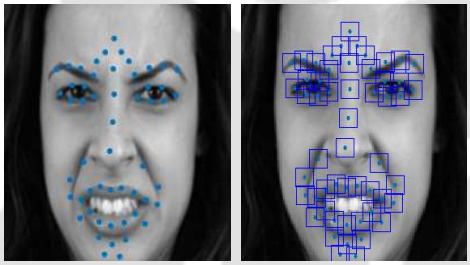
As explained earlier, the deformation of a neutral face image causes different facial expressions. The facial regions undergo different movements for showing facial expressions, *i.e.*, facial expressions are the manifestations of different facial muscular movements. Thus, geometric features can differentiate facial expressions on the basis of facial shapes. In view of this, following geometrical features are extracted with the help of the proposed face model as illustrated in Figure 4.4. Table 4.1 shows all the geometrical features extracted from our proposed face model, where f_i is the i^{th} feature, e_1, e_2 and m are the center points of two eyes, and mouth respectively. These center points are determined using simple geometrical analysis. For example, e_1 can be obtained by solving the equation of lines joining the points 2, 5 and 3, 6 in the face model. The term $d\,(a, b)$ represents Euclidean distance between the points a and b. The points a and b are the landmark points used in our proposed face model (Figure 4.4). All the geometrical features $f_1 - f_{16}$ are normalized by the distance $d\,(1, 7)$ to minimize the effect of shape variations among different subjects. Thus, a 16-dimensional feature vector is extracted from the proposed face model.

Table 4.1: Geometrical features extracted from the proposed face model.

Upper half part of a face	Lower half part of a face
$f_1 = [d(e_1, 15) + d(e_1, 16)]/2$	$f_8 = [d(30, 36) + d(31, 35) + d(41, 37)]/3$
$f_2 = [d(e_2, 8) + d(e_2, 9)]/2$	$f_9 = d(32, 40)$, $f_{10} = d(33, 39)$, $f_{11} = d(34, 38)$
$f_3 = d(4, 13)$, $f_4 = d(10, 17)$, $f_5 = d(13, 26)$	$f_{12} = [d(m, 48) + d(m, 49)]/2$
$f_4 = d(10, 17)$, $f_5 = d(13, 26)$	$f_{13} = [d(m, 52) + d(m, 53)]/2$
$f_5 = d(13, 26)$	$f_{14} = d(40, 45)$
$f_6 = d(23, 26)$	$f_{15} = d(38, 46)$
$f_7 = d(13, 17)$	$f_{16} = d(39, 43)$

4.4.2 Proposed texture feature extraction

Facial texture features are extracted from each of the informative regions defined in our proposed face model. As discussed in Chapter 3, LBP-based texture feature is extracted. For this, a



(a) Geometric (b) Texture

Figure 4.5: Landmark points in the proposed face model for extracting geometrical and texture features

12×12 block is selected around each of the informative regions as shown in Figure 4.5, and subsequently, 59-dimensional uniform LBP features are extracted. As our model comprises of 54 fiducial points, total number of features per image would be $59 \times 54 = 3186$.

Although PCA is a popular technique to reduce dimensionality of a feature vector, it is not suitable to capture discriminative features, as it also maximizes intra-class variance [3]. Hence, multiple discriminant analysis (MDA) is applied to reduce the dimension of the feature vector. MDA requires a set of labelled feature vectors, which are not readily available in the videos having both basic and non-basic expressions. This is due to the fact that the prior information of an expression is not available in hand. So, we proposed to use labelled images of some

Table 4.2: Grouping of the sub-regions on the basis of the landmark points.

Principal regions	Sub-region indices	Description of the regions
R_1	$r_1 - r_6$, $r_{13} - r_{16}$	Left eye region
R_2	$r_7 - r_{12}$, $r_{17} - r_{20}$	Right eye region
R_3	$r_{21} - r_{28}$, r_4, r_{10}, r_{13}, r_{17}	Between eyebrow
R_4	$r_{29} - r_{33}$, $r_{39} - r_{43}$, r_{45}, $r_{47} - r_{50}$	Left side of mouth
R_5	$r_{33} - r_{39}$, $r_{43} - r_{44}$, r_{46}, $r_{51} - r_{54}$	Right side of mouth

standard datasets (MUG [119], JAFFE [117] and CK+ [118]) to extract a set of basis vectors which can be subsequently used to reduce the dimensionality of feature vectors of unlabelled expressions of a video.

Each face image is first labelled with our proposed face landmarking scheme using fast-AAM [137] (**Appendix A.1**). Then, facial regions are segregated into five groups on the basis of the position of the landmark points. The grouping of sub-regions and region descriptions are given in Table 4.2. In Table 4.2, R_i represents i^{th} region, which consists of several sub-regions shown in the second column of Table 4.2. Each sub-region is represented by r_j, $j = 1, 2, ..., 54$ [4]. Subsequently, MDA is applied on the region R_i, $i = 1, 2, ..., 5$ separately. For C-class problem, MDA gives $C - 1$ basis vectors per region, and hence the total number of basis vectors for all the regions would be $5\,(C - 1)$. Finally, feature vector of each of the frames of a video is projected onto MDA basis vectors to obtain reduced discriminative features of each of the frames. Hence for $C = 7$, each frame of a video is represented by $5 \times (7 - 1) = 30$ dimensional feature vector. The proposed texture feature gives almost equivalent performance as that of geometric features. The texture features together with the geometrical features can give better performance in recognizing expressions both from images or videos. A more detailed analysis of our proposed face model and the extracted features is presented in the experimental section.

[4]This work has been published in *Journal on Multimodal User Interfaces, Springer 2016* (Refer item 2 in Page 143 for details)

4.5 Experimental Results

To validate our proposed face model and the associated features, expressions are recognized from some static images. Subsequently, some of the additional experiments are performed to recognize gestures by recognizing facial expressions in a video.

In our first experiment, efficacy of the proposed face model is validated by recognizing basic expressions from static face images. MUG dataset [119] is used for this purpose. This dataset contains images of 86 subjects which includes the facial expressions of 35 women and 51 men. Each image of MUG database is of resolution 896×896, which is resized to 120×120 for our experiment. Out of 86 subjects, 51 subjects (31 women and 20 men) are randomly selected. In our experiment, images from 15-20 subjects are used for training, and images of remaining subjects are used for testing. For each class, five levels of expressions are considered for each of the subjects. Discriminative property of texture features extracted from informative regions marked in our proposed face model was demonstrated in Chapter 3 (Section 3.3).

To compare the performance of our proposed face model with other face models, the proposed face model is modified by excluding the face boundary points to get almost equivalent face models corresponding to other standard face models shown in Figure 4.1. For example, models shown in Figure 4.1 can be easily generated with the help of our proposed face model by deleting or adding few landmark points. So, most of the existing face models are the subset of proposed face model in terms of facial landmark points. Subsequently, features are extracted from each of the existing face models to evaluate their performances in terms of facial expression recognition. The comparison among different face models for texture features is shown in Table 4.3. Table 4.3 clearly shows that the proposed model gives better recognition accuracy as compared to other face models.

4.5.1 Case study: (Recognition of gestures only with the help of facial expressions)

In this experiment, we consider a problem of recognizing a set of gestures with the help of facial expressions. Not much works related to gesture recognition on the basis of facial expressions

Table 4.3: Performance evaluation of different face models using texture features.

Methods	Recognition Rates in (%)						
	Happy	Surprise	Fear	Anger	Disgust	Sad	ARR
Agris *et al.* [14]	92.52	97.02	78.11	65.62	72.52	82.01	81.30
Ismail *et al.* [135]	92.88	96.89	84.63	70.62	72.52	82.01	83.25
Yang *et al.* [136]	82.00	78.30	71.58	67.81	70.68	68.56	73.15
Rudovic *et al.* [33]	78.66	79.00	58.80	53.78	67.80	61.00	66.50
Walecki *et al.* [12]	87.20	82.88	71.06	55.22	69.70	68.50	72.42
Walecki *et al.* [12]	93.50	88.00	79.80	69.44	79.61	74.00	80.72
Proposed model	**98.60**	**98.88**	**84.70**	**94.40**	**95.40**	**92.00**	**93.99**

has been reported till date. Facial features used for sign gestures are mainly geometric features. Agris *et al.* examined the role of facial features in the context of automatic sign language recognition [14]. They proposed a face model having 50 landmark points, which are automatically tracked by using AAM [89]. Subsequently, a 16-dimensional geometric feature is extracted from each of the frames of a sign language corpus. Their method achieved 40%-48% recognition rates, when only geometric features are utilized. Ismail *et al.* tracked 116 landmark points using a variant of active shape model, whereas Nguyen *et al.* tracked 21 fiducial points using Kanade-Lucas-Tomasi tracker along with Bayesian feedback mechanism [135], [138]. Further, tracked facial points are used to extract geometric features from different frames of a video. Finally, these features are used to classify a sign corpus. However, performances of all these methods extremely depend on the accuracy of tracking algorithms and the sequence classifiers used to model the sign gestures.

Hidden Markov Model (HMM) is a commonly used sequence classifier for gesture recognition. Hidden states of HMM capture spatial and temporal characteristics of a gesture [139], [140]. Several other variants of HMM such as factored-HMM, parallel-HMM, and coupled-HMM have been proposed [141, 142]. However, Markovian constrains on these models restrict them to capture long-run dependencies in the gesture sequences. Conditional random field (CRF) is another variants of probabilistic model which avoids inter-independence assumption between gesture sequences, and allows non-local dependencies between the states and the observations [143]. However, CRF requires observable hidden states as it models the distribution of hidden states conditioned on the observation. For this, manual intervention is required to tag

each of the observation sequences with the emotion labels. This step has to be done for all the gesture sequences. Apparently, this step is time consuming, and even more challenging in the context of sign language recognition. Wang *et al.* proposed another variant of discriminative model known as Hidden-CRF (HCRF) which combines the HMM to capture spatial and temporal variability, and the CRF to capture long-run dependencies of a gesture [144]. A special HCRF is proposed by Kim *et al.* which can impose an ordinal constrain on the hidden states to model intensity levels of a facial expression [132]. However, such an ordinal constrain on states may be suitable only when a particular expression is in its active state, and hence this condition restricts its practical applicability for sign language recognition. In reality, an expression may or may not be in active state. Walecki *et al.* proposed another variant of HCRF, where an additional hidden state is included which automatically decides whether an expression is in its nominal state or ordinal state [12]. However in [12] and [132], it is assumed that component emotions remain unchanged for a composite expression. But in sign language, component expressions may change during the entire period of gesturing. Hence, ordinal constrain may not be always applicable for sign language expression modeling. Consequently, HCRF is more suitable for modelling facial expressions of sign language video. Therefore, HCRF is used as a sequence classifier in this context (Case study: recognition of signs with the help of only facial expressions).

4.5.1.1 Sequence classification using Hidden Conditional Random Field

Recognition of sign is basically a C-class time-series sequence classification problem, where the objective is to assign a class label $y \in \{1, 2, ..., C\}$ to a test measurement sequence $\mathbf{x} = \{x_1, x_2, ..., x_T\} \in D \times T$ for a given model. The class label y is further represented by a sequence of hidden states $h \in \{1, 2, ..., K\}$, where K is the number of hidden states. The length of a sequence T can vary from one sign to another sign, whereas dimension of a feature vector of each frame will remain the same for the entire sequence.

HCRF is a probabilistic classification model $P(y|\mathbf{x})$, which models temporal dynamics of facial expressions in sign language gesturing as a sequence of hidden states relating facial features to the class label [12, 132]. The general framework of HCRF can be considered as a combi-

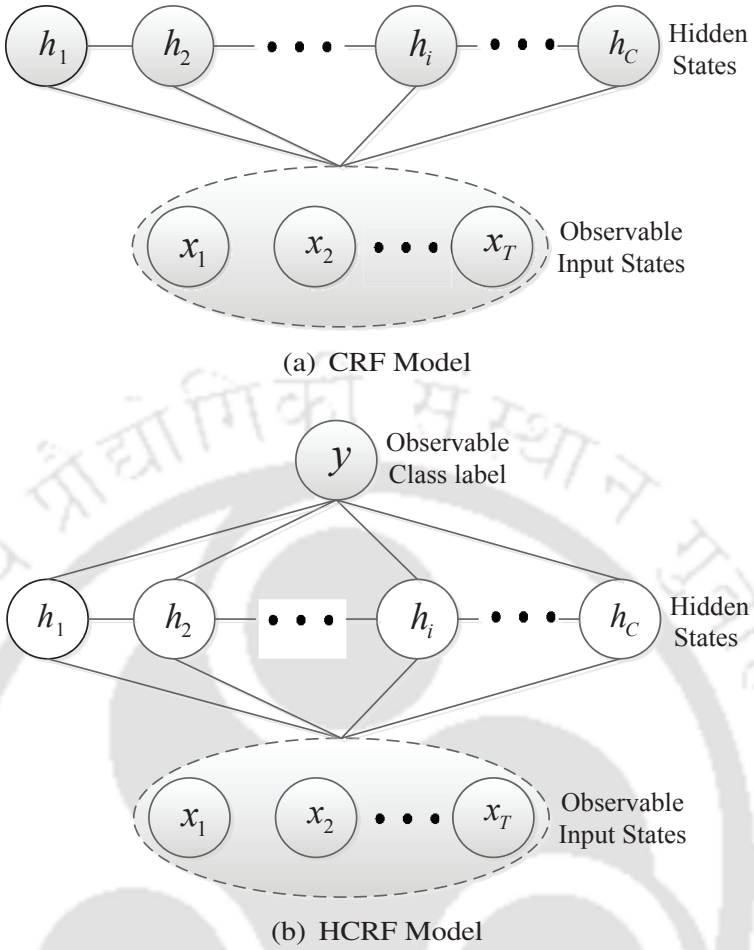


Figure 4.6: Structure of HCRF, which has an extra observable node (shaded node) at the top in contrast to CRF. The latent/unobservable states are shown as white background, whereas observable states are shown as shaded/gray background. Hidden states of HCRF are unobservable, whereas they are observable in CRF.

nation of K-CRFs – one for each class. So, HCRFs are suitable for multi-class classification. For multi-category classification, an extra node in the graph structure of CRF is introduced as shown in Figure 4.6. This extra node represents the class label, where the hidden/latent states $\mathbf{h}$ are now assumed to be unobserved variables. More specifically, HCRF aims to model the joint distribution of the class label y and the hidden states $\mathbf{h}$ conditioned on the input feature vector $\mathbf{x}$. The mathematical representation of the conditional distribution $p(y, \mathbf{h}|\mathbf{x},\Omega)$

of HCRF takes the Gibbs form clamped on the observation $\mathbf{x}$ as:

$$p\left(y, \mathbf{h}|\mathbf{x}, \Omega\right) = \frac{1}{Z\left(\mathbf{x}\right)} \exp\left\{s\left(y, \mathbf{h}, \mathbf{x}; \Omega\right)\right\} \tag{4.1}$$

As hidden states $\mathbf{h}$ are unobservable, they can be integrated out from the expression shown in Eqn. (4.1), which will result a class conditional distribution given by Eqn. (4.2).

$$\begin{aligned} p\left(y|\mathbf{x}, \Omega\right) &= \sum_{\mathbf{h}} p\left(y, \mathbf{h}|\mathbf{x}, \Omega\right) \\ &= \frac{1}{Z\left(\mathbf{x}\right)} \sum\nolimits_{\mathbf{h}} \exp\left\{s\left(y, \mathbf{h}, \mathbf{x}; \Omega\right)\right\} \end{aligned} \tag{4.2}$$

The argument of exponential function of Eqn. (4.1) is known as score function of HCRF, which is basically the combination of score functions of K-CRF – one for each class. Thus, the score function of HCRF can be written as:

$$s\left(y, \mathbf{h}, \mathbf{x}; \Omega\right) = \sum_{k=1}^{K} I\left(k = y\right) . s\left(\mathbf{h}, \mathbf{x}; \theta_y\right) \tag{4.3}$$

where, $s\left(\mathbf{h}, \mathbf{x}; \theta_y\right)$ is the score function of y^{th} CRF model, $\Omega = \{\theta_k\}_{k=1}^{K}$ denotes model parameters, and $I\left(k = y\right)$ is the indicator function whose value is 1 iff $k = y$, otherwise 0.

Evaluation of class-conditional density $p\left(y|\mathbf{x}, \Omega\right)$ given in Eqn. (4.2) depends on the partition of $Z\left(\mathbf{x}\right)$ and the joint distribution of class labels, and hidden states *i.e.*, $p\left(k, h_r, h_s|\mathbf{x}\right) = p\left(h_r, h_s|\mathbf{x}, k\right) p\left(k|\mathbf{x}\right)$. These terms can be computed using independent assumption in K-CRFs. The learning of HCRF parameters Ω in Eqn. (4.2) is carried out by maximizing the negative log-likelihood of class-conditional distribution function of Eqn. (4.2). The optimum value of parameters Ω^* is a point where the score function attains the maxima. These parameters can be estimated using the approaches mentioned in [12,132]. Finally, given the model Ω^*, the class label y^* of an unknown test sequence $\mathbf{x}^*$ is obtained by Eqn. (4.4).

$$y^* = \arg\max_{y} p\left(y|\mathbf{x}^*, \Omega^*\right) \tag{4.4}$$

Gesture recognition only using facial expressions: In this experiment, we have ana-

lyzed the significance of facial expressions in non-verbal communication. RWTH-BOSTON-50 database [145] is used for our analysis. This database has 483 utterances of 50 words of American sign language. The signs are recorded by multiple cameras placed at different positions with a frame rate of 30. The size of all the images is 312×242. All the samples of the words are annotated by three signers: one man and two female signers. We have altogether selected 15-20 sign words for our experiment, where each word comprises of 10-15 sign utterances. The selected words include simple (single handed) as well as complex (double handed) gestures, which additionally require facial expressions either to complete or to boost the meaning of a gesture. A sample word "CAN" and the corresponding frames are shown in Figure 4.7, where manual

Figure 4.7: Showing importance of facial expressions in the "CAN" word in contrast to the manual parameters (movement of the hands).

parameters (hand shapes and their locations; movement and orientation) are comparatively less significant as compared to the non-manual parameters (facial expressions). Figure 4.8 shows few examples of localization of salient points of the proposed face model using fast-AAM.

We performed four experiments to validate our proposed face model and the associated features. In the first experiment, we mainly analyze the spatio-temporal variability present in the facial gestures. This is carried out in terms of the changes in the latent states of the models

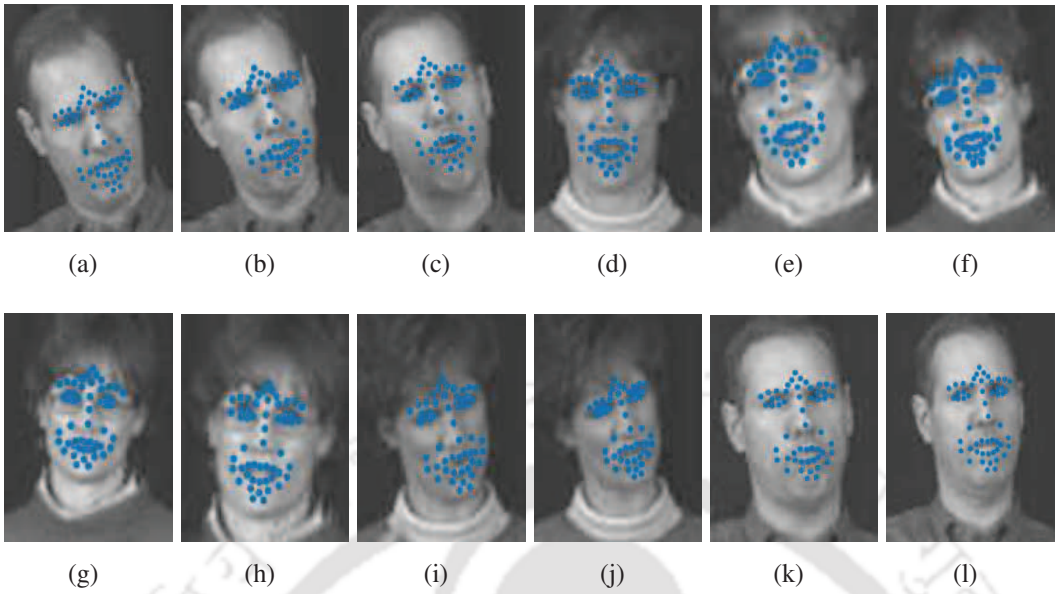


Figure 4.8: Localization of landmark points of our proposed face model on different frames of RWTH sign language dataset using fast-AAM.

(HCRF and HMM) over the gesturing time (profile of a gesture). For the second experiment, utterances of selected words are divided into three parts. Two out of three parts are randomly selected to estimate the model parameters of HCRF, and the third part of remaining data is used for testing. For third experiment, utterances are separated based on the signers *i.e.*, all the instances of two signers are used to learn the model parameters, whereas samples of third signer are employed for testing. In the forth experiment, we extracted geometric features as discussed in Section 4.4.1. In the proposed method, geometrical features and 30-dimensional texture features are concatenated to get more information of facial regions. This combined feature in-turn gives more accuracy.

Profile Analysis: The profile is a graphical representation of a gesture, where some characteristics of a gesture are plotted over time based on observations. The latent states of HCRF captures the spatial-temporal variability present in a sequence of facial gestures. The changes in the latent states of the model over the period of time indirectly indicate the corresponding changes of the facial gestures. Profiles of the words "CAN", "WHAT", "LOVE" and "WHO" are shown in Figure 4.9. These gestures are correctly recognized by HCRF with the likelihood of 0.9 or more. It is clear from Figure 4.9 that the profiles of these gestures are significantly

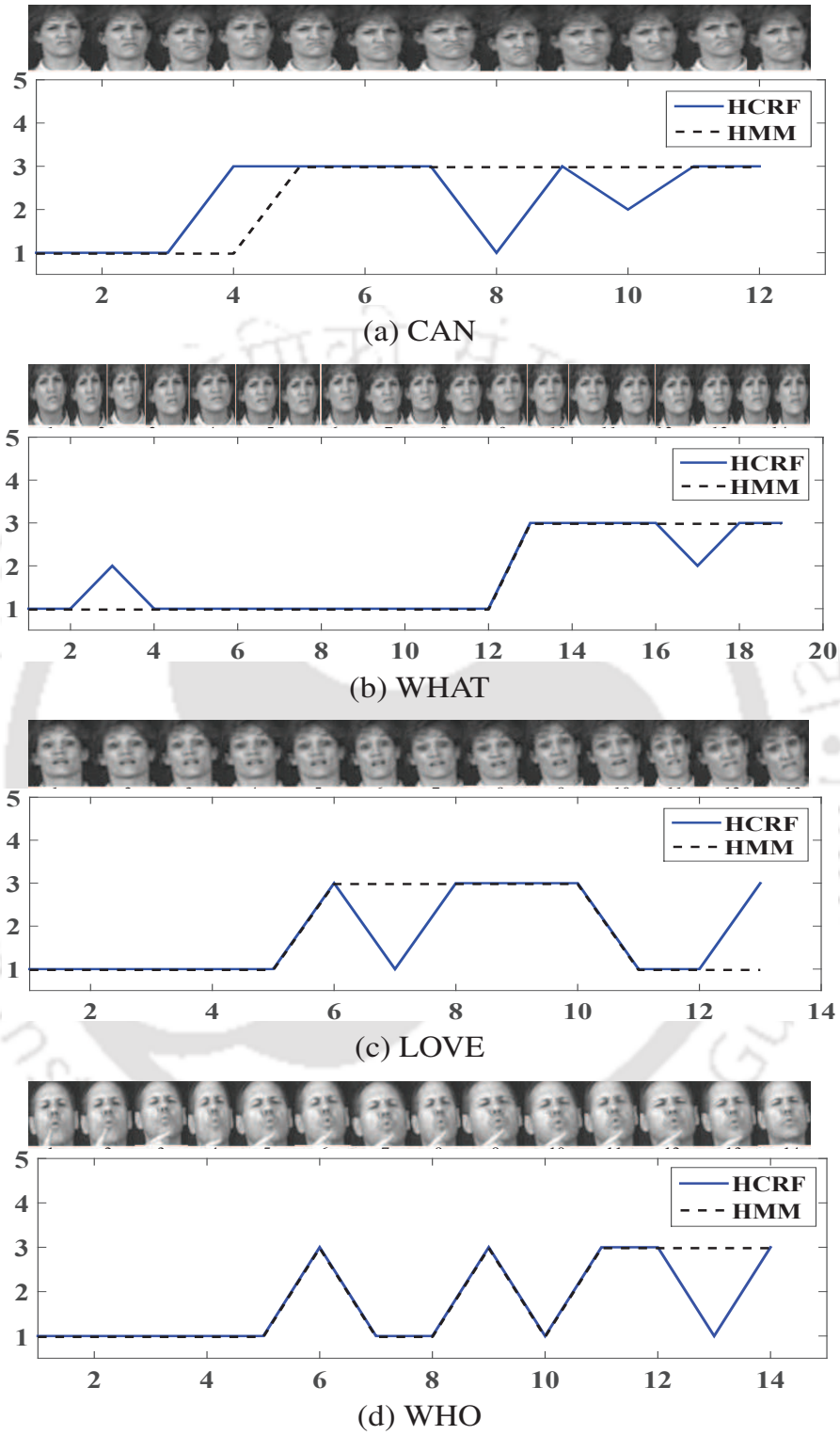


Figure 4.9: Showing the profile of few sign language words, where vertical axis represents latent states of HCRF or HMM (3-states) and horizontal axis represents the frames of a video.

Table 4.4: Average accuracies in recognizing different sign language facial gestures.

Experimental procedure			Recognition Rates in (%)			
Training samples	Features	Classifiers	$C = 10$	$C = 15$	$C = 20$	Avg
2-fold of C class dataset	TF	HMM	56.20	37.02	27.81	40.34
2-fold of C class dataset	GF	HMM	51.46	45.00	34.65	43.70
2-fold of C class dataset	TF	HCRF	85.54	70.62	61.55	**72.57**
2-fold of C class dataset	GF	HCRF	86.30	63.44	61.12	**70.29**
2-fold singers of C class	TF	HMM	22.67	16.66	12.86	17.40
2-fold singers of C class	GF	HMM	40.22	30.45	27.10	32.59
2-fold singers of C class	TF	HCRF	67.88	51.28	48.64	**55.93**
2-fold singers of C class	GF	HCRF	68.00	65.56	52.32	**61.96**
2-fold of C class dataset	TF + GF	HMM	59.00	37.00	35.50	43.83
2-fold of C class dataset	TF + GF	HCRF	93.98	78.20	68.00	**80.06**
2-fold singers of C class	TF + GF	HMM	49.80	48.67	33.87	44.11
2-fold singers of C class	TF + GF	HCRF	78.80	67.70	52.22	**66.24**

TF: Texture Features
GF: Geometrical Features

different from each other. That is why, facial expressions can discriminate the signs having almost similar hand movements and positions. Also, the latent states of HCRF can capture gesture variations more precisely than the hidden states of HMM.

Performance Analysis: For our experimental validation, average accuracy is calculated. The accuracy is computed by using Eqn. (4.5), where c_{ij} is the $(i,j)^{th}$ element of a $C \times C$ confusion matrix.

$$Accuracy = \frac{\sum_{i=1}^{C} c_{ii}}{\sum_{i,j=1}^{C} c_{ij}} \times 100\% \tag{4.5}$$

Table 4.4 shows the accuracy for different experimental procedures. All the experiments are repeated for different values of C, different numbers of isolated words, and also for different types of facial features (geometrical and/or texture). Table 4.4 shows that both texture-based facial features and geometrical-based facial features can give almost similar performances. However, combined geometrical and texture features can significantly give better performance as compared to geometrical or texture features. Additionally, the third experiment is again repeated for a small set of gestures *i.e.*, for $C = 6$. The corresponding confusion matrices are shown in Tables 4.5 and 4.6. The proposed method is analyzed with HMM and HCRF, and we found that HCRF performs better as compared to HMM in this kind of sequence classification problem.

This may be because of the Markovian constraints on HMM, which restrict HMM to capture long-run dependencies of the facial gestures.

Table 4.5: Confusion matrix showing gesture recognition rates. All the sequences of the signs are divided based on singers as described in third experimental procedure, and then HMM is applied to recognize sign language gestures.

	Can	What	Love	Who	Buy	Book
Can	78.5	6.0	0	0	15.5	0
What	1	82.0	0	0	13.0	0
Love	0	14	61.0	16	0	9
Who	0	0	6.5	87.2	0	6.3
Buy	0	19.0	0.4	0	80.6	0
Book	0	0	2	13.2	14	72.8
	Average Accuracy = 77.02%					

Table 4.6: Confusion matrix for sign gesture recognition. All the sequences of the signs are divided based on singers as described in third experimental procedure, and then HCRF is applied to recognize the corresponding gestures.

	Can	What	Love	Who	Buy	Book
Can	93.0	0	0	0	7	0
What	1	96.0	0	0	4.0	0
Love	0	0	90.2	7	0	2.8
Who	0	0	2	98	0	0
Buy	0	6.2	0	0	93.8	0
Book	0	0	3	1	0	96
	Average Accuracy = 94.5%					

4.6 Conclusion

In this Chapter, an efficient face model is proposed to extract both geometrical and texture features for expression recognition. For this, the method developed in Chapter 3 is extended to identify important landmark points in a face. The proposed face model consists of 54 landmark points which indicates a set of 54 informative regions of a face. Our proposed face model is

suitable to extract both geometric and texture features. Experimental validation of our proposed face model is carried out by recognizing some of the facial expressions from static face images. As a case study, the proposed face model is employed to recognize some of the selected signs (sign language gestures) only with the help of facial expressions. Experimental results show that the proposed model can give an improvement of about $8-10\%$, when the face model is used to recognize expressions from frontal face images. Also, experiments on sign language recognition show the significance of facial expressions on non-verbal communication. Experimental results signify that the proposed face model can play a significant role in distinguishing different signs with the help of only facial features. We also found that HCRF gives better performance as compared to HMM.

5

Uncorrelated Multiview Discriminant LPP Analysis for MvFER

The objective of this chapter is to utilize informative region-based face model and the associated features for recognizing expressions from multi-view face images, and thereby the versatility of previously developed recognition scheme is enhanced. The state-of-the-art methods extract a discriminative common space for multi-view FER (MvFER). However, these methods are not suitable for finding discriminative space if data exhibits multi-modal characteristics. In such cases, Locality Preserving Projection (LPP) and/or Local Fisher Discriminant Analysis (LFDA) are found to be more appropriate to capture discriminative directions. Also, uncorrelated constraint onto common space improves classification accuracy. Inspired from the above findings, we propose an uncorrelated multi-view discriminant locality preserving projection (UMvDLPP)-based approach. The proposed method searches a common uncorrelated discriminative space for multiple observable spaces. Moreover, the proposed method can also handle the multi-modal characteristics of the data. Hence, the proposed method is effectively more efficient for MvFER. Experimental results show the efficacy of the proposed UMvDLPP-based method.

5.1 Recognition of Multi-view Facial Expressions

In practical scenarios, images of facial expressions can be captured by the cameras which can be fixed at different points in world coordinate system. So, the captured expressive face images may have arbitrary head poses [111]. Because of this, the facial features extracted for one viewing angle may differ from the facial features obtained for other viewing angles. This scenario imposes several challenges in facial expression recognition. Some of the challenges include face alignment, face occlusion, discriminative feature extraction, and localization of facial points [8]. Few research works have been reported till date on recognizing expressions from multi-view and/or view-invariant face images. The reported techniques can be grouped into three categories: 1) methods which perform pose-wise recognition [36,90,91,97], 2) methods which perform view-normalization before recognition [8, 33, 107, 108], and 3) methods which learn a single discriminative space using the observations of multiple views [1, 134, 146].

The first group of methods select view-specific classifiers for FER. Moore and Bowden [36] performed multi-view FER by learning a view-specific supervised support vector machine (SVM) for each of the views [147]. Hu *et al.* performed multi-view FER by extracting 2D displacement vectors of 38 landmark facial points between the expressive face images and the corresponding neutral face images [90]. Subsequently, these features are fed into different view-specific classifiers. In [91], the authors investigated three kinds of appearance-based features, namely, scale invariant feature transform (SIFT) [96], histogram of oriented gradient (HOG) [98], and local binary pattern (LBP) [75] for multi-view (0°, 30°, 45°, 60°, and 90°) FER. Hesse *et al.* performed multi-view FER by extracting different appearance-based features, *e.g.,* SIFT, HOG, and discrete cosine transform (DCT) [105] around 83 landmark facial points [97]. The major shortcoming of the above-mentioned methods is that the correlation which exists across different views of the expressions are not at all considered. Since separate view-specific classifiers are learned for FER, and so, the overall classification strategy is sub-optimal.

The second group of methods mainly follow a three step procedure *i.e.,* head-pose estimation, head-pose normalization, and FER from the frontal pose. Rudovic *et al.* [33, 107, 108] recognize expressions from non-frontal views of facial images. For this, 39 facial points are localized on each of the non-frontal/multi-view facial images, and then head-pose normalization is

performed. The objective of head-pose normalization is to learn the mapping functions between a discrete set of non-frontal poses and the frontal pose. In order to learn the robust mapping functions, a coupled Gaussian process regression-based framework is proposed by considering pair-wise correlations between different views. However, learning of mapping functions is performed on the observation space, and hence, improper mapping functions can adversely affect the classification accuracy. View-normalization or multi-view facial feature synthesis method is also proposed in [8]. In this, block-based texture features are extracted from multi-view facial images to learn the mapping functions between any two different views of facial images. So, the features can be extracted from several off-regions, on-regions, and on/off-regions of a face. Their method has a limitation as the weight assignment for on/off-region is not defined. The major shortcoming of methods of this group is that the head-pose normalization and the learning of expression classifiers are carried out independently, which may eventually affect the final classification accuracy.

The third group of methods has several advantages. One important advantage is that a single classifier is learned instead of view-specific classifiers [1, 134, 146]. So, head-pose normalization is not needed. In [1], it was assumed that different views of a facial expression are just different manifestations of the same underlying facial expression, and hence the correlations which exist among different views of expressions are considered during the training phase. They proposed discriminative shared Gaussian process latent variable model (DS-GPLVM) to learn a single non-linear discriminative subspace. However, discriminative nature of a Gaussian process depends on the kind of prior. They proposed Laplacian-based prior [112], which can give better performance than the Linear Discriminant Analysis (LDA)-based prior. Laplacian-based prior preserves within-class local topology of the data onto the reduced space by minimizing the sum of squared distances between the latent positions of the examples of the intra-class. However, it ignores the effect of inter-class variations of data, which results a sub-optimal latent space. Although, DS-GPLVM can give better performance, the method proposed in [111] is a linear non-parametric projection based approach, and hence it is comparatively simpler than the parametric DS-GPLVM-based approach.

For leaning of a common discriminative space (latent space) shared by all the views, sev-

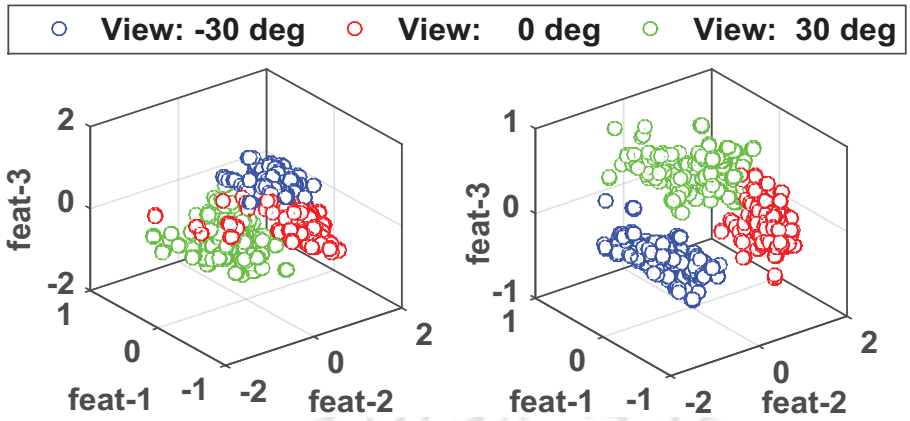


Figure 5.1: Plots showing multi-modal characteristics of "Happy" [Left] and "Surprise" [Right] facial expressions.

eral research works [109, 111, 148–150] have been proposed. Eleftheriadis *et al.* [1] showed that the above-mentioned methods can be efficiently applied for multi-view FER. Hence, inspired from the state-of-the-art multi-view learning-based method proposed in [111] and the method proposed in [115], we proposed a more efficient objective function for multi-view FER. The proposed method is termed as Uncorrelated Multi-view Discriminant Locality Preserving Projection (UMvDLPP) analysis. The proposed objective function of UMvDLPP is formulated in such a way that it can preserve the intra-class topology of intra-view as well as inter-view onto the common space. Also, it can maximize the local between-class separation of intra-view and inter-view samples. The motivation behind generalizing LPP and local between-class scatter matrix (LBCSM) in our proposed method is that they both are capable of handling multi-modal characteristics of multi-view facial data [116]. On the other hand, the simple LDA-based approach fails to capture discriminative directions when data of different classes have several local maxima and minima [116]. So, our approach entails extracting an uncorrelated common discriminative space. Figure 5.1 shows the multi-modal characteristics of multi-view "happy" and "surprise" expressions. To handle multi-modal data, we adopt LBCSM defined in [3] to maximize the local between-class separability of the data, and Laplacian-based approach to minimize the within-class local geometric structure of the data in our proposed UMvDLPP

approach.

Performance of any FER system mainly depends on the availability of most discriminative features. Apparently, informative/active regions of a face can provide most discriminative features [47, 102, 103]. In Chapter 4, we proposed an efficient face model by extracting informative regions of a face. Our proposed face model consists of 54 facial landmark points, and the features are extracted only from the salient/informative regions of a face. It was shown in our earlier work [101] that our proposed face model outperformed several existing facial models [12, 14, 33]. So, we employed our proposed informative region-based face model [101] to extract features from multi-view facial images. The proposed UMvDLPP method is elaborately discussed in the following Section.

5.2 Proposed Method

Our proposed UMvDLPP-based method generalizes Uncorrelated Discriminative LPP (UDLPP) [115] along with Local Fisher Discriminant Analysis (LFDA) [116] to learn a robust uncorrelated discriminative common space from the multi-view facial images [5]. The detail about UDLPP is discussed in **Appendix A.2**.

5.2.1 Proposed UMvDLPP

Let $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^v] \in \Re^{D \times vN}$ be a D-dimensional data space extracted from v-views of facial expressions. The k^{th} view of $\mathbf{X}$ *i.e.,* $\mathbf{X}^k$ is given by: $\mathbf{X}^k = \left\{\mathbf{x}_{ic}^k | i = 1, 2, ..., N; c = 1, 2, ..., C\right\}$, where $\mathbf{x}_{ic}^k$ indicates i^{th} sample of c^{th} class extracted from k^{th} view of facial images. The corresponding reduced features in common space is given by $\mathbf{Y} = [\mathbf{Y}^1, \mathbf{Y}^2, ..., \mathbf{Y}^v] \in \Re^{d \times vN}$, where $\mathbf{y}_{ic}^k$ of $\mathbf{Y}^k$ indicates i^{th} sample of c^{th} class in common space corresponding to $\mathbf{x}_{ic}^k$. Here, N is the number of samples in each of the views.

The proposed UMvDLPP approach aims to learn a set of v-linear transformation matrices $\left\{\mathbf{V}^k | k = 1, 2, ..., v\right\}$ as shown in Figure 5.2, which projects samples from all the views to the common discriminative space while preserving intra-class topology of intra-view and inter-view

[5] This work has been published in *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), ACM 2016* (Refer item 1 in Page 143 for details)

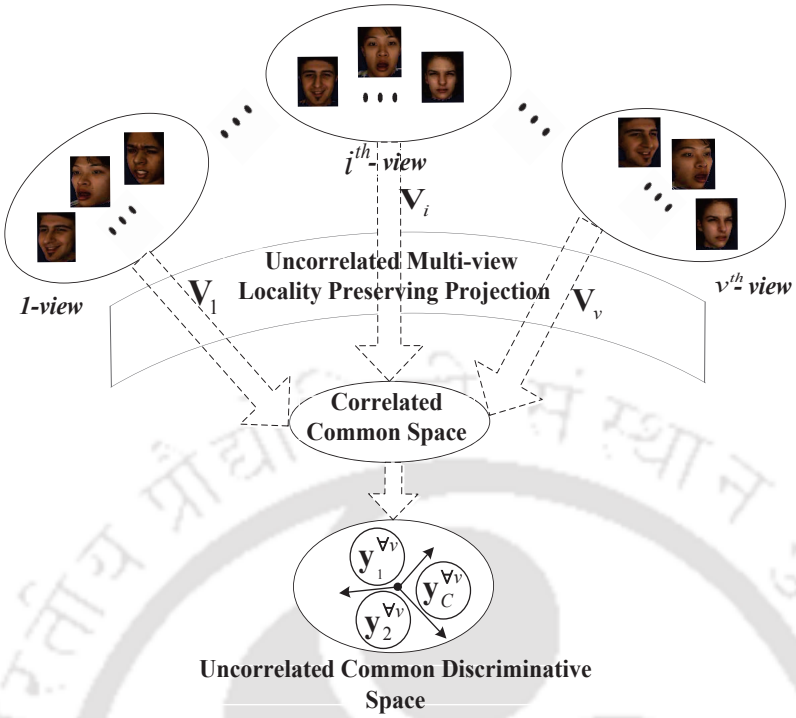


Figure 5.2: Overall representation of the proposed UMvDLPP method for multi-view facial expression recognition.

of the data space. Additionally, the proposed method maximizes the local between-class scatter matrix of intra-view and inter-view. This formulation is more suitable for multi-view facial expression recognition, as it can handle multi-modal characteristics of multi-view observations (Figure 5.1). Finally, the proposed objective function of UMvDLPP is formulated as a trace ratio minimization problem, and so it can be represented as:

$$\begin{aligned}\left[\mathbf{V}^1, \mathbf{V}^2, ..., \mathbf{V}^v\right]_{opt} &= \arg \min_{\mathbf{V}^1,\mathbf{V}^2,...,\mathbf{V}^v} \frac{tr\left(\mathbf{V}^T\mathbf{X}\mathbf{L}_{eq}\mathbf{X}^T\mathbf{V}\right)}{tr\left(\mathbf{V}^T\mathbf{X}\mathbf{B}_{eq}\mathbf{X}^T\mathbf{V}\right)} \\ &= \arg \min_{\mathbf{V}^1,\mathbf{V}^2,...,\mathbf{V}^v} \frac{tr\left(\mathbf{V}^T\mathbf{P}_{eq}\mathbf{V}\right)}{tr\left(\mathbf{V}^T\mathbf{Q}_{eq}\mathbf{V}\right)}\end{aligned} \tag{5.1}$$

where, $\mathbf{P}_{eq} = \mathbf{X}\mathbf{L}_{eq}\mathbf{X}^T$ and $\mathbf{Q}_{eq} = \mathbf{X}\mathbf{B}_{eq}\mathbf{X}^T$ are $Dv \times Dv$ matrices, which are represented as:

$$\mathbf{P}_{eq} = \begin{pmatrix} \mathbf{P}_{11} & \cdots & \mathbf{P}_{1v} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{v1} & \cdots & \mathbf{P}_{vv} \end{pmatrix} \text{ and } \mathbf{Q}_{eq} = \begin{pmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1v} \\ \vdots & \ddots & \vdots \\ \mathbf{Q}_{v1} & \cdots & \mathbf{Q}_{vv} \end{pmatrix}$$

The diagonal blocks of $\mathbf{P}_{eq}$ *i.e.,* $\mathbf{P}_{kk}$, $k = 1, 2, ..., v$, whose trace represent the sum of the distances between the samples of the same class of k^{th} view (intra-view), whereas trace of off-diagonal blocks $\mathbf{P}_{kl}$, $(k \neq l)$ indicates sum of the distances between the samples of the same class corresponding to two different views (inter-view) *i.e.,* k^{th} view and l^{th} view. Mathematically, we define the block matrix $\mathbf{P}_{kl}$ as:

$$\mathbf{P}_{kl} = \begin{cases} \mathbf{X}^k \mathbf{L}_k \mathbf{X}^{k^T}; \text{ if } k = l \\ \begin{bmatrix} \mathbf{X}^k & \mathbf{X}^l \end{bmatrix} \mathbf{L}_{kl} \begin{bmatrix} \mathbf{X}^k & \mathbf{X}^l \end{bmatrix}^T; \text{ if } k \neq l \end{cases} \tag{5.2}$$

where, $\mathbf{L}_k$ is the intra-view Laplacian matrix, and it is evaluated as $\mathbf{L}_k = \mathbf{D}_k - \mathbf{A}_k$. The elements of $\mathbf{A}_k$ (similarity matrix for k^{th} view) can be obtained by applying RBF kernel on i^{th} and j^{th} elements of the original data space $\mathbf{X}^k$. In this, $(i, j)^{th}$ element of $\mathbf{A}_k$ is defined as follows:

$$\mathbf{A}_{k,ij} = \begin{cases} \exp\left(-\frac{||\mathbf{x}_i^k - \mathbf{x}_j^k||^2}{\sigma^2}\right) & , \text{ if } c_i = c_j \\ 0, \text{ otherwise} \end{cases} \tag{5.3}$$

However, the inter-view Laplacian matrix $\mathbf{L}_{kl}$ is evaluated by using the joint observations of k^{th} view and l^{th} view of facial features.

Similarly, the diagonal blocks of $\mathbf{Q}_{eq}$ *i.e.,* $\mathbf{Q}_{kk}$ represents the intra-view local between-class scatter matrix, whereas off-diagonal block matrices $\mathbf{Q}_{kl}$ $(k \neq l)$ represent inter-view local between-class scatter matrix. Formulation of inter-view $\mathbf{Q}_{kl}$ can be mathematically defined as follows:

$$\mathbf{Q}_{kl} = \begin{cases} \frac{1}{2} \sum_{i,j=1}^{2N} \mathbf{B}_{ij} \left(\mathbf{x}_i^v - \mathbf{x}_j^v\right)\left(\mathbf{x}_i^v - \mathbf{x}_j^v\right)^T; v = k \text{ and } l, \ k \neq l \\ \frac{1}{2} \sum_{i,j=1}^{N} \mathbf{B}_{ij} \left(\mathbf{x}_i^v - \mathbf{x}_j^v\right)\left(\mathbf{x}_i^v - \mathbf{x}_j^v\right)^T; v = k \text{ or } l, \ k = l \end{cases} \tag{5.4}$$

where,

$$\mathbf{B}_{ij} = \begin{cases} \frac{\mathbf{A}_{ij}}{n_c}, \text{ if } c_i = c_j \\ 0 \ , \text{ if } c_i \neq c_j \end{cases} \tag{5.5}$$

The proposed objective function of UMvDLPP is further modified in the form of ratio trace problem. Since, the ratio trace problem can be converted into a generalized eigenvalue equation, hence there exists a global optimal solution [111]. The converted ratio trace form of UMvDLPP can be written as follows:

$$\left[\mathbf{V}^1, \mathbf{V}^2, ..., \mathbf{V}^v\right]_{opt} = \arg \min_{\mathbf{V}^1, \mathbf{V}^2, ..., \mathbf{V}^v} tr \left(\frac{\mathbf{V}^T \mathbf{P}_{eq} \mathbf{V}}{\mathbf{V}^T \mathbf{Q}_{eq} \mathbf{V}} \right) \tag{5.6}$$

Then, the d^{th} column of $\mathbf{V}$ *i.e.,* v_d can be obtained by solving the following generalized eigenvalue equation:

$$\mathbf{P}_{eq} v_d = \lambda_d \mathbf{Q}_{eq} v_d \tag{5.7}$$

where, λ_d is the d^{th} lowest eigenvalue. Finally, reduced feature vector onto the common space is obtained by using the following transformation:

$$\mathbf{Y}^k = \mathbf{V}^{k^T} \mathbf{X}^k; \ k = 1, 2, ..., v \tag{5.8}$$

Although the obtained common space is quite discriminative, the different components of the reduced feature vectors may be correlated, and so, we call this space as a Correlated Common Space (CCS), and it is denoted by $\mathbf{Y}_{ccs}$. In our proposed approach, instead of classifying directly from CCS, we first transformed features of the correlated common space $\mathbf{Y}_{ccs}$ to the Uncorrelated Common Space (UCS) $\mathbf{Y}_{ucs}$, and then classification is done. For this, UCS is obtained from the CCS with the help of the transformation matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_d]$ [115]. The columns of $\mathbf{U}$ (eigenvectors) are essentially the solutions of the following generalized eigenvalue equation corresponding to the first d lowest eigenvalues:

$$\mathbf{Y}_{ccs} \left(\mathbf{L}_{ccs} + \mathbf{B}_{ccs} \right) \mathbf{Y}_{ccs}^T \mathbf{u} = \lambda_{ccs} \mathbf{Y}_{ccs} \mathbf{G}_{ccs} \mathbf{Y}_{ccs}^T \mathbf{u} \tag{5.9}$$

where, $\mathbf{L}_{ccs}$ and $\mathbf{B}_{ccs}$ are Laplacian and between-class transformation matrices respectively. These matrices are obtained from the CCS. The matrix $\mathbf{G}_{ccs} = \mathbf{I} - (1/vN)\, \mathbf{e}\mathbf{e}^T$, where $\mathbf{I}$ is an identity matrix, and $\mathbf{e} = (1, 1, ..., 1)^T$. Therefore, the transformed UCS can be obtained by

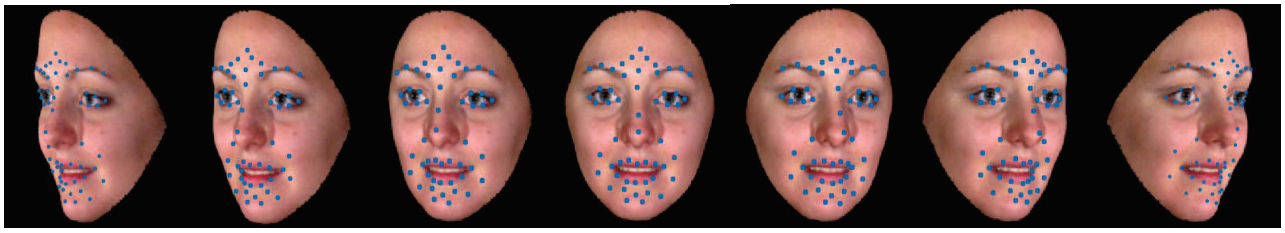


Figure 5.3: A part of BU3DFE dataset with 54 landmark points. Locations of these facial points show the informative regions of a multi-view facial image as suggested in [101].

linear projection:

$$\mathbf{Y}_{ucs} = \mathbf{U}^T \mathbf{Y}_{ccs} \tag{5.10}$$

Finally, we use kNN classifier for classification onto the uncorrelated common discriminative space. The learning of kNN is straightforward. During inference process, the test sample $\mathbf{x}_{test}^k$ of k^{th} view is first projected onto the CCS by using learned view-specific transformation matrix ($\mathbf{V}^k$) followed by a projection onto the UCS by using the transformation matrix $\mathbf{U}$. This sequence of projections can be represented as:

$$\mathbf{y}_{test}^k = \mathbf{U}^T \left(\mathbf{V}^{k^T} \mathbf{x}_{test}^k \right) \tag{5.11}$$

The class-label of the test sample $\mathbf{x}_{test}^k$ is obtained on the basis of the labels of k-nearest samples of $\mathbf{y}_{test}^k$ onto the uncorrelated common discriminative space.

5.3 Experimental Results

BU-3D Facial Expression (BU3DFE) dataset [37] is used to validate our proposed UMvDLPP-based method. This dataset comprises of 3D facial images of seven basic expressions *i.e.*, "happy", "surprise", "fear", "anger", "disgust", "sad" , and "neutral" expressions. Expressions of BU3DFE dataset are captured at four different levels of intensity levels ranging from onset/offset level to peak level of expressions. For our experimentation, 2D facial images corresponding to seven views *i.e.*, $-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, and $45°$ yaw/pan angles are

considered. A part of BU3DFE dataset is shown in Figure 5.3, where a single subject is showing "happy" expression at seven different pan angles. In total, 1800 images/view *i.e.*, $1800 \times 7 = 12600$ images are considered for performance evaluation of our method. Each view comprises of images of six basic expressions *i.e.*, *anger, disgust, fear, happy, sad, and surprise*, where 300 images are considered for each of the classes of expressions. The face part of 2D textured expressive images are manually cropped, and down-sampled to get image size of 160×140. Subsequently, 54 facial landmark points are localized in the face images as per our earlier proposal discussed in Chapter 3 and Chapter 4 [101]. Localization of facial points for the images of views $-45°$ and $45°$ are carried out manually, whereas images of other views ($-30°$, $-15°$, $0°$, $15°$, and $30°$) are automatically annotated using Active Appearance Model (AAM) [89]. Finally, a grid of 15×15 is considered at each of the facial points to extract feature vector from each of the salient regions of a face. In our method, LBP^{u2} operator is applied on sub-blocks around each of the landmark points to extract the feature vectors. As LBP^{u2} gives a 59-dimensional feature vector for a facial sub-region, and so, the overall feature dimension is $54 \times 59 = 3186$. Subsequently, view-wise alignment is done using procrustes analysis [129]. However in our case, instead of using a set of landmark points to align facial images, we considered feature vector of each sub-block as a point in a 59^{th}-dimensional space, and then procrustes analysis is applied to align each of the faces. The first level of dimensionality reduction of LBP-based appearance feature is performed by using principal component analysis, so that 95% of the total variance of the data is preserved. Finally, the proposed UMvDLPP is applied on the principal components to obtained the respective uncorrelated discriminative common space.

All the experiments are carried out using 10-fold leave-one-out-cross-validation strategy, and hence we randomly divide images of each of the views into 10 subsets. Out of which, nine subsets are used to train the model, whereas the remaining set of each of the views is used for testing. Finally, we use 1-nearest neighbour (1-NN) classifier onto UCS to evaluate the performance of the proposed UMvDLPP method.

We performed four experiments to validate our proposed method. In our first experiment, we found out optimal dimensionality of the common space. Subsequently, rest of the three

experiments are performed on this optimal common space to evaluate the performances of the proposed method. In our second experiment, inter-class and intra-class separation of different class of expressions are analyzed. Confusion matrices for each of the views are analyzed in our third experiment. Finally, comparison of the proposed method with the existing learning-based methods is presented in the forth experiment. The detail of the above mentioned experiments are discussed below.

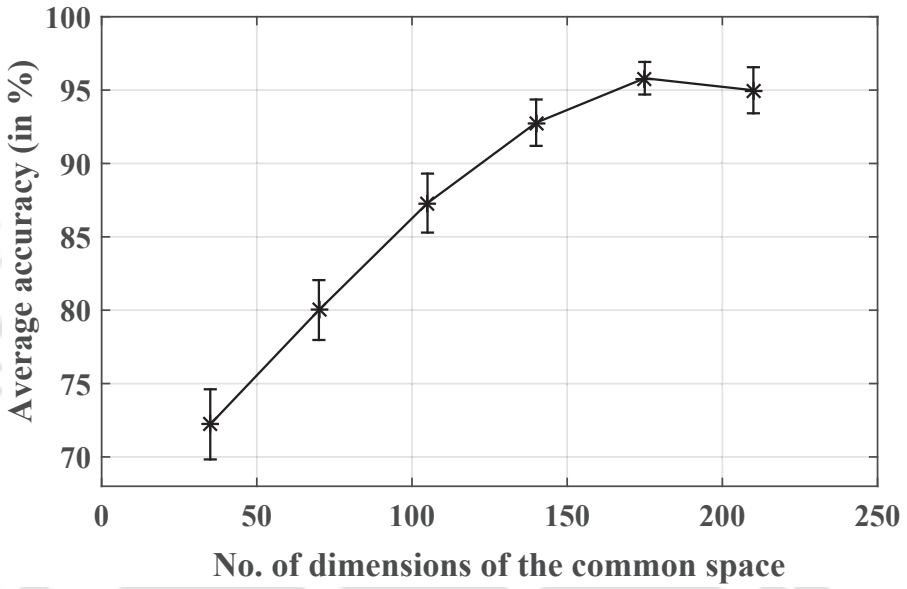


Figure 5.4: Effect of dimensionality of common space on average accuracy of UMvDLPP framework across all the seven poses of BU3DFE dataset.

Experiment-1: The aim of our proposed method is to find an optimal uncorrelated discriminative space. In our first experiment, we analyzed how the classification accuracy depends on dimensionality of the common space. In this view, average classification accuracy across all the seven views are calculated by varying the dimensionality of the common space as shown in Figure 5.4. Horizontal axis of Figure 5.4 represents number of dimensions of the common space. This analysis shows that UMvDLPP gives highest recognition rate at the dimension of 175. Apparently, a 175 dimensional space captures almost 98% variance of the data. Hence in our rest of the analysis, we used 175^{th} dimensional uncorrelated common space to obtain accuracy of each of the views of multi-view facial images.

Experiment 2: In this experiment, we showed intra-class and inter-class variations obtained from the samples of the uncorrelated common space. Figure 5.5 shows intra-class and inter-class variations for each of the classes. The overall intra-class variations of c^{th} class of expression is obtained by taking average of all intra-class distances obtained across all the views of the expression under consideration. Intra-class distance of c^{th} class of expression for v^{th} view is obtained by taking mean of pair-wise distances between samples of the c^{th} class. On the other hand, inter-class variation is obtained using *One-vs-All* strategy, where one of the samples in the pair belongs to c^{th} class, and other sample belongs to the rest of the classes. Finally, average is taken for all the poses to obtain the overall inter-class variations. It can be clearly observed from the Figure 5.5 that inter-class distances are higher than the intra-class for each of the expressions, which ensures separability of different classes. It is also observed that intra-class

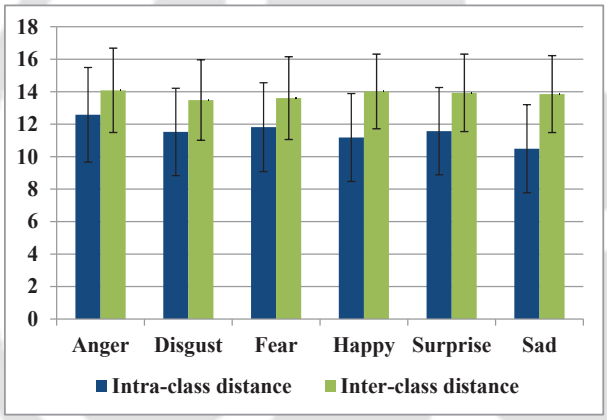


Figure 5.5: Overall separation between intra-class and inter-class of each expression across all the views .

and inter-class differences are comparatively more in "happy", "sad", and "surprise" expressions as compared to "anger", "fear", and "disgust" expressions. To validate this distribution of training samples on a 3D plane obtained from the samples of uncorrelated common space is plotted. PCA is used to project samples from higher dimensional space (UCS) to the lower dimensional space (3D space) by selecting first three principal components corresponding to the first three largest eigenvalues. The left column of Figures 5.6 5.7 and 5.8 show distribution of samples for different views (0°, −15°, −30°, −45°, 15°, 30°, 45°) obtained by our proposed UMvDLPP-based method. On the other hand, right column of the above mentioned figures

shows distribution of samples obtained by state-of-the-art MvDA-based method. These plots clearly show that the proposed method represents data in a more compact way. Furthermore, the proposed method also gives slightly better separability in a 3D-space, and hence, even better class separation may be achieved in a high dimensional space. Figure 5.9 shows distribution of samples from all the views ($-45°$, $-30°$, $-15°$, $0°$, $45°$, $30°$, and $15°$) onto the learned common space by our proposed UMvDLPP and MvDA-based methods. This also indicates that proposed method learned better discriminative common space as compared to MvDA-based approach.

Experiment 3: In this experiment, performance of our proposed method is analyzed for different views. Table 5.1 shows the confusion matrices for each of the views ($-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, and $45°$). It is observed that UMvDLPP gives comparatively better recognition rates for negative views as compared to positive views of facial images. This analysis once again validates the findings of the psycho-visual experiments presented in [151]. The psycho-visual experiments presented in [151] highlighted the importance of left part of a face for negative expressions like "disgust" and "anger", and the importance of right part of a face for positive expressions like "happy".

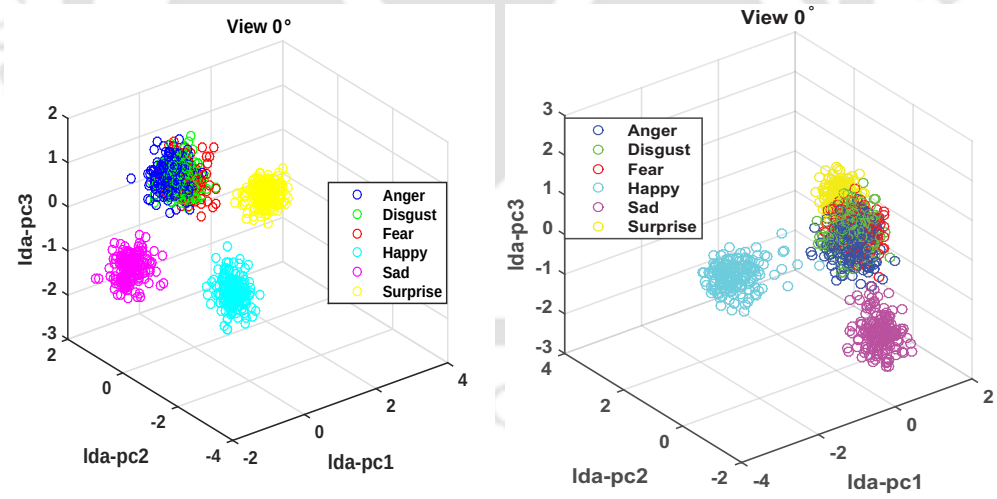


Figure 5.6: Distribution of sample points for frontal view facial images (0° view) obtained by UMvDLPP [Left] and MvDA [Right] respectively.

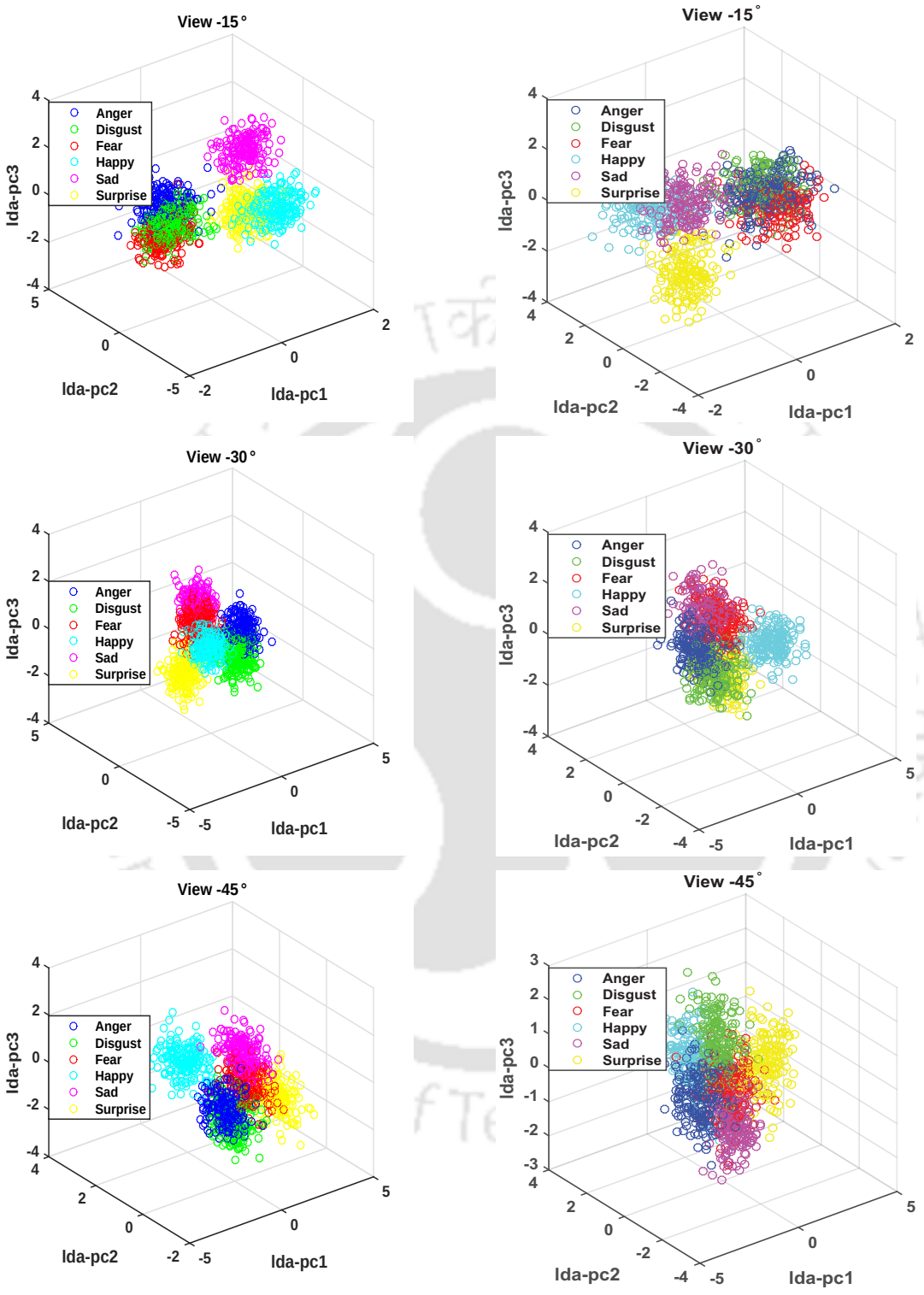


Figure 5.7: Distribution of sample points for facial images of $-15°$, $-30°$, and $-45°$ views obtained by UMvDLPP [Left column] and MvDA [Right column] respectively.

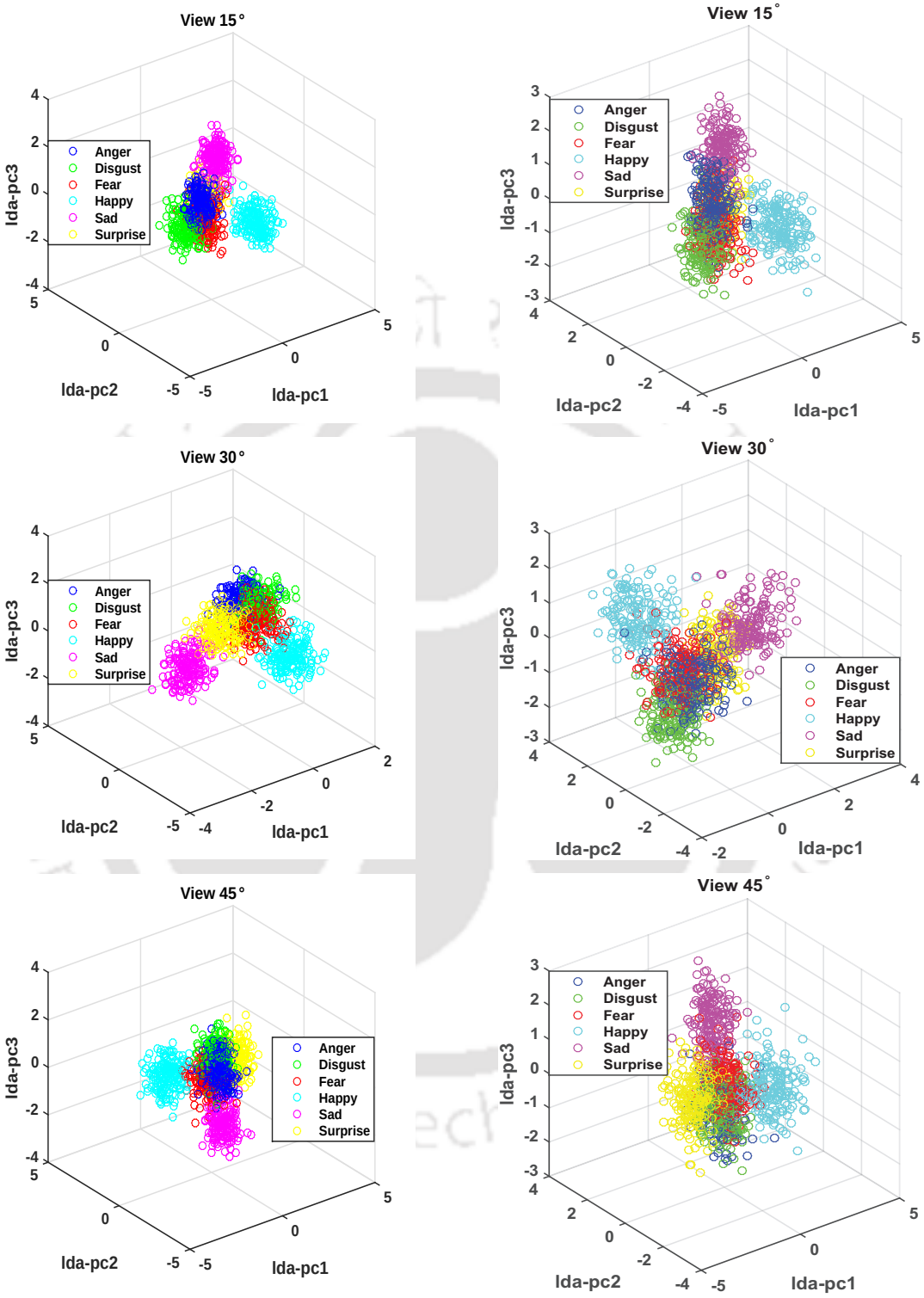


Figure 5.8: Distribution of sample points for facial images of 15°, 30°, and 45° views obtained by UMvDLPP [Left column] and MvDA [Right column] respectively.

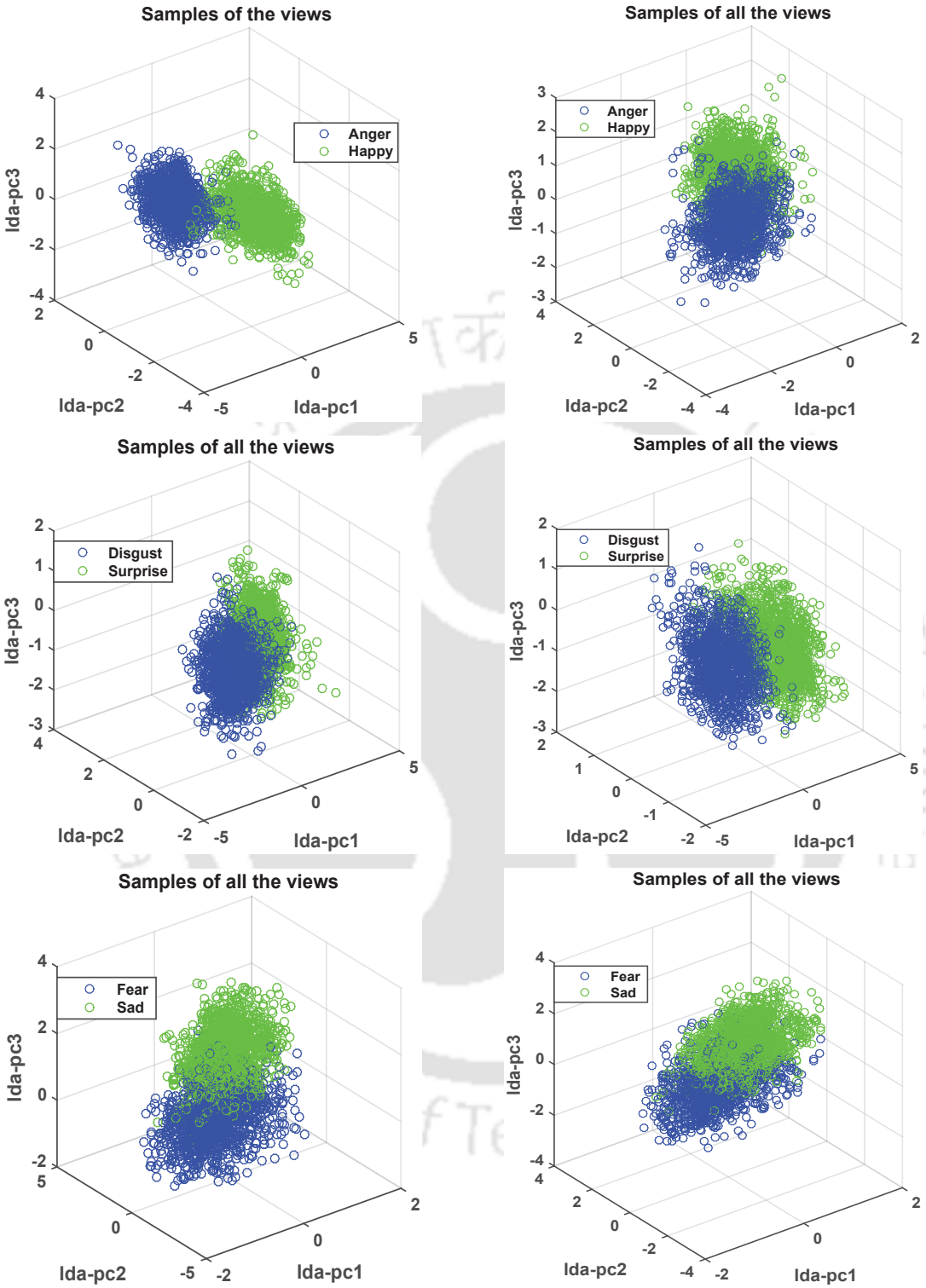


Figure 5.9: Distribution of samples of all the views ($-45°$, $-30°$, $-15°$, $0°$, $45°$, $30°$, and $15°$) onto the learned common space using our proposed UMvDLPP method [Left column] and the stae-of-the-art MvDA method [Right column], respectively.

Table 5.1: View-wise confusion matrix for six basic expressions

Confusion matrix obtained using LBP+PCA+UMvDLPP features							
View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of 0°	Anger	98.48	1.52	0	0	0	0
	Disgust	0	98.48	1.52	0	0	0
	Fear	0	1.52	98.48	0	0	0
	Happy	0	0	0	98.48	1.52	0
	Sad	0	0	0	0.12	99.88	0
	Surprise	0	0	0.08	0	0	99.92

View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of −45°	Anger	95.45	0	0	4.55	0	0
	Disgust	0	99.10	0	0.9	0	0
	Fear	0	3.03	96.97	0	0	0
	Happy	0	0.36	0	99.64	0	0
	Sad	0	0	1.52	0	96.96	1.52
	Surprise	0	0	3.03	1.51	0	95.46

View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of −30°	Anger	95.45	3.03	0	0	1.52	0
	Disgust	0	98.49	1.51	0	0	0
	Fear	3.00	0	97.00	0	0	0
	Happy	1.52	0	0	98.48	0	0
	Sad	1.52	0	0	0	98.48	0
	Surprise	0	0	1.52	0	0	98.48

View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of −15°	Anger	98.48	0	0	0	1.52	0
	Disgust	1.52	96.96	0	0	0	1.52
	Fear	1.52	1.52	96.96	0	0	0
	Happy	0	1.52	0	98.48	0	0
	Sad	0	0	3.04	0	96.96	0
	Surprise	0.3	0	0	0	0	99.70

Pan of 45°	Anger	80.32	4.54	1.51	0	13.63	0
	Disgust	3.03	90.93	1.51	1.51	1.51	1.51
	Fear	7.57	3.03	80.31	0	6.06	3.03
	Happy	3.03	0	1.51	95.46	0	0
	Sad	0	0.3	0	0	99.70	0
	Surprise	0	4.54	3.03	0	0	92.43
Pan of 30°	Anger	92.43	6.06	0	0	1.51	0
	Disgust	1.51	90.92	3.03	1.51	0	3.03
	Fear	4.54	3.03	87.89	0	0	4.54
	Happy	0	1.51	3.03	95.46	0	0
	Sad	3.03	3.03	0	0.01	93.93	0
	Surprise	3.03	0	3.03	0	1.51	92.43
Pan of 15°	Anger	96.96	1.52	0	1.52	0	0
	Disgust	1.51	90.92	4.54	0	3.03	0
	Fear	3.04	7.57	89.39	0	0	0
	Happy	0	0.2	0	99.80	0	0
	Sad	1.52	1.52	0	0	96.96	0
	Surprise	1.8	0	0	0	0	98.20

Experiment 4: Table 5.2 shows comparative results of the proposed method with the existing multi-view learning-based methods. The view-wise average accuracy and average recognition rates for all the seven poses are used to compare the performance of the proposed method with the existing methods. The performance of kNN ($k = 1$) is considered as a baseline approach, and its performance is evaluated on the original feature space. LDA [114] and LPP [152], can capture 98% variance of the data. Finally, 1-NN was used to evaluate their view-specific accuracy. In a multi-view framework, expressions from all the views are used to learn a common discriminative subspace. Generalized Multi-view Analysis (GMA) [109] is a well known approach which can find a common discriminative space. Further, GMA-based approach can be integrated with PCA, LDA, Canonical Correlation Analysis (CCA)-based approaches. The corresponding methods are termed as GMPCA, GMLDA, and GMCCA respectively. The average recognition rates obtained using the variants of GMA are slightly better than the view-wise LDA and LPP-based approaches. Additionally, we extended CCA and LDA in a pair-wise

fashion to recognize expressions from multi-view facial images. In this view, pair-wise CCA (PW-CCA) [153] and pair-wise LDA (PW-LDA) approaches are investigated. The pair-wise methods learn ${}^{v}C_2$ number of common subspaces – each of the subspaces represents a common discriminative subspace for two different views of facial images. Finally, average accuracy is calculated for obtaining view-wise accuracy. Table 5.2 shows that pair-wise classification strategy gives slightly less recognition rates as compared to multi-view learning-based approaches. The performance of MvDA [111] is closer to the proposed UMvDLPP-based approach, however UMvDLPP gives better average accuracy as compared to MvDA. This improvement may be because of the ability of UMvDLPP to capture optimal discriminative directions of multimodal data, which may not be efficiently captured by MvDA. Hence, MvDA is sub-optimal for multi-view facial expression recognition. In a nutshell, our proposed UMvDLPP gives better performance than the existing learning-based approaches.

Table 5.2: Comparative performance of proposed UMvDLPP method with the existing learning-based methods.

Methods	View-wise Recognition Rate (RR) (in %)							
	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	Avg RR
kNN	65.15	78.53	78.78	85.85	79.54	65.40	66.41	74.24
LDA	84.84	95.70	86.11	82.82	88.38	78.28	63.88	82.86
LPP [152]	94.19	94.69	90.65	93.68	89.39	87.12	85.85	90.80
GMPCA [109]	91.16	87.37	98.98	95.20	87.87	83.83	83.08	89.64
GMLDA [109]	91.91	98.48	96.71	97.97	89.89	81.06	82.32	91.19
GMCCA [109]	90.65	96.71	97.47	97.47	90.40	85.35	85.35	91.91
PW-CCA [153]	90.00	83.33	90.00	96.67	73.33	80.00	76.67	84.28
PW-LDA	93.77	95.15	88.72	91.83	86.57	81.43	80.85	88.33
MvDA [111]	95.70	96.21	94.44	97.72	94.69	92.92	89.64	94.48
MvDLPP	98.73	98.48	95.90	96.56	94.34	91.81	88.68	94.94
UMvDLPP	**97.47**	**98.23**	**96.71**	**99.24**	**93.93**	**93.93**	**90.15**	**95.67**

5.4 Conclusion

In this chapter, we addressed the problem of recognizing facial expressions from multi-view facial images. In this view, a linear non-parametric-based approach termed as uncorrelated multi-view discriminant locality preserving projection (UMvDLPP) analysis is proposed. The

proposed objective function of UMvDLPP learns a uncorrelated common discriminative subspace from multiple observations. Moreover, the proposed method is quite efficient to capture discriminative subspace even when the data exhibits multi-modal characteristics. Performance of the proposed method and other existing state-of-the-art learning-based approaches are evaluated on BU3DFE dataset. In all the cases, uniform LBP-based features are extracted around 54 informative facial landmark points as discussed in Chapter 3 and Chapter 4. To validate our proposed method, four experiments are performed, and finally average accuracy is used to compare the proposed method with the existing multi-view learning-based approaches. In our first experiment, an optimal uncorrelated common space is extracted, and subsequently rest of the experiments are performed on this common space. The compactness and separability of data are analyzed in our second experiment. In the third experiment, the performance of our proposed scheme is quantified, and the role of two sides of a face in different facial expressions is analyzed. Finally, comparative performance of UMvDLPP with existing learning-based methods is presented in our fourth experiment to show the efficacy of our proposed scheme.

Hierarchical classification strategy improves classification accuracy in recognizing frontal view facial expressions, and this has been investigated in [42] for expression recognition. No prior works on hierarchical-based approaches for recognizing multi-view facial expressions are reported in the literature. So, there is scope to implement our proposed UMvDLPP approach in a hierarchical fashion, and this has been investigated in Chapter 6.

6

Multilevel Uncorrelated Discriminative Shared Gaussian Process for MvFER

In MvFER, discriminative shared Gaussian process latent variable model (DS-GPLVM) gives better performance than the existing linear and non-linear multi-view learning-based methods. However, Laplacian-based prior used in DS-GPLVM only captures topological structure of data without considering inter-class separability of the data, and hence the extracted latent space is sub-optimal. So, we proposed a multi-level uncorrelated DS-GPLVM (ML-UDSGPLVM) model which searches a common uncorrelated discriminative latent space learned from multiple observable spaces. In this, a novel prior is proposed, which not only depends on the topological structure of intra-class data but also on the local-between-class-scatter-matrix of the data onto the latent manifold. The proposed approach employs a hierarchical framework, in which, expressions are first divided into three sub-categories. Subsequently, each of the sub-categories are further classified to identify the constituent basic expressions. In this chapter, UMvDLPP proposed in chapter 5 is also implemented in a hierarchical fashion, and its performance is compared with ML-UDSGPLVM and other existing approaches. Experimental results show the efficacy of the proposed methods.

6.1 Introduction

In Chapter 5, we recognized facial expressions from multi-view facial images. For this, Uncorrelated Multi-view Discriminant Locality Preserving Projection (UMvDLPP) analysis was proposed. As discussed in Chapter 5, UMvDLPP gives comparatively better performance than the existing linear learning-based methods. Also, performance of UMvDLPP is comparable to Discriminative Shared Gaussian Process Latent Variable Model (DS-GPLVM) [1]. However, DS-GPLVM uses a very low dimensional space ($d = 5$) for classification, whereas UMvDLPP or MvDA uses 175 dimensional space for multi-view expression classification. Hence, testing of DS-GPLVM takes comparatively less time than UMvDLPP for recognition. Moreover, DS-GPLVM also aims to find a common shared space for recognition. Motivated by these facts, DS-GPLVM framework is employed in our proposed scheme by embedding additional information in the prior of existing DS-GPLVM.

In [1], it is considered that different views of a facial expression are just different manifestations of a same facial expression. Hence, correlations between different views of facial expressions are exploited during learning of a common shared space. More specifically, DS-GPLVM generalizes discriminative-GPLVM (D-GPLVM) [154] along with shared Gaussian process [155, 156] to learn a discriminative manifold. Nevertheless, discriminative nature of Gaussian process depends on the prior. In [154], a prior based on Linear Discriminant Analysis (LDA) [114] was proposed to replace a standard spherical Gaussian-based prior. A more general prior based on the notion of graph Laplacian matrix was proposed in [112] [113]. However, their prior does not include between-class separation, and hence latent manifold obtained by these approaches is sub-optimal. Laplacian-based prior is further generalized for multi-view FER in DS-GPLVM [1], and so shared space obtained by this approach may not be optimal. Also, samples of latent space may be correlated, which may further affect the classification accuracy of the DS-GPLVM-based multi-view FER [115].

To address all the above issues, we proposed uncorrelated discriminative shared Gaussian process latent variable model (UDSGPLVM), and subsequently it is extended to a hierarchical framework termed as multi-level UDSGPLVM (ML-UDSGPLVM) for multi-view FER. In ML-UDSGPLVM, a more generalized discriminative prior is proposed, which is based on graph

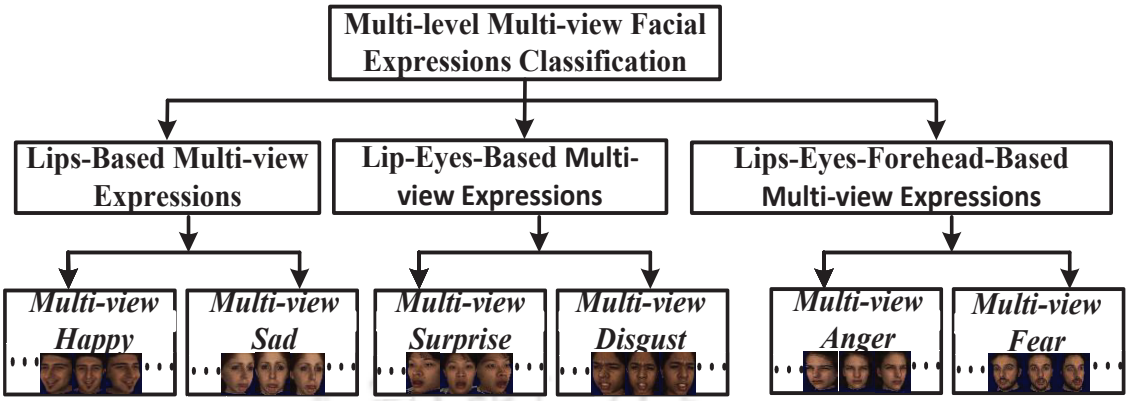


Figure 6.1: Different levels of multi-level multi-view facial expression classifications.

Laplacian matrix [112] and a transformation matrix. The transformation matrix is derived from the local between-class scatter matrix (LBCSM) of data [3, 116]. The advantage of our proposed prior is that it can provide better separation of data onto the latent manifold. The proposed prior depends on both intra-class geometric structure of data captured by Laplacian matrix and local inter-class variability of data provided by LBCSM. Hence, the proposed prior is more efficient than Laplacian-based prior [1]. Moreover, discriminative non-linear latent manifold (feature space) obtained by Gaussian process might be correlated, and hence classification performed directly on correlated manifold may reduce classification accuracy [115]. So in our proposed ML-UDSGPLVM approach, correlated shared manifold is first transformed into uncorrelated shared manifold, and then classification is performed.

To implement a multi-level classification scheme, expressions of multi-view face images are recognized in two steps as shown in Figure 6.1. In the first step, all the basic expressions are grouped into three categories, namely Lips-based, Lips-Eyes-based, and Lips-Eyes-Forehead-based expressions. This classification of expressions is done with the help of the regions of a face which mostly contribute to an expression. Then, category-wise training and testing are performed using the proposed UDSGPLVM. In the second step, a separate UDSGPLVM is applied on each of the sub-categories to further classify the basic expressions embedded in the above mentioned three classes of expressions. The proposed 2-level-UDSGPLVM follows this approach as against the method used in 1-level-DS-GPLVM or simply DS-GPLVM.

In our proposed method, we employed our earlier developed face model [101] to extract

features only from the informative regions of a face, as most discriminative features are only attainable from the informative/active regions of a face [47, 102, 103]. The proposed method is elaborately discussed in the following Sections.

6.2 Proposed Methodology

Shape-based method is employed in our proposed method to extract texture features from the active/informative regions of a face. We proposed to use our earlier developed face model, as it was derived from informative regions of a face [101]. Subsequently, LBP features are extracted from a 15×15 block around each of the facial points of our proposed face model. Next, expressions are divided into three classes based on the movements of lips, eyes, and forehead as stated in [42, 157]. The corresponding reduced non-linear subspace is learned using 1-UDSGPLVM as shown in Figure 6.2 (a). Subsequently, a 2-UDSGPLVM is learned for each of the expressions embedded in each of the sub-categories. Hence, three different 2-UDSGPLVMs have to be learned for final level of classification. The class-label of the test sample obtained by the first-level of ML-UDSGPLVM and kNN *i.e.,* 1-UDSGPLVM+kNN is used to select a specific 2-UDSGPLVM out of three 2-UDSGPLVMs. So, first-level of classification is performed using 1-UDSGPLVM and kNN. The first-level of classification is basically a three-class problem, and hence, the classifier identifies the appropriate sub-category. Any specific expression is finally identified by 2-UDSGPLVM and kNN. Our proposed ML-UDSGPLVM is discussed in the following Section.

6.3 Proposed ML-UDSGPLVM

In our method, a more accurate low-dimensional manifold is derived for multi-view FER. We first give a brief overview of DS-GPLVM [1]. The impact of the state-of-the-art priors on latent manifold is analyzed, and then we proposed a new prior to nullify some of the limitations of the existing priors. Finally, we introduce our proposed ML-UDSGPLVM model as shown in Figure 6.3, in which a more generalized discriminative prior is proposed. Also, uncorrelated constraint onto the latent manifold is imposed. All the steps of ML-UDSGPLVM are discussed below.

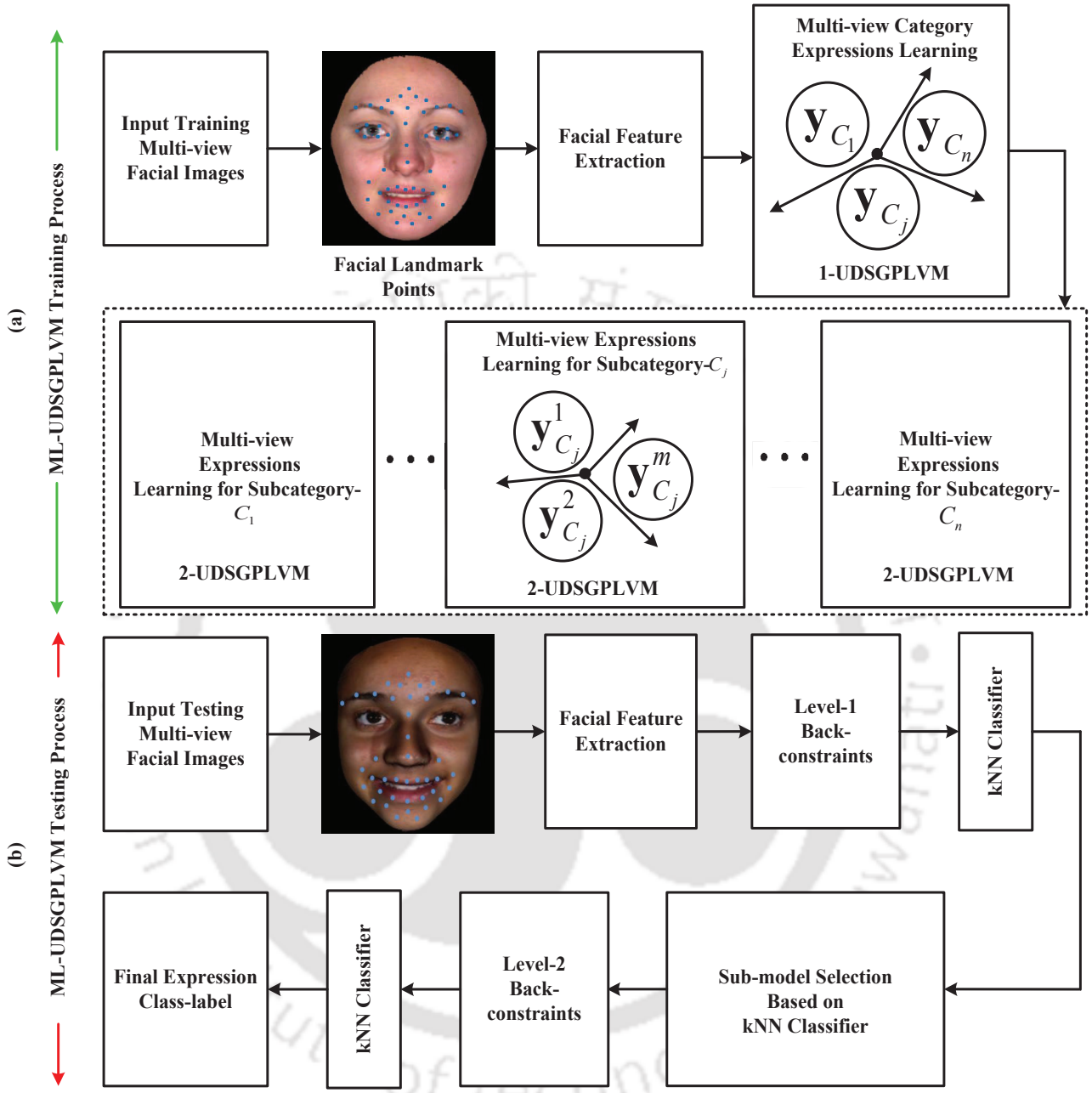


Figure 6.2: Proposed ML-UDSGPLVM for multi-view expression recognition: (a) training phase includes facial feature extraction and non-linear dimensionality reduction using *l*-UDSGPLVM. 1-UDSGPLVM learns first level of discriminative features for group-level facial expression classification, and 2-UDSGPLVM comprises of distint features for constituent expressions of the respective sub-groups, (b) classification stages of the proposed scheme.

6.3.1 DS-GPLVM

DS-GPLVM is a state-of-the-art approach for multi-view FER [1]. More specifically, DS-GPLVM generalizes D-GPLVM [154] using the framework of shared Gaussian process [155,156] to simultaneously learn a single non-linear discriminative manifold of multiple observation spaces. The problem formulation of DS-GPLVM as a multi-view FER can be stated as follows:

Let $\mathbf{X} = \left\{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^V\right\}$ be the set of V observation spaces of size $VN \times D$, where N is the number of samples in each of the observation spaces and D is the dimension of each feature vector. Then, the objective of DS-GPLVM is to learn a d-dimensional manifold $\mathbf{Y} \in \mathbb{R}^{N \times d}$ with $d << D$, which is assumed to be the shared information across all the views. The learning of low-dimensional manifold $\mathbf{Y}$ of DS-GPLVM and its mapping to the v^{th} observation space $\mathbf{X}^v$ is modeled using the framework of a shared Gaussian process. More specifically, it tries to learn the covariance function $k\left(\mathbf{y}_i, \mathbf{y}_j\right)$ of the shared manifold. In shared Gaussian process, each observation space is generated from the shared manifold via a separate Gaussian process, and hence the joint likelihood of the observed $\mathbf{X}$ given the shared manifold $\mathbf{Y}$ is factorized as follows:

$$p\left(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}\right) = p\left(\mathbf{X}^1|\mathbf{Y}, \boldsymbol{\theta}_1\right) \cdots p\left(\mathbf{X}^V|\mathbf{Y}, \boldsymbol{\theta}_V\right) \tag{6.1}$$

where, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_V\}$ are the kernel parameters of the shared observation space. The v^{th} factor of Eqn. (6.1) represents likelihood of v^{th} observation space $\mathbf{X}^v$ given the shared manifold $\mathbf{Y}$ *i.e.,* $p\left(\mathbf{X}^v|\mathbf{Y}, \boldsymbol{\theta}_v\right)$, which is defined as:

$$p\left(\mathbf{X}^v|\mathbf{Y}, \boldsymbol{\theta}_v\right) = \frac{1}{\sqrt{(2\pi)^{ND}|\mathbf{K}_v|^D}} \exp\left\{-\frac{1}{2} tr\left(\mathbf{K}_v^{-1}\mathbf{X}^v\mathbf{X}^{v^T}\right)\right\} \tag{6.2}$$

where, $\mathbf{K}_v$ is the kernel covariance matrix associated with v^{th} view of input space $\mathbf{X}^v$, whose $(i,j)^{th}$ element can be obtained by using the covariance function $k\left(\mathbf{y}_i, \mathbf{y}_j\right)$ defined as the sum of the Radial Basis Function (RBF) kernel, bias, and noise term. Hence, $k\left(\mathbf{y}_i, \mathbf{y}_j\right)$ can be

represented as follows:

$$k\left(\mathbf{y}_i, \mathbf{y}_j\right) = \theta_{v1} \exp\left(-\frac{\theta_{v2}}{2}||\mathbf{y}_i - \mathbf{y}_j||^2\right) + \theta_{v3} + \frac{\delta_{i,j}}{\theta_{v4}} \tag{6.3}$$

where, $\boldsymbol{\theta}_v = \{\theta_{v1}, \theta_{v2}, \theta_{v3}, \theta_{v4}\}$ are the kernel parameters of covariance function and $\delta_{i,j}$ is the Kronecker delta function. Finally, the distribution of shared manifold $\mathbf{Y}$ can be obtained by imposing a prior $p\left(\mathbf{Y}\right)$ over the shared manifold, and then applying the Bayes law. Thus, the posterior distribution of $\mathbf{Y}$ given $\mathbf{X}$ can be written as follows:

$$\begin{aligned} p\left(\mathbf{Y}, \boldsymbol{\theta}|\mathbf{X}\right) &= \frac{p\left(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}\right) p\left(\mathbf{Y}\right)}{p\left(\mathbf{X}\right)} \\ &\propto p\left(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}\right) p\left(\mathbf{Y}\right) \end{aligned} \tag{6.4}$$

The learning of the shared manifold is accomplished by minimizing the negative log-likelihood of the posterior distribution given in Eqn. (6.4) with respect to the latent positions of the shared manifold $\mathbf{Y}$. The negative log-likelihood of Eqn. (6.4) can be written as:

$$L_s = \sum_{v=1}^{V} L_v - \log\left(p\left(\mathbf{Y}\right)\right) \tag{6.5}$$

where, L_v is given by:

$$L_v = \frac{D}{2} \ln|\mathbf{K}_v| + \frac{1}{2} tr\left(\mathbf{K}_v^{-1} \mathbf{X}^v \mathbf{X}^{v^T}\right) + \text{constant} \tag{6.6}$$

6.3.2 Effect of priors on Gaussian process

The effectiveness of GPLVM depends on the kind of prior used in the manifold. In this direction, first attempt was explored in [154], where a simple spherical Gaussian prior is replaced by a discriminative prior based on LDA. Hence, it maximizes the between-class separability ($\mathbf{S}_b$) and minimizes the within-class separability ($\mathbf{S}_w$) of the latent space. The LDA-based prior is defined as:

$$p\left(\mathbf{Y}\right) = \frac{1}{Z_g} \exp\left\{-\frac{1}{\sigma_g} J^{-1}\left(\mathbf{Y}\right)\right\} \tag{6.7}$$

where, $J(\mathbf{Y}) = tr(\mathbf{S}_w^{-1}\mathbf{S}_b)$. In [113], a more general prior based on the notion of graph Laplacian matrix has been proposed. The Laplacian matrix of v^{th} view is defined as $\mathbf{L}^v = \mathbf{D}^v - \mathbf{W}^v$, where $\mathbf{D}^v$ is a diagonal matrix with $\mathbf{D}^v_{ii} = \sum_j \mathbf{W}^v_{ij}$. The weight $\mathbf{W}^v_{ij}$ is defined as:

$$\mathbf{W}^v_{ij} = \begin{cases} \exp\left(-\frac{||\mathbf{x}^v_i - \mathbf{x}^v_j||^2}{\sigma^v}\right) & ; \text{ if } c_i = c_j \\ 0 \ ; \text{ otherwise} \end{cases} \tag{6.8}$$

Also, [1] generalizes the Laplacian-based prior to obtain a prior for multi-view facial images. The net Laplacian matrix $\mathbf{L}_{net}$ in [1] is obtained by summing all the normalize Laplacian matrices corresponding to each of the views. Hence, mathematically $\mathbf{L}_{net}$ can be represented as:

$$\mathbf{L}_{net} = \mathbf{L}^1_{nor} + \mathbf{L}^2_{nor} + \cdots + \mathbf{L}^V_{nor} + \xi\mathbf{I} \tag{6.9}$$

where,

$$\mathbf{L}^v_{nor} = (\mathbf{D}^v)^{-1/2}\mathbf{L}^v(\mathbf{D}^v)^{-1/2}$$

Here, $\mathbf{I}$ indicates the identity matrix, and ξ is the regularization parameter which ensures positive-definiteness of $\mathbf{L}_{net}$ [158]. Finally, the discriminative shared-space prior is defined as:

$$p(\mathbf{Y}) = \prod_{v=1}^{V} p(\mathbf{Y}|\mathbf{X}^v)^{\frac{1}{V}} = \frac{1}{V.Z_d}\exp\left\{-\frac{\beta}{2}tr\left(\mathbf{Y}^T\mathbf{L}_{net}\mathbf{Y}\right)\right\} \tag{6.10}$$

where, Z_d is a normalization constant and β (reciprocal of the variance) is the precision parameter.

6.3.3 Proposed ML-UDSGPLVM model

In the previous section, we introduced the impact of state-of-the-art priors on Gaussian processes. In this section, we derive a more generalized discriminative prior function [6]. Also, influences of the prior function on the likelihood function are analyzed to obtain a more accurate posterior distribution. The prior based on Laplacian matrix ($\mathbf{L}_{net}$) given in Eqn. (6.10)

[6]This work is in under review in *IEEE Transaction on Circuits Systems and Video Technology 2016* (**(Date of communication : 03-Nov-2016)**)

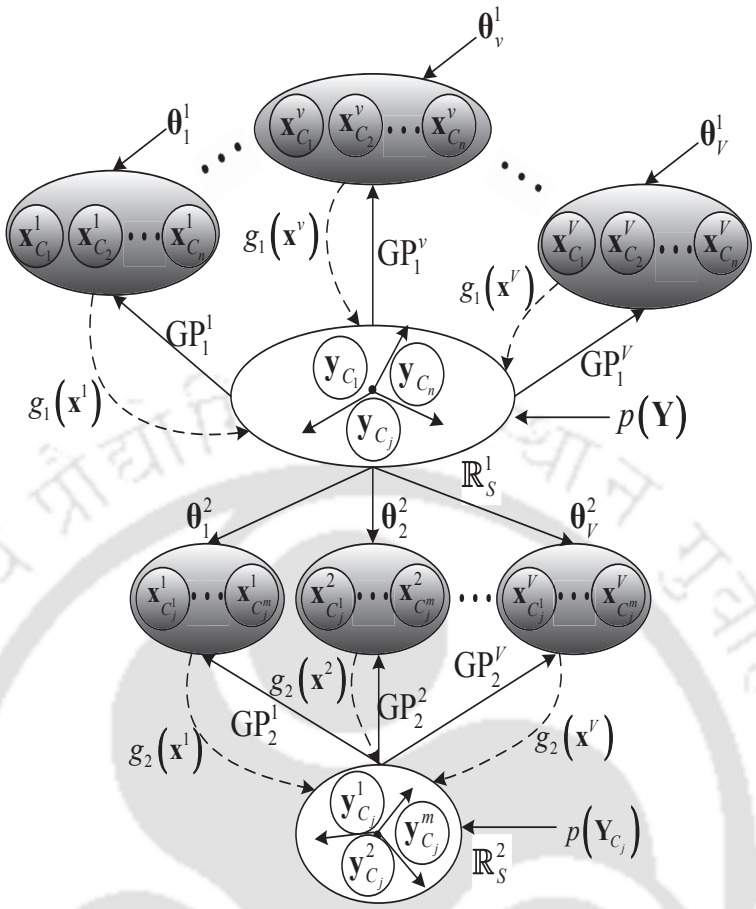


Figure 6.3: Proposed ML-UDSGPLVM.

essentially preserves the within-class geometric structure of the data. It uses RBF kernel to obtain weights between the data samples. So, it can also handle the multi-modalities present in the data. However, this approach did not consider the impact of between-class variability while defining the prior, and hence the prior proposed in Eqn. (6.10) makes the Gaussian process suboptimal for classification. But, the impact of between-class scatter matrix is crucial for all sorts of classification problems. So for our proposed prior, we incorporate a centering transformation matrix $\mathbf{B}$. This matrix is derived based on local between-class scatter matrix ($\mathbf{S}_{lb}$) as defined in [3]. Our proposed prior considers the joint impact of net Laplacian matrix ($\mathbf{L}_{net}$) and the net $\mathbf{B}$ *i.e.,* $\mathbf{B}_{net}$ onto the shared manifold. The reason behind the use of local between-class scatter matrix in our proposed method is that it can handle the multi-model

characteristics of the data. Mathematically, for v^{th} view, the LBSCM ($\mathbf{S}_{lb}^{v}$) can be represented as follows:

$$\begin{aligned}\mathbf{S}_{lb}^{v} &= \frac{1}{2}\sum_{i,j=1}^{N}\mathbf{W}_{lb,ij}^{v}\left(\mathbf{x}_{i}^{v}-\mathbf{x}_{j}^{v}\right)^{T}\left(\mathbf{x}_{i}^{v}-\mathbf{x}_{j}^{v}\right)\\ &= \mathbf{X}^{v^{T}}\mathbf{B}^{v}\mathbf{X}^{v}\end{aligned} \tag{6.11}$$

where, $\mathbf{B}^{v}=\mathbf{D}_{lb,ii}^{v}-\mathbf{W}_{lb,ij}^{v}$ and $\mathbf{D}_{lb,ii}^{v}=\sum_{j}\mathbf{W}_{lb,ij}^{v}$. The term $\mathbf{W}_{lb,ij}^{v}$ is defined as follows [3,116]:

$$\mathbf{W}_{lb,ij}^{v}=\begin{cases}\left(\frac{1}{N}-\frac{1}{n_{c}^{v}}\right)\exp\left(-\frac{||\mathbf{x}_{i}^{v}-\mathbf{x}_{j}^{v}||^{2}}{\sigma_{i}^{v}\sigma_{j}^{v}}\right) & ;\text{ if } c_{i}=c_{j}\\ \frac{1}{N} & ;\text{ otherwise}\end{cases} \tag{6.12}$$

The parameter n_{c}^{v} is the number of samples belongs to c^{th} class in v^{th} view, and σ_{i}^{v} is the local scaling around $\mathbf{x}_{i}$ in v^{th} view, which is defined as $\sigma_{i}^{v}=||\mathbf{x}_{i}^{v}-\mathbf{x}_{i}^{vk}||_{2}$. The term $\mathbf{x}_{i}^{vk}$ is the k-nearest-neighbor of $\mathbf{x}_{i}^{v}$. We use $k=7$ in our proposed work [159]. Thus, the proposed regularized net local between-class transformation-matrix $\mathbf{B}_{net}$ is defined as:

$$\mathbf{B}_{net}=\mathbf{B}_{nor}^{1}+\mathbf{B}_{nor}^{2}+\cdots+\mathbf{B}_{nor}^{V}+\xi\mathbf{I}=\sum_{v}\mathbf{B}_{nor}^{v}+\xi\mathbf{I} \tag{6.13}$$

where,

$$\mathbf{B}_{nor}^{v}=\left(\mathbf{D}_{lb,ii}^{v}\right)^{-1/2}\mathbf{B}^{v}\left(\mathbf{D}_{lb,ii}^{v}\right)^{-1/2}$$

Finally, the proposed prior for ML-UDSGPLVM is defined as:

$$p\left(\mathbf{Y}\right)=\frac{1}{V.Z_{d}}\exp\left\{-\frac{\beta}{2}tr\left(\frac{\mathbf{Y}^{T}\mathbf{L}_{net}\mathbf{Y}}{\mathbf{Y}^{T}\mathbf{B}_{net}\mathbf{Y}}\right)\right\} \tag{6.14}$$

Hence, the proposed prior is more general and suitable for classification as compared to the earlier priors proposed in [1] and [154]. So, class-separation in the low-dimension manifold is being learned from the class-separability of all the views. Additionally, it can also preserve local structure of the data on the reduced manifold. Incorporating the proposed prior in Eqn. (6.5), the proposed negative log-likelihood of ML-UDSGPLVM is given in Eqn. 6.15.

$$L_s = \sum_{v=1}^{V} L_v + \frac{\beta}{2} tr\left(\frac{\mathbf{Y}^T \mathbf{L}_{net} \mathbf{Y}}{\mathbf{Y}^T \mathbf{B}_{net} \mathbf{Y}}\right) \tag{6.15}$$

where, L_v is defined in Eqn. (6.6). To obtain the optimal latent space, we need to find the derivative of Eqn. (6.15) w.r.t $\mathbf{Y}$, which is given as:

$$\frac{\partial L_s}{\partial \mathbf{Y}} = \sum_{v=1}^{V} \frac{\partial L_v}{\partial \mathbf{Y}} + \frac{\beta}{2} \varphi\left(\mathbf{Y}\right) \tag{6.16}$$

where,

$$\boldsymbol{\varphi}\left(\mathbf{Y}\right) = \left(\frac{2\mathbf{L}_{net}\mathbf{Y}}{\mathbf{Y}^T \mathbf{B}_{net} \mathbf{Y}}\right) - \left(\frac{2\mathbf{B}_{net}\mathbf{Y}}{\mathbf{Y}^T \mathbf{B}_{net} \mathbf{Y}}\right)\left(\frac{\mathbf{Y}^T \mathbf{L}_{net} \mathbf{Y}}{\mathbf{Y}^T \mathbf{B}_{net} \mathbf{Y}}\right) \tag{6.17}$$

As the Gaussian process follows an iterative procedure to find an optimal latent space, we need to evaluate $\boldsymbol{\varphi}\left(\mathbf{Y}\right)$ in each of the iterations, which is computationally expensive. Also, latent states obtained by this approach is fluctuating, and hence convergence rate will be slower than that of LPP-based prior [1]. To overcome these limitations of our proposed method, the proposed prior is slightly modified as follows:

$$p_{\,\mathrm{mod}}\,\left(\mathbf{Y}\right) = \frac{1}{V.Z_d} \exp\left\{-\frac{\beta_1}{2} tr\left(\mathbf{Y}^T \mathbf{L}_{net} \mathbf{Y}\right) + \frac{\beta_2}{2} tr\left(\mathbf{Y}^T \mathbf{B}_{net} \mathbf{Y}\right)\right\} \tag{6.18}$$

The corresponding proposed negative log-likelihood and its derivative w.r.t. latent space $\mathbf{Y}$ can reformulated as follows:

$$L_s^{\,\mathrm{mod}} = \sum_{v=1}^{V} L_v + \frac{\beta_1}{2} tr\left(\mathbf{Y}^T \mathbf{L}_{net} \mathbf{Y}\right) - \frac{\beta_2}{2} tr\left(\mathbf{Y}^T \mathbf{B}_{net} \mathbf{Y}\right) \tag{6.19}$$

$$\frac{\partial\, L_s^{\,\mathrm{mod}}}{\partial \mathbf{Y}} = \sum_{v=1}^{V} \frac{\partial L_v}{\partial \mathbf{Y}} + \left(\beta_1 \mathbf{L}_{net} - \beta_2 \mathbf{B}_{net}\right) \mathbf{Y} \tag{6.20}$$

This representation is simple, and also it allows smooth convergence of the latent space. This is due to the absence of denominator terms, which change the latent space abruptly. Hence, the proposed method is comparatively more suitable than the existing methods in terms of obtaining optimal latent subspace. This directly improves the recognition accuracy.

Moreover, test samples come from high-dimensional subspace need to be mapped onto the lower-dimensional latent manifold during the inference process of GPLVM. For this, back-constrain (learning of inverse mapping) has been defined such that the topology of data space are preserved in the latent manifold [160]. In [1], two kinds of back-constraints are defined for multi-views, namely independent back-projection (I_{bp}) and single back-projection (S_{bp}). For I_{bp}, separate inverse functions are learned for each of the views, whereas for S_{bp}, a single inverse mapping function is learned from all the views to the shared space. They are defined as:

$$\mathbf{Y} = \begin{cases} \mathbf{K}_{ibc}^{v}\mathbf{A}_{ibc}^{v}; \forall v = 1, 2, ..., V & : \text{ for } \mathrm{I}_{bc} \\ \left(\sum_{v=1}^{V} w_v \mathbf{K}_{bc}^{v}\right)\mathbf{A}_{sbc} = \mathbf{K}_{sbc}\mathbf{A}_{sbc} & : \text{ for } \mathrm{S}_{bc} \end{cases} \tag{6.21}$$

where, $(i,j)^{th}$ element of $\mathbf{K}_{ibc}^{v}$ *i.e.,* $k_{bc}^{v}\left(\mathbf{x}_i^v, \mathbf{x}_j^v\right)$ which is given by:

$$k_{bc}^{v}\left(\mathbf{x}_i^v, \mathbf{x}_j^v\right) = \exp\left(-\frac{\gamma^v}{2}||\mathbf{x}_i^v - \mathbf{x}_j^v||^2\right) \tag{6.22}$$

$\mathbf{A}_{ibc}^{v}$ and $\mathbf{A}_{sbc}$ are the regression matrices and w_c is the weight corresponding to the v^{th} view. Finally, these constraints are incorporated in the objective function of Eqn. (6.19), and then the minimization problem takes of the following form:

$$\begin{aligned} &\underset{\mathbf{Y},\theta_v,\mathbf{A}}{\arg\min}\ L_s^{\text{mod}} + R(\mathbf{A}) \\ &\text{s.t.} \begin{cases} \mathbf{Y} - \mathbf{K}_{ibc}^{v}\mathbf{A}_{ibc}^{v} = \mathbf{0},\ v = 1, 2, ..., V \text{ for } \mathrm{I}_{bc} \\ \mathbf{Y} - \mathbf{K}_{sbc}\mathbf{A}_{sbc} = \mathbf{0}, w_v \geqslant 0, \sum_v w_v = 1, \text{for } \mathrm{S}_{bc} \end{cases} \end{aligned} \tag{6.23}$$

where $R(\mathbf{A})$ is a regularization term, which controls the over-fitting of the model to the data. An efficient way of solving this constraint optimization problem is given in [1], where the minimization problem is first divided into a set of subproblems by employing Alternative Direction Method (ADM) [161]. Next, an iterative approach (conjugate gradient algorithm) [162] is applied to solve each of the subproblems separately with respect to their associated model parameters. We follow the same procedure to obtain the optimal latent manifold and other model parameters.

6.3.4 Uncorrelated latent space

In spite of using non-linear based approach to reduce the dimensionality of original feature space to the latent space, there may exist correlations between features. This may further affect the classification accuracy of the FER system. So in our proposed approach, instead of classifying directly from the correlated latent space, we first transformed features of the shared space $\mathbf{Y}$ to another shared space $\mathbf{Y_{uc}}$, where features are uncorrelated. Then classification is performed. We obtained a non-linear uncorrelated discriminative manifold from the non-linear correlated discriminative manifold (original latent manifold) via transformation matrix $\boldsymbol{\chi} = [\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, ..., \boldsymbol{\upsilon}_d]$. The columns of $\boldsymbol{\chi}$ are essentially the solutions (eigenvectors) of the following generalized eigenvalue equation corresponding to the first d lowest eigenvalues [115]:

$$\phi\left(\mathbf{Y}\right)\left(\mathbf{L}_s + \mathbf{B}_s\right)\phi(\mathbf{Y})^T\boldsymbol{\upsilon} = \lambda\phi\left(\mathbf{Y}\right)\mathbf{G}\phi(\mathbf{Y})^T\boldsymbol{\upsilon} \tag{6.24}$$

where, $\phi\left(\mathbf{Y}\right) = \left[\phi\left(\mathbf{y}_1\right), \phi\left(\mathbf{y}_2\right), ..., \phi\left(\mathbf{y}_N\right)\right]$. $\mathbf{L}_s$ and $\mathbf{B}_s$ are the Laplacian and the local-between-class-transformation matrices respectively (similar to Eqn. (6.9) and Eqn. (6.13)) obtained from the shared manifold. The matrix $\mathbf{G} = \mathbf{I} - \left(1/NV\right)\mathbf{e}\mathbf{e}^T$, where $\mathbf{I}$ is an identity matrix and $\mathbf{e} = \left(1, 1, ..., 1\right)^T$. Further, since eigenvectors of Eqn. (6.24) should lie in the span of $\phi\left(\mathbf{y}_1\right), \phi\left(\mathbf{y}_2\right), ..., \phi\left(\mathbf{y}_N\right)$, there exists a vector $\boldsymbol{\alpha}_d$ such that $\boldsymbol{\upsilon}_d = \phi\left(\mathbf{Y}\right)\boldsymbol{\alpha}_d$, where $\boldsymbol{\alpha}_d = \left[\alpha_1^d, \alpha_2^d, ..., \alpha_N^d\right]^T$. Hence, for d^{th} eigenvector, Eqn. (6.24) can be rewritten in terms of $\boldsymbol{\alpha}_d$ as follows:

$$\phi\left(\mathbf{Y}\right)\left(\mathbf{L}_s + \mathbf{B}_s\right)\phi(\mathbf{Y})^T\phi\left(\mathbf{Y}\right)\boldsymbol{\alpha}_d = \lambda\phi\left(\mathbf{Y}\right)\mathbf{G}\phi(\mathbf{Y})^T\phi\left(\mathbf{Y}\right)\boldsymbol{\alpha}_d \tag{6.25}$$

Multiplying both sides of Eqn. (6.25) by $\phi(\mathbf{Y})^T$ and by simple substitution, the following generalized eigenvalue equation is obtained:

$$\mathbf{M}\left(\mathbf{L}_s + \mathbf{B}_s\right)\mathbf{M}\boldsymbol{\alpha}_d = \lambda\mathbf{M}\mathbf{G}\mathbf{M}\boldsymbol{\alpha}_d \tag{6.26}$$

where, $\mathbf{M} = \phi(\mathbf{Y})^T\phi\left(\mathbf{Y}\right)$ is the kernel matrix with $\mathbf{M}_{ij} = \exp\left(-||\mathbf{y}_i - \mathbf{y}_j||/\sigma\right)$. Let $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_d$ be the solutions of Eqn. (6.26), then transformed uncorrelated non-linear manifold can be ob-

tained as follows:

$$\begin{aligned}\mathbf{Y}_{uc} &= [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_d]^T \phi(\mathbf{Y}) \\ &= [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_d]^T \phi(\mathbf{Y})^T \phi(\mathbf{Y}) \\ &= [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_d]^T \mathbf{M}\end{aligned} \tag{6.27}$$

Similarly, for a given new sample $\mathbf{y}^*$ of correlated manifold $\mathbf{Y}$, the corresponding position onto the uncorrelated manifold can be obtained by using the following equation:

$$\mathbf{y}^*_{uc} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_d]^T [\mathbf{M}_{1*}, \mathbf{M}_{2*}, ..., \mathbf{M}_{N*}]^T \tag{6.28}$$

where, $\mathbf{M}_{k*} = \exp(-||\mathbf{y}_k - \mathbf{y}^*||/\sigma)$.

6.4 Experimental Validation

The BU3DFE is a widely used dataset to evaluate the performance of multi-view and/or view-invariant FER methods. This database comprises of 3D facial images of Happy (HA), Surprise (SU), Fear (FE), Anger (AN), Disgust (DI), Sad (SA), and Neutral (NA) expressions. The images of BU3DFE database has 100 subjects, which includes 56% of female and 44% of male candidates. Also, expressions of BU3DFE dataset are captured at four different intensity levels ranging from onset/offset level to peak level of a expression. As the database has 3D images, we first rendered the 3D face models using OpenGL to obtain the 2D textured facial images. 3D face model is first rotated by a user-defined angle, and then the corresponding 2D textured images are obtained. In our proposed approach, we obtained 2D facial images for seven views *i.e.,* $-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, and $45°$ yaw angles. A part of BU3DFE dataset is shown in Figure 5.3, where a single subject is showing the "happy" expression for seven different viewing angles.

The proposed ML-UDSGPLVM algorithm is validated on BU3DFE dataset. In our experiment, images from all the 100 subjects of BU3DFE dataset are employed. Also, expressions from all the intensity levels are considered for our experiment. So, altogether 1800 images/view *i.e.,*

$1800 \times 7 = 12600$ images are considered to evaluate the performance of our proposed method. Each view of the multi-view facial images comprises of six basic expressions *i.e.,* anger, disgust, fear, happy, sad, and surprise. For our experimentation, 300 images are taken for each of the expressions. The face part of 2D textured expressive images are manually cropped, and then down-sampled to get an image size of 160×140. Subsequently, the proposed 54 facial landmark points are localized. Localization of the facial points for the views $-45°$ and $45°$ are carried out manually, whereas images for the views ($-30°$, $-15°$, $0°$, $15°$, and $30°$) are automatically annotated using Active Appearance Model (AAM). Out of 54 landmark points, 5 stable points *i.e.,* left and right corners of the respective eyes, tip of the nose, and corners of the mouth are used to align the facial images using Procrustes analysis [129]. Finally, a grid of 15×15 is considered at each of the facial points to extract a feature vector from salient regions of a face. LBP^{u2} operator is applied to each of the sub-blocks around each of the landmark points to obtain a feature vector. LBP^{u2} gives a 59-dimensional feature vector corresponding to each of the facial sub-regions, and hence the overall feature dimension of an image is $54 \times 59 = 3186$. The first

Table 6.1: View-wise recognition rates (RR) for ML-UDSGPLVM on BU3DFE database

Stage1 model evaluation using LBP + PCA + ML-UDSGPLVM features								
Ist-level of expression classes	Recognition Rate (RR) (in %)							
	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	Avg RR
Lip-based	94.32	96.00	95.80	95.80	95.80	90.88	89.20	93.97
Lip-Eye-based	96.08	92.70	95.00	98.30	91.70	92.53	95.84	94.59
Lip-Eye-Forehead-based	96.25	95.26	96.90	98.05	93.11	93.33	95.26	95.45
Average accuracy = 94.67%								
Stage1 model evaluation using LBP + LPP + ML-UDSGPLVM features								
Ist-level of expression classes	Recognition Rate (RR) (in %)							
	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	Avg RR
Lip-based	98.02	98.30	98.20	99.20	99.01	99.20	97.00	97.30
Lip-Eye-based	97.86	97.05	98.20	99.00	99.70	99.00	98.20	98.43
Lip-Eye-Forehead-based	98.50	98.20	99.00	99.00	98.30	98.30	98.20	98.50
Average accuracy = 98.07%								

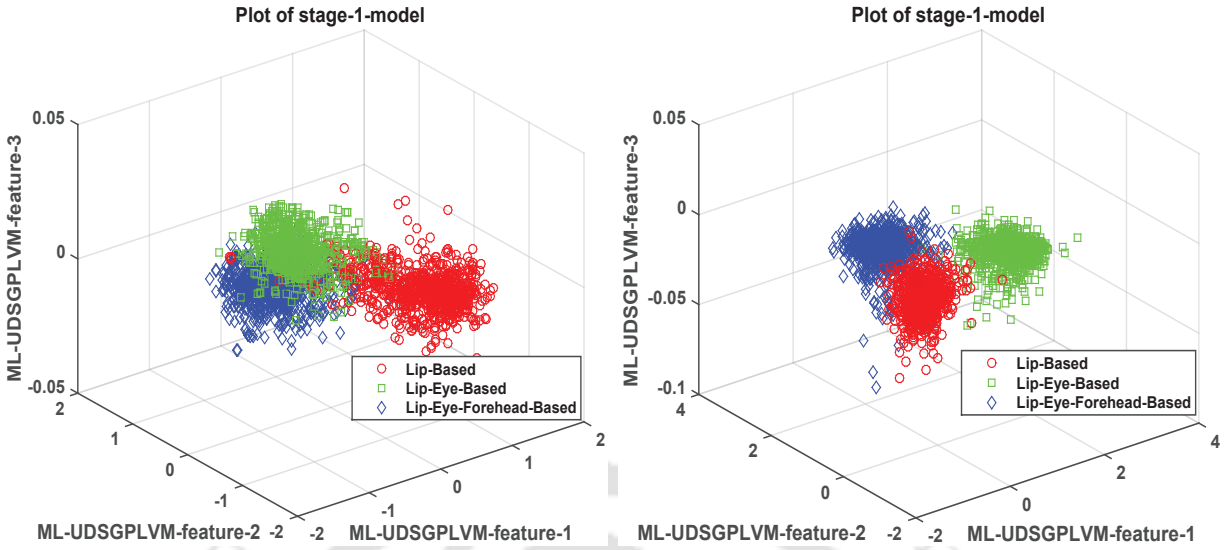
level of dimensionality reduction of LBP-based appearance feature is performed using PCA, in which 95% of total variance of the data is preserved. As the features corresponding to the data

(original feature space) are obtained for different views, so they may form altogether different clusters. Thus, the overall data space may be multi-modal. Hence, LPP-based dimensionality reduction approach would be more suitable in case of multi-view facial expression recognition. LPP-based dimensionality reduction technique is more capable in handling multi-modal data. In our proposed method, LPP-based approach is utilized to extract a set of discriminative features.

Table 6.2: View-wise expressions recognition rates (RR) for ML-UDSGPLVM on BU3DFE database

Stage2 model evaluation using LBP + PCA + ML-UDSGPLVM features									
Model	Expressions	Recognition Rate (RR) (in %)							
		−45°	−30°	−15°	0°	15°	30°	45°	Avg RR
Stage2-model1	Happy	83.40	93.30	96.70	96.70	99.60	96.70	99.00	95.05
	Sad	93.30	73.40	76.70	96.70	80.00	90.00	83.40	84.78
Stage2-model2	Surprise	99.10	90.00	96.70	98.00	93.30	99.80	93.40	95.75
	Disgust	93.30	93.40	96.70	96.70	83.30	99.20	93.30	93.70
Stage2-model3	Anger	83.40	80.00	96.40	93.30	76.70	96.70	86.70	87.60
	Fear	80.00	73.30	76.70	96.70	83.30	90.00	90.00	84.28
Average accuracy = 90.20%									
Stage2 model evaluation using LBP + LPP + ML-UDSGPLVM features									
Model	Expressions	Recognition Rate (RR) (in %)							
		−45°	−30°	−15°	0°	15°	30°	45°	Avg RR
Stage2-model1	Happy	99.80	99.50	99.50	99.80	83.40	99.60	99.20	97.25
	Sad	79.20	98.80	91.60	94.40	93.30	94.80	96.00	92.58
Stage2-model2	Surprise	95.60	94.80	94.40	97.60	99.70	97.60	99.60	97.04
	Disgust	99.60	98.80	98.40	98.40	93.30	99.20	96.00	97.67
Stage2-model3	Anger	95.60	96.00	98.40	96.40	83.40	96.40	99.60	95.11
	Fear	90.00	96.80	98.00	99.20	80.00	94.00	95.60	93.37
Average accuracy = 95.51%									

The experiments are carried out using 10-fold cross validation strategy, and hence we first divide images of each of the views into 10 subsets. Out of which, 9 subsets are used to train the model, whereas the remaining set is used for testing. The experiments are repeated for 10 times such that testing subset is selected exactly ones in each of the iterations. Then, average accuracy is obtained for all the experiments. In all the experiments, we use 1-nearest neighbor (1-NN) classifier to evaluate the performance of the proposed method.



(a) (b)

Figure 6.4: 3D distribution of test samples of three expressions: (a) trained by LBP, PCA, and ML-UDSGPLVM features(b) trained by LBP, LPP, and ML-UDSGPLVM features.

We used the same parameter settings as used in [1]. The parameters γ^v (back-projection parameter) are learned through Leave-One-Out cross validation procedure. Finally, the optimum values of the two parameters *i.e.,* β (in our case β_1) and d (dimension of latent space) are found as $\beta = 300$ and $d = 5$. So, we used these parameter values to get optimal performance of our proposed ML-UDSGPLVM. The only extra parameter which is used in our proposed algorithm is the weight of the prior β_2, which controls the inter-class variance of the data onto the shared manifold. This parameter is learned experimentally by varying β_2 from 10 to 0.01, and found to be optimal at $\beta_2 = 0.8$.

The proposed ML-UDSGPLVM approach is a multi-level framework, where first level of proposed model *i.e.,* 1-UDSGPLVM is first trained by three sets of expression categories *i.e.,* Lips-based = {happy, sad}, Lips-Eyes-based = {surprise, disgust}, and Lips-Eyes-Forehead-based = {anger, fear}.

Table 6.3: View-wise confusion matrices for six basic expressions

Confusion matrix obtained using LBP + PCA + ML-UDSGPLVM features							
View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of $-45°$	Anger	83.4	0	0	0	13.3	3.3
	Disgust	6.7	93.3	0	0	0	0
	Fear	3.3	0	80	0	10	6.7
	Happy	0	0	0	83.4	13.3	3.3
	Sad	0	0	6.7	0	93.3	0
	Surprise	0	0	0	0.9	0	99.1
Pan of $-30°$	Anger	80.0	0	0	0	6.7	13.3
	Disgust	3.3	93.4	0	0	3.3	0
	Fear	6.7	0	73.3	0	20	0
	Happy	0	0	0	93.3	6.7	0
	Sad	3.3	0	20	0	73.4	3.3
	Surprise	0	0	0	0	10	90
Pan of $-15°$	Anger	96.4	0	3.0	0	0	0.6
	Disgust	0	96.7	0	0	3.3	0
	Fear	0	0	76.7	0	23.3	0
	Happy	0	0	0	96.7	0	3.3
	Sad	0	0	23.3	0	76.7	0
	Surprise	0	0	0	0	3.3	96.7
Pan of $0°$	Anger	93.3	0	0	0	0	6.7
	Disgust	0	96.7	0	0	0	3.3
	Fear	0	0	96.7	0	3.3	0
	Happy	0	0	0	96.7	3.3	0
	Sad	0	0	3.3	0	96.7	0
	Surprise	0	0	0	2	0	98
Pan of $15°$	Anger	76.7	0	0	0	10	13.3
	Disgust	3.3	83.3	0	0	13.3	0
	Fear	6.7	0	83.3	0	0	10
	Happy	0	0	0	99.6	0	0.4
	Sad	10	0	10	0	80	0
	Surprise	0	0	0	0	6.7	93.3
Pan of $30°$	Anger	96.7	0	0	0	3.3	0
	Disgust	0	99.2	0	0	0	0.8
	Fear	0	0	90	0	10	0
	Happy	0	0	0	96.7	3.3	0
	Sad	0	0	10	0	90	0
	Surprise	0	0	0	0.2	0	99.8
Pan of $45°$	Anger	86.7	0	0	0	13.3	0
	Disgust	0	93.3	0	0	0	6.7
	Fear	0	0	90	0	10	0
	Happy	0	0	0	99	0	1
	Sad	3.3	0	13.3	0	83.4	0
	Surprise	0	3.3	0	0	3.3	93.4

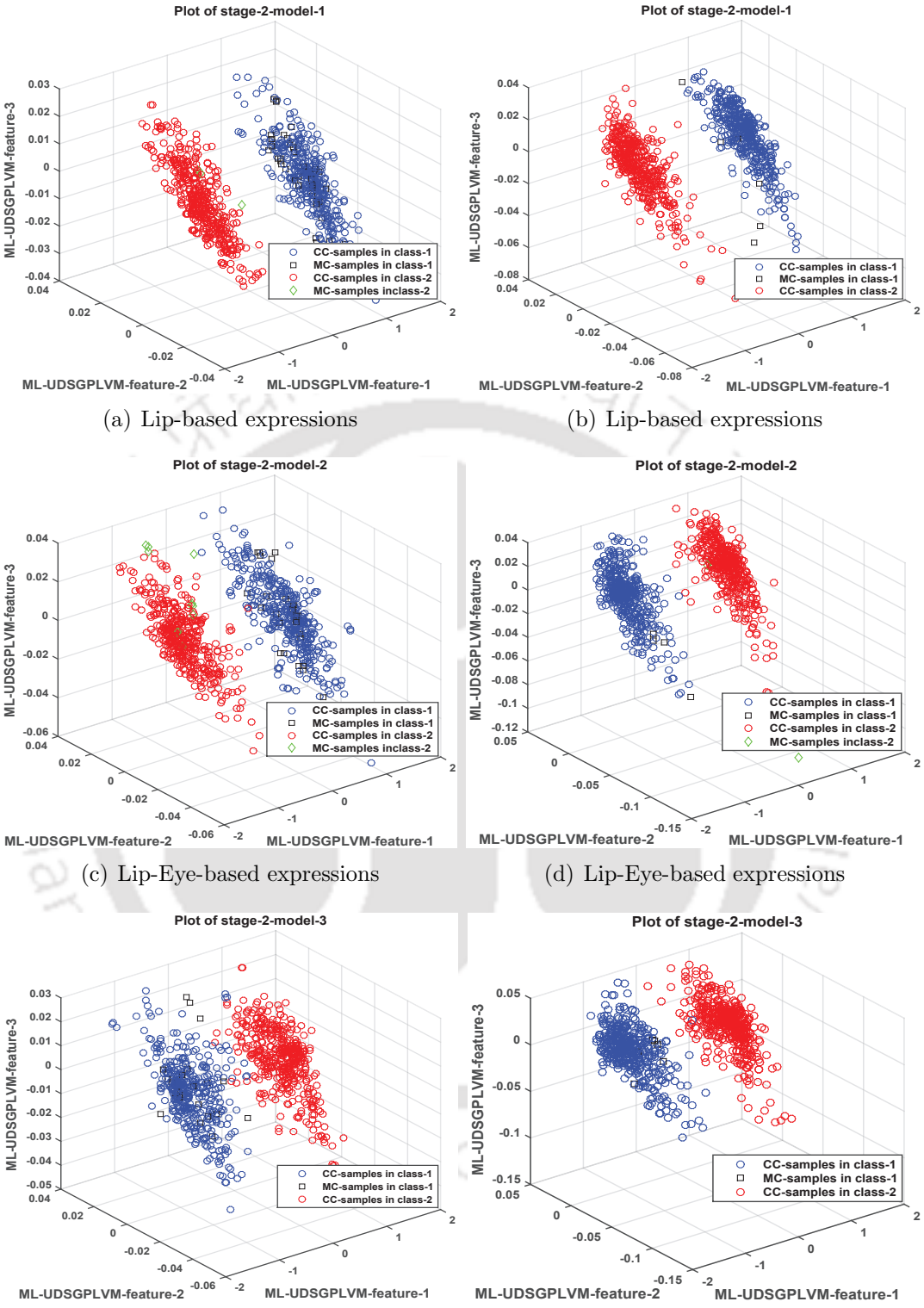
Table 6.4: View-wise confusion matrices for six basic expressions

Confusion matrix obtained using LBP + LPP + ML-UDSGPLVM features							
View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of −45°	Anger	95.6	0	0	4.0	0	0.4
	Disgust	0	99.6	0	0	0.4	0
	Fear	0.4	0	90.0	0	9.6	0
	Happy	0	0	0	99.8	0	0.2
	Sad	0	0	20.4	0	79.2	0.4
	Surprise	0	4.4	0	0	0	95.6
Pan of −30°	Anger	96.0	0	0	1.2	2.4	0.4
	Disgust	0.8	98.8	0	0	0	0.4
	Fear	0.8	0	96.8	0	1.2	1.2
	Happy	0	0	0	99.5	0	0.5
	Sad	0	0	1.2	0	98.8	0
	Surprise	0	4.4	0	0	0.8	94.8
Pan of −15°	Anger	98.4	0	0	0.4	0.4	0.8
	Disgust	0.4	98.4	0	0	1.2	0
	Fear	0.8	0	98.0	0	1.2	0
	Happy	0	0	0	99.5	0	0.5
	Sad	0.4	0	8.0	0	91.6	0
	Surprise	0	5.6	0	0	0	94.4
Pan of 0°	Anger	96.4	0	0	1.2	0.8	1.6
	Disgust	0.4	98.4	0	0	1.2	0
	Fear	0	0	99.2	0	0.8	0
	Happy	0	0	0	99.8	0.2	0
	Sad	0	0	5.2	0	94.4	0.4
	Surprise	0.4	2.0	0	0	0	97.6
Pan of 15°	Anger	83.4	0	0	0	13.3	3.3
	Disgust	6.7	93.3	0	0	0	0
	Fear	3.3	0	80	0	10	6.7
	Happy	0	0	0	83.4	13.3	3.3
	Sad	0	0	6.7	0	93.3	0
	Surprise	0	0	0.3	0	0	99.7
Pan of 30°	Anger	96.4	0	0	0.8	1.2	1.6
	Disgust	0.4	99.2	0	0	0.4	0
	Fear	1.2	0	94.0	0	4.0	0.8
	Happy	0	0	0	99.6	0.4	0
	Sad	0.4	0	4.4	0	94.8	0.4
	Surprise	0	1.6	0	0	0.8	97.6
Pan of 45°	Anger	99.6	0	0	0	0.4	0
	Disgust	0.4	96.0	0	0	0.4	3.2
	Fear	0.4	0	95.6	0	4.0	0
	Happy	0	0	0	99.2	0.4	0.4
	Sad	0	0	4.0	0	96.0	0
	Surprise	0	0.4	0	0	0	99.6

Subsequently, a second level of ML-UDSGPLVM *i.e.,* 2-UDSGPLVM is trained for the expressions, and hence, three 2-UDSGPLVMs are trained in the second level of proposed ML-UDSGPLVM. Two different approaches *i.e.,* PCA and LPP are applied to reduce the dimensionality of LBP features. For this, both PCA and LPP are applied on 90% of the samples (*i.e.*, 10-fold cross-validation strategy) of each of the views to obtain the principal directions, and subsequently those direction vectors are used to project both training and testing samples to the initial reduced subspace. In PCA, we reduce feature dimension in such a way that 95% variance of the data can be captured. In case of LPP, we restrict the feature set to be a 100-dimensional subspace. Finally, we apply our proposed ML-UDSGPLVM onto the reduced feature set to obtain a sufficiently lower dimensional non-linear discriminative subspace.

Furthermore, features in the discriminative latent space may be correlated, and hence, we performed another transformation on features of the correlated latent space. The first three components of ML-UDSGPLVM features are applied on two sets of features *i.e.,* "LBP + PCA + ML-UDSGPLVM" and "LBP + LPP + ML-UDSGPLVM". The distribution of the test samples of all the views for these two cases are shown in Figure 6.4(a) and Figure 6.4(b), respectively. These distribution plots show that first level of proposed "ML-UDSGPLVM + LBP + LPP" provides better separability than the combination of "ML-UDSGPLVM + LBP + PCA" [1]. The view-wise average recognition rates for all the three types of expressions are shown in Table 6.1. From Table 6.1, it is clear that proposed LBP + LPP followed by ML-UDSGPLVM gives an improvement of about 4% as compared to "LBP + PCA + ML-UDSGPLVM"-based approach [1].

As discussed earlier, three class problem is considered at the first stage of ML-UDSGPLVM. In the second stage, we need three 2-UDSGPLVM – one for each expression. Each 2-UDSGPLVM is trained using the same training samples of the respective expression class. For example, 2-UDSGPLVM corresponding to Lip-based expressions are trained using the samples of the respective sub-classes *i.e.,* "happy" and "sad". Furthermore, the samples which were used for testing of 1-UDSGPLVM are again used for 2-UDSGPLVM.



(a) Lip-based expressions (b) Lip-based expressions

(c) Lip-Eye-based expressions (d) Lip-Eye-based expressions

(e) Lip-Eye-Forehead-based expressions (f) Lip-Eye-Forehead-based expressions

Figure 6.5: 3D distribution of test samples of the basic expressions. First these figures show the plot of test samples when 2-UDSGPLVM is applied on LBP followed by PCA, and the second column shows the distribution when 2-UDSGPLVM is applied on LBP followed by LPP-based features, respectively. CC and MC stand for correctly classified and missclassified test samples.

The samples which were misclassified in the first stage of ML-UDSGPLVM are tested by the respective 2-UDSGPLVM in the second level of ML-UDSGPLVM. So, misclassified samples of 1-UDSGPLVM (stage-1) and 2-UDSGPLVM (stage-2) are accounted for finding the overall misclassified samples. The misclassified samples are shown in Figure 6.5. The overall view-wise classification accuracies for different basic expressions are shown in Table 6.2, and the corresponding distributions of test samples for two sets of features are shown in Figure 6.5. It is even perceptually clear from the distribution plots that second level of ML-UDSGPLV and LBP + LPP provide better separability than that of LBP + PCA based features, and the overall improvement is about 5%. Tables 6.3 shows the recognition accuracies for different views *i.e.*,($-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$,and $45°$) for the above mentioned two feature sets. Table 6.5 shows the comparison of DS-GPLVM [1] and our proposed ML-UDSGPLVM. The performance of DS-GPLVM is evaluated by imposing it with LDA-based prior, LPP-based prior, and the prior proposed in Eqn. (6.18). It is observed that the performance of DS-GPLVM with the proposed prior is better than LPP-based prior, and the improvement is even more significant ($> 5\%$) than LDA-based prior. Our proposed ML-UDSGPLVM gives an overall average accuracy of 95.51%, which is about 3% better than the original DS-GPLVM (DS-GPLVM with LPP-based prior). This significant improvement is due to the use of multi-lavel

Table 6.5: Comparison of proposed method with the state-of-the-art DS-GPLVM-based methods on BU3DFE dataset in terms of average recognition rates with average standard deviation.

Comparison using LBP + PCA + Shared features								
Methods	Recognition Rate (RR) (in %)							
	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	Avg RR
DS-GPLVM with LDA-based prior	81.04	76.20	74.79	83.87	81.04	79.23	78.83	79.29 ± 0.027
DS-GPLVM with LPP-based prior	90.92	85.88	85.68	93.95	81.04	87.50	77.01	86.00 ± 0.021
DS-GPLVM with proposed prior	90.32	83.87	84.07	95.16	85.08	85.28	83.46	86.75 ± 0.021
ML-UDSGPLVM with proposed prior	**88.75**	**83.90**	**89.98**	**96.35**	**86.03**	**95.40**	**90.96**	**90.19 ± 0.011**
Comparison using LBP + LPP + Shared features								
Methods	Recognition Rate (RR) (in %)							
	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	Avg RR
DS-GPLVM with LDA-based prior	96.97	91.53	94.95	91.73	93.75	87.90	84.87	91.67 ± 0.025
DS-GPLVM with LPP-based prior	96.37	90.12	91.33	97.37	92.74	91.53	88.50	92.56 ± 0.015
DS-GPLVM with proposed prior	95.86	92.13	94.15	96.87	95.86	91.73	89.61	93.75 ± 0.015
ML-UDSGPLVM with proposed prior	**93.30**	**97.45**	**96.71**	**97.63**	**88.85**	**96.93**	**97.66**	**95.51 ± 0.014**

Table 6.6: Comparison of proposed method with the state-of-the-arts methods on BU3DFE dataset.

State-of-the-art-methods									Proposed method
GMPCA	GMLDA	GMLPP	GMCCA	PW-CCA	MCCA	MvDA	D-GPLVM	DS-GPLVM	ML-UDSGPLVM
89.64	91.19	92.03	91.91	84.28	89.32	93.48	88.33	92.56	**95.51**

framework of uncorrelated DS-GPLVM. The proposed ML-UDSGPLVM on LBP + LPP-based feature gives better performance as compared to DS-GPLVM.

Table 6.9 shows the comparison of several state-of-the-art multi-view learning-based methods [109, 111, 148, 150] with the proposed ML-UDSGPLVM. In this, performance of MvDA is better than DS-GPLVM with LPP-based prior, and it is very close to DS-GPLVM with our proposed prior. However, common spaces in all the multi-view-based linear approaches [109,111,148,150] were obtained by taking 98% of total variance, which corresponds to 175 eigenvectors. This is relatively very high dimensional common space than non-linear DS-GPLVM latent space. Finally, our proposed ML-UDSGPLVM-based approach gives an overall improvement of about 2% than MvDA-based approach. In summary, the proposed ML-UDSGPLVM-based approach can efficiently find the low-dimensional discriminative shared manifold for multi-view FER.

6.5 Hierarchical-UMvDLPP vs ML-UDSGPLVM

Hierarchical-UMvDLPP (H-UMvDLPP) is a multi-level extension of our previous work proposed in Chapter 5. The methodology of H-UMvDLPP is same as that of ML-UDSGPLVM (discussed in Section 6.2). The architecture of H-UMvDLPP is shown in Figure 6.6 [7]. In H-UMvDLPP, expressions are also initially divided into three categories on the basis of the movements of lips, eyes, and forehead as stated in [42, 157]. The corresponding common space is learned using 1-UMvDLPP (Category-wise uncorrelated common space) as shown in Figure 6.6. Further, a 2-UMvDLPP is learned for the constituent expressions present in each of the

[7] This work is in under review in *IET Biometrics* (**To be communicated**)

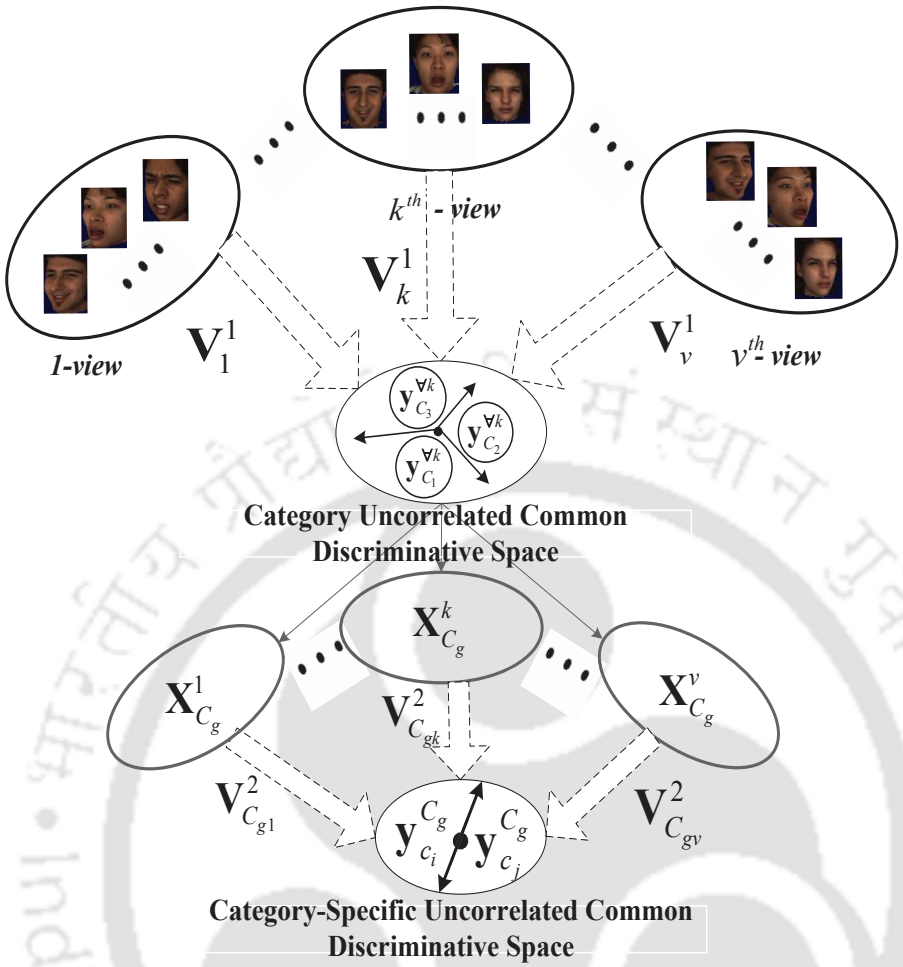


Figure 6.6: Representation of proposed H-UMvDLPP.

subcategories (Category-Specific Uncorrelated Common Space). The basic experimental setup for H-UMvDLPP is same as that of UMvDLPP, however analysis of parameters like finding optimal uncorrelated space, accuracy and/or confusion matrices for different category-expressions and/or basic expressions are presented in the following paragraphs.

Figure 6.7(a) shows the accuracies obtained at two different levels of H-UMvDLPP as a function of dimensionality of the common space. Horizontal axis of Figure 6.7 shows number of dimensions of the common space. The differences between the accuracies obtained at the first level of H-UMvDLPP and the final level of H-UMvDLPP is shown in Figure 6.7(b). The slight reduction in accuracies in the final/second level of the proposed method is due to miss-

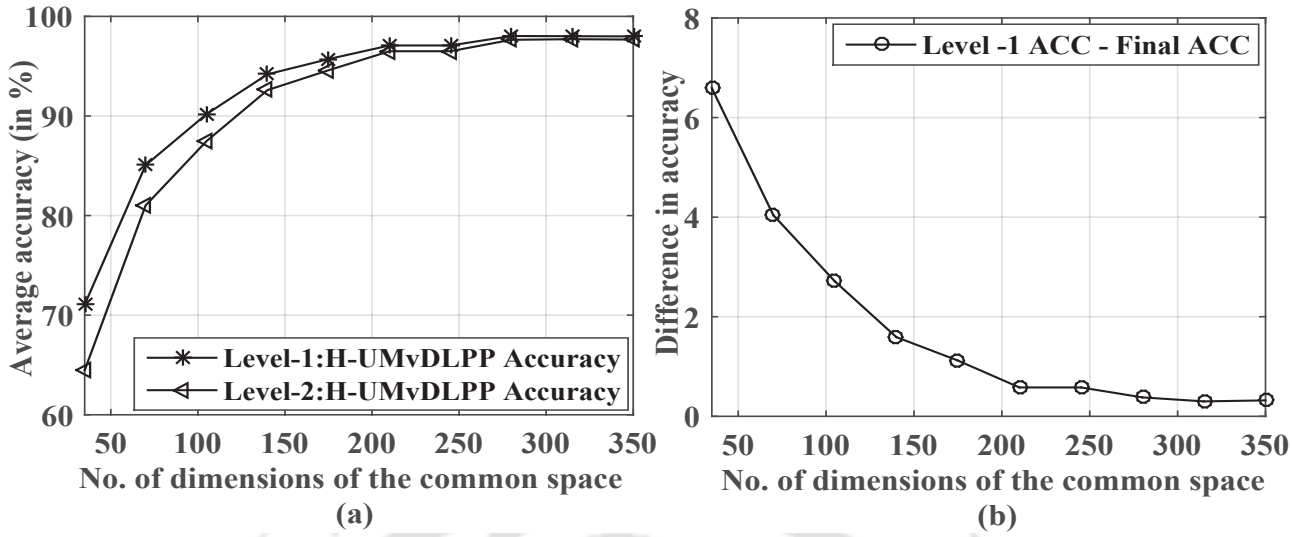


Figure 6.7: Average accuracy of H-UMvDLPP across all the seven poses of BU3DFE dataset vs dimensionality of common space.

classifications occurs in the final stage. In the final stage, samples which are correctly classified in the first stage are only considered for classification. The overall misclassified samples include misclassified samples of both the stages. We experimentally found that 315 eigenvectors captures almost 98% of total variance of the training data. The proposed H-UMvDLPP model gives a recognition rate of 97.66% for all the seven poses, which is better than the existing multi-view-based approaches [1, 111].

In our second experiment, we focused on analyzing inter-class separations between the samples of different classes at different stages of the proposed hierarchical approach. In this process, samples of different classes of the common space are projected onto a plane by eigenvectors of LDA. In the first stage of H-UMvDLPP, the six class problem is converted into three class problem (category-wise classification) *i.e.,* Lips-based, Lips-Eyes-based, and Lips-Eyes-Forehead-based. Subsequently, category-specific UMvDLPP is learned in the second phase of H-UMvDLPP. Figure 6.8 shows the view-wise inter-class separations of test samples in the first level of H-UMvDLPP *i.e.*, 1-UMvDLPP. It is observed that the samples of different categories are quite well separable even in a two dimensional space, and hence even better separation is achievable in a 315-dimensional common space. The distributions of test samples of category-specific expressions (Lips-based, Lips-Eyes-based, and Lips-Eyes-Forehead-based) are shown

Table 6.7: Confusion matrices for category expressions obtained by using 1-UMvDLPP on BU3DFE database. The category expressions are Lips-based (LB), Lips-Eyes-based (LEB), and Lips-Eyes-Forehead-based (LEFB)

Stage1 model evaluation using LBP + PCA + 1-UMvDLPP features (in %)						
Category	Pan of $-45°$			Pan of $-30°$		
Classes	LB	LEB	LEFB	LB	LEB	LEFB
Lips-based	99.03	0.65	0.32	98.81	0.97	0.22
Lips-Eyes-based	1.19	98.27	0.54	0.97	98.92	0.11
Lips-Eyes-Forehead-based	1.08	0.65	98.27	1.08	0.22	98.70
	Average accuracy = 98.52			**Average accuracy** = 98.81		
	Pan of $-15°$			Pan of $0°$		
Lips-based	99.03	0.76	0.22	98.16	0.76	1.08
Lips-Eyes-based	0.54	99.46	0	0.32	99.35	0.32
Lips-Eyes-Forehead-based	0.22	0.76	99.03	0.54	0	99.46
	Average accuracy = 99.17			**Average accuracy** = 98.99		
	Pan of $15°$			Pan of $30°$		
Lips-based	97.62	1.62	0.76	97.84	1.08	1.08
Lips-Eyes-based	0.65	98.70	0.65	1.95	96.86	1.19
Lips-Eyes-Forehead-based	0.87	1.73	97.40	2.49	1.19	96.32
	Average accuracy = 97.91			**Average accuracy** = 97.01		
	Pan of $45°$			Average of all the views		
Lips-based	94.81	2.16	3.03	97.90	1.14	0.96
Lips-Eyes-based	2.60	96.54	0.87	1.18	98.30	0.53
Lips-Eyes-Forehead-based	2.92	0.65	96.43	1.31	0.74	97.94
	Average accuracy = 95.92			**Average accuracy** = 98.05		

in Figure 6.9. In this, we projected expressions of all the views to their respective learned category-specific models to infer the overall inter-class separation. Lips-based and Lips-Eyes-based expressions show better separation than the Lips-Eyes-Forehead-based (*i.e.*, anger and fear) expressions. Moreover, inter-view compactness of each of the expressions is quite good, which shows the effectiveness of introducing intra-view and inter-view LPP-based framework in our proposed method. The horizontal axis (feat-1) and vertical axis (feat-2) of Figures 6.8 and 6.9 represent first and second principal components obtained by LDA projection vectors. In our approach, the samples which are correctly classified in the first stage are also correctly classified using their corresponding category-specific models with an accuracy of more than 99%. Hence, the proposed model effectively reduces six class problem into a three class classification problem.

Classification accuracies for different expression categories are shown in Table 6.7 for different views. This analysis clearly shows that first level of H-UMvDLPP can give an overall

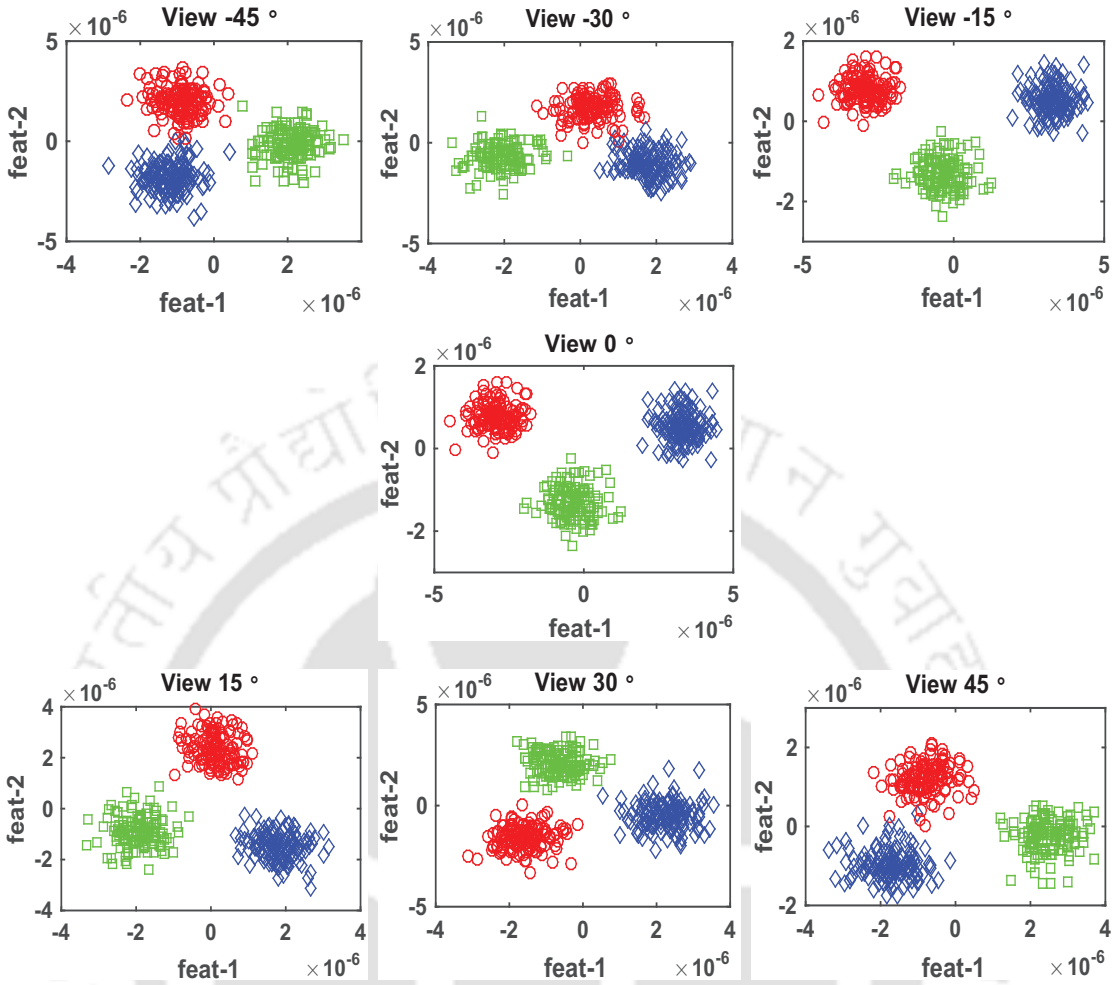


Figure 6.8: Category-wise separations for different views of facial expressions obtained at the first level of H-UMvDLPP. Here, the abbreviations "feat-1" and "feat-2" represent LDA feature 1 and LDA feature 2 respectively.

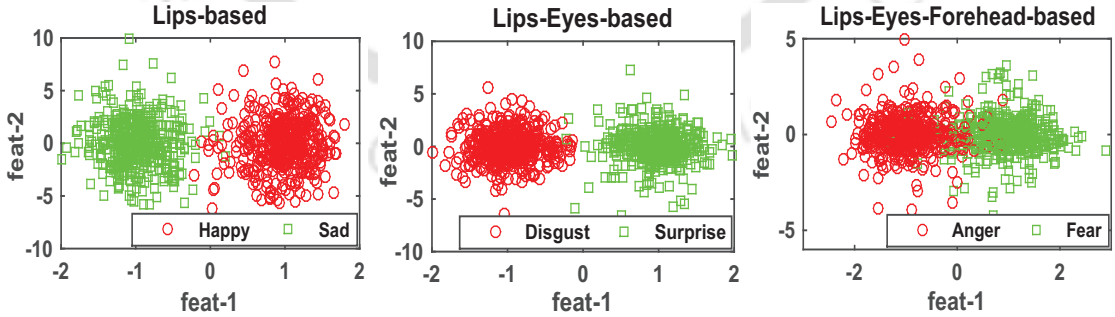


Figure 6.9: Overall compactness/separation of/between the samples of different expressions for all the views in the second level of H-UMvDLPP. Here, the abbreviations "feat-1" and "feat-2" represent LDA feature 1 and LDA feature 2 respectively.

average accuracy of 98.05%. In the second level, we have three category specific learned models, which needs to be selected for final classification. The selection of category-specific models is based on the class labels obtained at the first level, and hence the proposed method is automatic. The view-wise confusion matrices for six basic expressions are shown in Table 6.8. The overall recognition rates of UMvDLPP and H-UMvDLPP are found to be 95.67% and 97.66% respectively. The performance of UMvDLPP closely matches with the state-of-the-art methods, however H-UMvDLPP outperforms the existing state-of-the-art multi-view FER methods. Also, it can be observed from Table 6.8 that H-UMvDLPP gives comparatively higher recognition rates for negative pan angles (left side of a face) than the positive pan angles (right side of a face). These observations also support the findings presented in [151] – left part of a face is more informative for negative emotions such as "disgust", whereas right part of a face is more informative for positive emotions such as "happy" expression. In Table 6.9, we compared our proposed method with the state-of-the-art non-linear-based methods *i.e.*, D-GPLVM [154] and DS-GPLVM [1]. H-UMvDLPP is also compared with our proposed multi-level UDSGPLVM (ML-UDSGPLVM)-based approach. In summary, our proposed H-UMvDLPP-based method outperforms the state-of-the-art linear and non-linear multi-view facial expression recognition methods.

6.6 Conclusion

In this paper, a multi-level framework of uncorrelated discriminative shared Gaussian process latent variable model ML-UDSGPLVM is proposed to obtain a single non-linear uncorrelated discriminative shared manifold. More specifically, we proposed an efficient prior with the help of Laplacian matrix and the local-between-class-scatter-matrix. The reason behind the use of between-class-separability matrix is that it can handle the multi-modal characteristics of multi-view data similar to a Laplacian matrix. In our proposed ML-UDSGPLVM, instead of classifying a test sample directly on correlated shared space, we transformed it to a non-linear uncorrelated latent space, and then 1-NN classifier is used.

Table 6.8: Confusion matrices for proposed H-UMvDLPP method for different views

Confusion matrix obtained using LBP + PCA + H-UMvDLPP features							
View		Anger	Disgust	Fear	Happy	Sad	Surprise
Pan of −45°	Anger	97.84	0.43	1.73	0	0	0
	Disgust	0	98.27	0.87	0.22	0.65	0
	Fear	0	0.43	98.48	0	0.65	0.43
	Happy	0	0.22	0.87	98.92	0	0
	Sad	0	0.87	1.30	0.22	97.40	0.22
	Surprise	0.22	0	1.30	0	0.22	98.27
Pan of −30°	Anger	97.40	1.30	0.65	0	0.22	0.43
	Disgust	0.43	98.48	0	0.22	0	0.87
	Fear	0	0	99.57	0	0.22	0.22
	Happy	0	0	1.08	98.48	0	0.43
	Sad	0	0	1.08	0	98.92	0
	Surprise	0.65	0.22	0.87	0	0	98.27
Pan of −15°	Anger	98.27	0.65	0.65	0.43	0	0
	Disgust	0.22	99.13	0.65	0	0	0
	Fear	0.43	0.87	98.70	0	0	0
	Happy	0.22	0.65	0	98.92	0	0.22
	Sad	0.22	0	0	0	99.13	0.65
	Surprise	0.22	0	0	0	0	99.78
Pan of 0°	Anger	96.97	0.22	0.43	0.43	1.52	0.43
	Disgust	0.43	96.54	0.22	0	0.65	2.16
	Fear	0	0	98.92	0	0.22	0.87
	Happy	0	0	0	99.35	0.65	0
	Sad	0.65	0	0.43	0	98.92	0
	Surprise	0	0	0.12	0	0	99.88
Pan of 15°	Anger	96.10	1.52	1.08	0.22	0.65	0.43
	Disgust	0	99.13	0.65	0	0.22	0
	Fear	0.87	0.65	97.19	0	0.65	0.65
	Happy	0	1.95	0.22	97.62	0	0.22
	Sad	1.30	0.87	0.22	0	97.19	0.43
	Surprise	0	0.65	0.65	0	1.08	97.62
Pan of 30°	Anger	96.75	1.08	0.65	0.22	1.30	0
	Disgust	1.30	96.10	0.65	0.22	1.30	0.43
	Fear	1.73	0.43	96.54	0.22	0.43	0.65
	Happy	0.87	0.87	0.43	97.84	0	0
	Sad	2.16	0.87	1.52	0	94.81	0.65
	Surprise	0	0.43	1.95	0.43	0.43	96.75
Pan of 45°	Anger	93.07	2.60	0	0.22	3.46	0.65
	Disgust	1.30	94.16	1.95	0	1.30	1.30
	Fear	0.43	0.43	96.10	0.22	2.16	0.65
	Happy	0.22	0	0.87	98.92	0	0
	Sad	1.73	0.87	3.03	0	93.94	0.43
	Surprise	0.43	0.65	1.52	0	0.43	96.97

Table 6.9: Comparison with non-linear methods

D-GPLVM [154]	DS-GPLVM [1]	ML-UDSGPLVM	H-UMvDLPP
88.3%	92.56%	95.16%	**97.66**%

Also, the proposed approach adopts multi-level framework – the expressions are first divided into three basic categories *i.e.,* expressions by only Lips, expressions by Lips-Eyes, and expressions by Lips-Eyes-Forehead, which are recognized by first level of ML-UDSGPLVM (1-UDSGPLVM). Subsequently, a separate second level of ML-UDSGPLVM (2-UDSGPLVM) is learned for each of the sub-classes. So, three 2-UDSGPLVMs have to be learned to reach the final classification stage. Expressions are first classified on 1-UDSGPLVM manifold, and the corresponding 2-UDSGPLVM manifold is used for final level of classification. This multi-level decision strategy inherently improves the recognition accuracy. The performance of our proposed ML-UDSGPLVM is evaluated for six basic expressions obtained from seven different poses ($-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, and $45°$) of BU3DFE dataset. ML-UDSGPLVM approach gives an average recognition rate of 95.51% with LBP + LPP-based features. We also presented experimental analysis for extension of our previous work *i.e*, UMvDLPP (Chapter 5), and the underline method is termed as hierarchical-UMvDLPP (H-UMvDLPP). It is clear from the analysis that for the first level of classification (category-level classification), average recognition rates obtained by H-UMvDLPP and ML-UDSGPLVM are almost comparable. However, the overall recognition rate obtained by H-UMvDLPP is better than that of ML-UDSGPLVM due to slightly higher misclassification in the second stage of ML-UDSGPLVM. On the other hand, testing process of ML-UDSGPLVM is faster than that of H-UMvDLPP. This is due to the fact that ML-UDSGPLVM used a very low dimensional space ($d = 5$) as compared to the H-UMvDLPP which uses a common space of dimension 315 for classification. Furthermore, ML-UDSGPLVM uses samples of only one view learned by the samples of all the views for classification, whereas samples of all the views are projected onto the common space learned by the H-UMvDLPP. In summary, our proposed schemes (ML-UDSGPLVM and H-UMvDLPP) outperforms the state-of-the-art linear and non-linear-based multi-view learning techniques.

7

Conclusions and Future Work

There is a humbling amount of past work on facial expression recognition. This dissertation described our proposed facial expression recognition algorithms for both frontal and multi-view face images. In our method, informative facial regions are identified, and subsequently features are extracted from these regions. Also, a common discriminative shared space is derived for recognizing multi-view facial expressions in a hierarchical recognition framework. This chapter reflects on these contributions, discusses future work for facial expression recognition, and concludes. We hope that the future facial expression recognition platform developers find our contributions useful, and benefit from our informative region-based FER without having to re-invent their own.

7.1 Summary

In the beginning of the thesis, challenges faced by computer vision community for recognizing facial expressions automatically are mentioned. These challenges include identification of most informative regions of a face, generation of an efficient face model for extracting both geometrical and texture discriminative features, derivation of a common discriminative shared space for multi-view FER, computational complexity, inadequacy for uncontrolled environment *i.e.*, low resolution images, recognition of subtle and micro expressions etc. In this research work, we proposed different frameworks for facial expression recognition from both frontal and non-frontal multi-view face images to overcome some of the above mentioned challenges. The main goal of this dissertation is to make facial expression recognition techniques work better for different head-poses.

In summary, this thesis addresses the following issues:

(i) Extraction of informative regions of a face for extracting discriminative facial features.

(ii) Development of an efficient face model with the help of informative regions of a face.

(iii) Learning of an optimal discriminative common/shared space for multi-view facial expression recognition.

(iv) Handling multi-modal characteristics of multi-view facial data.

(v) Uncorrelating samples in the common/shared latent space.

(vi) Incorporating hierarchical framework for multi-view facial expression recognition.

The main contributions of this dissertation are as follows:

- A projection-based approach is proposed to analyze importance of different facial regions. Thereby, informative/salient regions of a face are selected on the basis of projection errors.
- A weighted local binary pattern-based feature is proposed, where features extracted only from the informative regions of a face are weighted on the basis of their respective importances.

- A more versatile face model is proposed to simultaneously extract geometrical and texture features.
- Uncorrelated multi-view discriminant locality preserving projection (UMvDLPP)-based approach is proposed to recognize expressions from a set of multi-view face images.
- Multi-level uncorrelated discriminative shared Gaussian process latent variable model (ML-UDSGPLVM) is proposed for recognizing multi-view expressions in a hierarchical framework.

The summary of all the chapters of this dissertation is highlighted as follows:

(i) In introduction chapter (**Chapter 1**), a typical facial expression recognition framework is illustrated. This chapter also highlights several major applications of facial expression recognition. Also, inter-connections between human emotions and facial gestures are explored. Further, two face parametrization approaches used for analyzing facial movements are discussed. In the first approach, expressions are identified on the basis of movements of different facial regions. The second parametrization approach is based on facial landmark points. This chapter is concluded by introducing the concept of multi-view and pose-invariant facial expression, followed by organization of the thesis.

(ii) In **Chapter 2**, a review on different facial expression recognition algorithms for frontal face images and multi-view or view-invariant face images is presented. The existing facial feature extraction techniques can be grouped into three main categories. These categories are: 1) face-shape-free-based methods, 2) face-shape-based methods, and 3) salient/informative-region-based feature extraction methods. In general, face-shape-free-based methods divide a face into a number of sub-blocks, and then features are extracted from each of the sub-blocks. In local methods, either features are extracted from all the sub-blocks or a subset of sub-blocks of a face. On the other hand, global-based methods use either intensity values or their transformed values to form a feature vector. Dimensions of feature vectors extracted by global methods are generally very high, and hence "course-of-dimensionality" is an issue for these methods. Literature also showed

that local-based methods can give better performance than global-based methods, and that is why we adopted a local feature extraction method in our proposed method.

Shape-based methods are suitable for extracting both geometrical and texture features. However, most of the existing standard face models are only developed with an objective to extract mainly geometric features, and hence texture features extracted using these face models may not be discriminative. The salient-region based methods focused on extracting facial features only from informative regions of a face. However, few research works have been reported in the direction of extracting texture features from active/informative regions of a face, and hence, we explored extracting features from the most informative regions of a face.

In **Chapter 2**, challenges of multi-view facial expression recognition are also discussed. In literature, multi-view FER is performed in three different ways: 1) view-wise extension of the frontal face-based FER methods, 2) view-normalization before recognition, and 3) learning of a common discriminative space. Finally, Chapter 2 is ended up with motivation, objectives, and discussion on datasets used in our experiments.

(iii) **Chapter 3** focused on developing a method for extracting informative regions of a face. This is accomplished by projecting different sub-regions of expressive face images onto their corresponding neutral face images. However, neutral face images may not be available in many practical situations, and hence, a method based on procrustes analysis is proposed to obtain a common/reference image. Furthermore, our proposed projection-based analysis assigns different weights to different facial regions based on their relative importance in facial expressions. Proposed method is evaluated on three standard benchmark datasets namely: MUG, JAFFE, and CK+, and it is observed that overall performance of our proposed method is better than the state-of-the-art salient region-based FER methods.

(iv) In **Chapter 4**, an informative region-based face model is proposed. The proposed face model comprises of 54 landmark points. One noticeable advantage of our proposed face model is that both geometrical and texture discriminative features can be extracted from

the landmark points marked in our face model. As a case study, the proposed face model is employed for recognizing few sign language gestures only with the help of associated facial expressions.

(v) In **Chapter 5**, an attempt has been made to recognize prototypic expressions from multi-view face images. For this, uncorrelated multi-view discriminant locality preserving projection analysis (UMvDLPP) is proposed to recognize expressions from a set of pre-defined views ($-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$, $45°$). More specifically, we addressed the issues of handling of multi-modal characteristics of multi-view data. The proposed approach uses locality preserving projection (LPP) to preserve local geometric structure of data in a low dimensional common space. Additionally, separation between classes is enhanced by utilizing local between-class scatter matrix (LBCSM). The performance of our proposed method is validated on BU3DFE dataset.

(vi) In **Chapter 6**, an efficient approach termed as multi-level uncorrelated discriminative shared Gaussian process latent variable model (ML-UDSGPLVM) is proposed for multi-view FER. More specifically, an efficient prior based on LPP and LBCSM is proposed to enhance discriminative ability of DS-GPLVM. Additionally, our proposed ML-UDSGPLVM is implemented in a hierarchical fashion, where underline prototypic expressions are first grouped into three categories. The common spaces for each of the categories learned by 1-UDSGPLVM. Further, a 2-UDSGPLVM is learned for each of the constituent expressions present in each of the categories. Hence, there are three different 2-UDSGPLVMs – one for each sub-category. Performance of ML-UDSGPLVM is evaluated on BU3DFE dataset, and it is compared with the state-of-the-art linear and non-linear learning-based methods.

7.2 Future work

Our proposed work addressed a number of existing issues of facial expression recognition. It also points to certain areas which could benefit from further research.

- First of all, my work did not explicitly address the issues of recognizing non-basic expressions, such as screaming, and yawning.
- Accurate localization of proposed facial landmark points increases discrimination of the extracted features, and this could be an important research direction.
- As our proposed face model is derived on the basis of the movements of different facial regions, and hence, it may be utilized to recognize action units as defined by Ekman *et al.* [2].
- Our multi-view FER methods recognized expressions for the views $-45°$ to $45°$, and so, the proposed model may be extended to recognize expressions for any arbitrary poses.
- There are few face detectors that deal with faces at various poses effectively. Fast and accurate face detection in wild and across poses for facial expression recognition is still a challenging research problem.
- Finally, our proposed methods are evaluated on laboratory collected data. Further research is needed to see how well the algorithms could be generalized in completely unconstrained environments.

A

Appendix

Contents

A.1 Active Appearance Model

Active Appearance Model (AAM) [89] is a principal component analysis (PCA)-based statistical approach, and it is used to model the variations in shape due to pose, expression, as well as variation in texture due to lighting conditions. So apparently, AAM is a combination of shape model (SM) and appearance/texture model (AM). Learning of AAM model requires a set of training instances having marked with facial landmark points.

Fitting of AAM to a given test image $\mathbf{I}$ is a non-linear optimization problem [163]. The computational complexity of the original active appearance model is of the order of $(n+m)^2N$, where n and m denote the number of shape and texture parameters. In general $m >> n$ and thus, the total cost becomes very high, which makes the fitting process very slow. A fast and accurate AAM is proposed in [164], where the total cost is reduced to only few times of mN. We employed the same fast-AAM technique in our proposed algorithm (Chapter 4). The objective of fast-AAM is to minimize the mean square error between the model instance and the given test face image over the model parameters. Mathematically, this optimization problem can be expressed as:

$$\arg \min_{\mathbf{p},\mathbf{c}} ||\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p})) - \mathbf{A}_0 - \mathbf{A}\mathbf{c}|| \tag{A.1}$$

where, $\mathbf{p}$ and $\mathbf{c}$ are the parameters of shape and texture models. The symbol $\mathbf{W}$ represents a piece-wise warping function. In Eqn. (A.1), $\left\{\mathbf{A}_0, \mathbf{A} \in \Re^{\{N,m\}}\right\}$ defines an appearance model, where $\mathbf{A}_0$ is the mean appearance, and $\mathbf{A}$ is a matrix of m eigenvectors corresponding to m largest eigenvalues obtained by applying PCA on shape-free texture training images. The shape-free texture images are obtained by warping each face image so that its landmark points match with the mean shape $\mathbf{s}_0$. This process removes spurious texture variations on account of shape differences. Similarly, the shape model SM: $\left\{\mathbf{s}_0, \mathbf{S} \in \Re^{\{2u,n\}}\right\}$ is defined by the mean shape $\mathbf{s}_0$ and the transformation matrix $\mathbf{S}$. The columns of $\mathbf{S}$ represent eigenvectors obtained by applying PCA on similarity-free shape instances.

Given the models, a test image $\mathbf{I}$ and its similarity-free shape $\mathbf{s}$, the model parameters can

be estimated by using Eqn. (A.2) and Eqn. (A.3) respectively.

$$\widehat{\mathbf{s}} = \mathbf{s}_0 + \mathbf{S}\mathbf{p},\ \mathbf{p} = \mathbf{S}'(\mathbf{s} - \mathbf{s}_0) \tag{A.2}$$

$$\widehat{\mathbf{I}} = \mathbf{A}_0 + \mathbf{A}\mathbf{c},\ \mathbf{c} = \mathbf{A}'(\mathbf{I} - \mathbf{A}_0) \tag{A.3}$$

The optimum values of $\mathbf{p}$ and $\mathbf{c}$ can be obtained by the method proposed in [164].

A.2 Uncorrelated Discriminant Locality Preserving Projection Analysis (UDLPP)

The UDLPP [115] seeks for a transformation matrix $\mathbf{V}$, which projects samples of high-dimensional data onto a low dimensional space such that it preserves the topology of intra-class samples of the observation space. Additionally, it also maximizes between-class-scatter-matrix of the reduced sample space. The above criterion is formulated as a minimization problem which is represented as follows:

$$J_{UDLPP}(\mathbf{V}) = \frac{J_1(\mathbf{V})}{J_2(\mathbf{V})} = \frac{tr\left(\mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V}\right)}{tr\left(\mathbf{V}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{V}\right)} \tag{A.4}$$

The numerator term $J_1(\mathbf{V}) = tr\left(\mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V}\right)$ of Eqn. (A.4) reflects the sum of the distances between the samples of the intra-class in the reduced subspace [1]. In Eqn. (A.4), $tr(\cdot)$ represents the trace of a matrix, the $N \times N$ matrix $\mathbf{L}$ is known as Laplacian matrix. The Laplacian matrix, $\mathbf{L}$, is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, and similarity matrix $\mathbf{A}$ is obtained by applying RBF kernel which is given as:

$$\mathbf{A}_{ij} = \begin{cases} \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma^2}\right) & \text{, if } c_i = c_j \\ 0\text{, otherwise} \end{cases} \tag{A.5}$$

where, σ represents width of the kernel function and c_i indicates the class of the i^{th} sample of the observation space. On the other hand, denominator term of the Eqn. (A.4) *i.e.,* $J_2(\mathbf{V}) = tr\left(\mathbf{V}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{V}\right)$ represents sum of distances of all pairs of distinct classes, and hence $J_2(\mathbf{V})$

has to be maximized for any classification problem. Furthermore, a statistical constraint has been imposed on Eqn. (A.4), so that any two components of the reduced feature vector become uncorrelated. The two feature components y_i and y_j $(i \neq j)$ of extracted feature vector $\mathbf{y} = \mathbf{V}^T\mathbf{x}$ are said to be uncorrelated iff

$$E\left[\left(y_i - E\left(y_i\right)\right)\left(y_j - E\left(y_j\right)\right)\right] = 0 \tag{A.6}$$

Hence, the final minimization problem of UDLPP becomes a constraint optimization problem which is further converted to the generalized eigenvalue problem. The eigenvectors corresponding to d smallest eigenvalues represent the first d-columns of the required transformation matrix.

LIST OF PUBLICATIONS

Journal Publications

1. **Sunil Kumar**, M.K. Bhuyan and B. K. Chakraborty, "Extraction of informative regions of a face for facial expression recognition", *IET Computer Vision*, vol. 10, no. 6, pp. 567-576, 2016.

2. **Sunil Kumar**, M.K. Bhuyan and B. K. Chakraborty, "Extraction of Texture and Geometrical Features from Informative Facial Regions for Sign Language Recognition", *Journal on Multimodal User Interfaces, Springer*, pp. 1-13, 2017.

Manuscripts under Review

1. **Sunil Kumar** and M.K. Bhuyan "Multilevel Uncorrelated Discriminative Shared Gaussian Process for Multiview Facial Expression Recognition", *IEEE Transactions on Circuits Systems Video Technology*, **(Date of communication : 03-Nov-2016)**.

2. **Sunil Kumar** and M.K. Bhuyan "Hierarchical Uncorrelated Multiview Discriminant Locality Preserving Projection for Multiview Facial Expression Recognition", *IET Biometrics*, **(To be communicated)**.

Conference Publications

1. **Sunil Kumar**, M.K. Bhuyan, and Biplab Ketan Chakraborty, "Uncorrelated multiview discriminant locality preserving projection analysis for multiview facial expression recognition", *Proceedings of Tenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2016).* ACM, 2016.

2. **Sunil Kumar**, M.K. Bhuyan and B. K. Chakraborty, "An efficient face model for facial expression recognition", *Proceedings of Twenty Second National Conference on Communication (NCC 2016), Guwahati.* pp. 1-6, 2016.

3. **Sunil Kumar**, and M.K. Bhuyan. "Neutral expression modeling in feature domain for facial expression recognition", *Proceedings of IEEE conference on Recent Advances in Intelligent Computational Systems (RAICS 2015)*, pp. 224-228, 2015.

Bibliography

[1] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *Image Processing, IEEE Transactions on*, vol. 24, no. 1, pp. 189–204, 2015.

[2] P. Ekman, W. V. Friesen, and C. P. Press, *Pictures of facial affect*. consulting psychologists press, 1975.

[3] Y. Rahulamathavan, R. C.-W. Phan, J. A. Chambers, and D. J. Parish, "Facial expression recognition in the encrypted domain based on local fisher discriminant analysis," *Affective Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 83–92, 2013.

[4] J. A. Russell, "The circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[5] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[6] L. Ma, "Facial expression recognition using 2-d dct of binarized edge images and constructive feedforward neural networks," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 4083–4088.

[7] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognition*, vol. 45, no. 1, pp. 80–91, 2012.

[8] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *Affective Computing, IEEE Transactions on*, vol. 5, no. 1, pp. 71–85, 2014.

[9] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, vol. 37, no. 1, pp. 1–9, 2007.

[10] V. Bettadapura, "Face expression recognition and analysis: the state of the art," *arXiv preprint arXiv:1203.6722*, 2012.

[11] L. Zhang and D. Tjondronegoro, "Facial expression recognition using facial movement features," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 219–229, 2011.

[12] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[13] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Exploring human visual system: study to aid the development of automatic facial expression recognition framework," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 49–54.

[14] U. Von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[15] K. J. Kantharia and G. I. Prajapati, "Facial behavior recognition using soft computing techniques: A survey," in *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. IEEE, 2015, pp. 30–34.

[16] A. Gupta and M. Garg, "A human emotion recognition system using supervised self-organising maps," in *Computing for Sustainable Global Development (INDIACom), 2014 International Conference on*. IEEE, 2014, pp. 654–659.

[17] E.-M. Seidela, U. Habela, M. Kirschnerc, R. C. Gurd, and B. Derntl, "The impact of facial emotional expressions on behavioral tendencies in females and males," *J Exp Psychol Hum Percept Perform*, vol. 36, no. 2, pp. 500–507, 2010.

[18] L.-G. Zhang, Y. Chen, G. Fang, X. Chen, and W. Gao, "A vision-based sign language recognition system using tied-mixture density hmm," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 198–204.

[19] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[20] F. Abdat, C. Maaoui, and A. Pruski, "Human-computer interaction using emotion recognition from facial expression," in *Computer Modeling and Simulation (EMS), 2011 Fifth UKSim European Symposium on*. IEEE, 2011, pp. 196–201.

[21] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction." in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5. IEEE, 2003, pp. 53–53.

[22] B. Reeves and C. Nass, *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK, 1996.

[23] M. A. Butalia, M. Ingle, and P. Kulkarni, "Facial expression recognition for security," *International Journal of Modern Engineering Research (IJMER)*, vol. 2, pp. 1449–1453, 2012.

[24] A. A. M. Al-modwahi, O. Sebetela, L. N. Batleng, B. Parhizkar, and A. H. Lashkari, "Facial expression recognition intelligent security system for real time surveillance," in *Proceedings of the International Conference on Computer Graphics and Virtual Reality (CGVR)*. WorldComp, 2012.

[25] D. Matsumoto and H. S. Hwang, "Psychological science agenda— may 2011," *Psychological Science*, 2011.

[26] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2015.

[27] C. Turkington, *The brain encyclopedia*. Facts on File, 1996.

[28] G.-B. Duchenne and R. A. Cuthbertson, *The mechanism of human facial expression*. Cambridge university press, 1990.

[29] A. Sharma and A. Dubey, "Facial expression recognition using virtual neutral image synthesis," in *National Conference on Computer Vision Pattern Recognition Image Processing and Graphics, Jaipur, India*, 2010.

[30] S. Kumar and M. Bhuyan, "Neutral expression modeling in feature domain for facial expression recognition," in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2015, pp. 224–228.

[31] M. Video and SNHC, "Text of iso/iec fdis 14 496-3: Audio," in *Doc. ISO/MPEG N2503*, Oct. 1998.

[32] J. Harrigan and R. Rosenthal, *New handbook of methods in nonverbal behavior research*. Oxford University Press, 2008.

[33] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1357–1369, 2013.

[34] L. Zhang, D. Tjondronegoro, and V. Chandran, "Evaluation of texture and geometry for dimensional facial expression recognition," in *digital image computing techniques and applications (DICTA), 2011 International Conference on*. IEEE, 2011, pp. 620–626.

[35] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[36] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 541–558, 2011.

[37] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.

[38] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2106–2112.

[39] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[40] H. K. Ekenel and R. Stiefelhagen, "Why is facial occlusion a challenging problem?" in *International Conference on Biometrics*. Springer, 2009, pp. 299–308.

[41] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing*, vol. 26, no. 7, pp. 1052–1067, 2008.

[42] 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," *Image Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 1386–1398, 2015.

[43] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2207–2216.

[44] A. Mehrabian, "Communication without words," *Psychol. Today,*, vol. 2, no. 9, pp. 52–55, 1968.

[45] N. Holt, A. Bremner, E. Sutherland, M. Vliek, M. Passer, R. Smith *et al.*, *Psychology: the science of mind and behaviour.-*. McGraw Hill Higher Education, 2012.

[46] E. Owusu, Y. Zhan, and Q. R. Mao, "A neural-adaboost based facial expression recognition system," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3383–3390, 2014.

[47] S. Kumar, M. Bhuyan, and B. Chakraborty, "Extraction of informative regions of a face for facial expression recognition," *IET Computer Vision*, vol. 10, no. 6, pp. 567–576, 2016.

[48] Y. Pang, Y. Yuan, and X. Li, "Iterative subspace analysis based on feature line distance," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 903–907, 2009.

[49] S. R. V. Kittusamy and V. Chakrapani, "Facial expressions recognition using eigenspaces," *Journal of Computer Science*, vol. 8, no. 10, p. 1674, 2012.

[50] J. Kalita and K. Das, "Recognition of facial expression using eigenvector based distributed features and euclidean distance based decision making technique," *arXiv preprint arXiv:1303.0635*, 2013.

[51] Z. Abidin and A. Harjoko, "A neural network based facial expression recognition using fisherface," *International Journal of Computer Applications*, vol. 59, no. 3, 2012.

[52] V. J. Mistry and M. M. Goyani, "A literature survey on facial expression recognition using global features," *Int. J. Eng. Adv. Technol*, vol. 2, no. 4, pp. 653–657, 2013.

[53] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, "Learning spatiotemporal features by using independent component analysis with application to facial expression recognition," *Neurocomputing*, vol. 93, pp. 126–132, 2012.

[54] J. K. Karande, N. S. Talbar, and S. S. Inamdar, "Face recognition using oriented laplacian of Gaussian (olog) and independent component analysis (ica)," in *Digital Information and Communication Technology and it's Applications (DICTAP), 2012 Second International Conference on*. IEEE, 2012, pp. 99–103.

[55] M. Z. Uddin, J. Lee, and T.-S. Kim, "An enhanced independent component-based human facial expression recognition from video," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2216–2224, 2009.

[56] L. Asiedu, F. O. Mettle, and E. N. Nortey, "Recognition of facial expressions using principal component analysis and singular value decomposition," *International Journal of Statistics and Systems*, vol. 9, no. 2, pp. 157–172, 2014.

[57] M. Kumbhar, A. Jadhav, and M. Patil, "Facial expression recognition based on image feature," *International Journal of Computer and Communication Engineering*, vol. 1, no. 2, p. 117, 2012.

[58] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized, parts-based representation," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–207.

[59] I. Buciu and I. Pitas, "Application of non-negative and local non negative matrix factorization to facial expression recognition," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1. IEEE, 2004, pp. 288–291.

[60] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[61] S. Zhang, X. Zhao, and B. Lei, "Facial expression recognition based on local binary patterns and local fisher discriminant analysis," *WSEAS Trans. Signal Process*, vol. 8, no. 1, pp. 21–31, 2012.

[62] 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*, vol. 29, no. 6, pp. 915–928, 2007.

[63] X. Huang, G. Zhao, W. Zheng, and M. Pietikainen, "Spatiotemporal local monogenic binary patterns for facial expression recognition," *IEEE Signal Processing Letters*, vol. 19, no. 5, pp. 243–246, 2012.

[64] W. Zhang, S. Shan, X. Chen, and W. Gao, "Local Gabor binary patterns based on kullback–leibler divergence for partially occluded face recognition," *IEEE signal processing letters*, vol. 14, no. 11, pp. 875–878, 2007.

[65] M. Kaur, R. Vashisht, and N. Neeru, "Recognition of facial expressions with principal component analysis and singular value decomposition," *International Journal of Computer Applications*, vol. 9, no. 12, pp. 36–40, 2010.

[66] A. Halder, A. Jati, G. Singh, A. Konar, A. Chakraborty, and R. Janarthanan, "Facial action point based emotion recognition by principal component analysis," in *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011*. Springer, 2012, pp. 721–733.

[67] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 257–258.

[68] C. Liu, "Gabor-based kernel pca with fractional power polynomial models for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 5, pp. 572–581, 2004.

[69] B. Oshidari and B. N. Araabi, "An effective feature extraction method for facial expression recognition using adaptive Gabor wavelet," in *Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on*, vol. 2. IEEE, 2010, pp. 776–780.

[70] V. Praseeda Lekshmi and M. Sasikumar, "Analysis of facial expression using Gabor and svm," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 1–43, 2009.

[71] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 454–459.

[72] T. S. Lee, "Image representation using 2d Gabor wavelets," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[73] S. M. Lajevardi and Z. M. Hussain, "Automatic facial expression recognition: feature extraction and selection," *Signal, Image and video processing*, vol. 6, no. 1, pp. 159–169, 2012.

[74] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Transactions on image processing*, vol. 16, no. 10, pp. 2617–2628, 2007.

[75] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[76] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[77] X. Zhao and S. Zhang, "Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding," *EURASIP journal on Advances in signal processing*, vol. 2012, no. 1, pp. 1–9, 2012.

[78] A. Majumder, L. Behera, and V. K. Subramanian, "Local binary pattern based facial expression recognition using self-organizing map," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 2375–2382.

[79] A. Vupputuri and S. Meher, "Facial expression recognition using local binary patterns and kullback leibler divergence," in *Communications and Signal Processing (ICCSP), 2015 International Conference on*. IEEE, 2015, pp. 0349–0353.

[80] W.-L. Chao, J.-J. Ding, and J.-Z. Liu, "Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection," *Signal Processing*, vol. 117, pp. 1–10, 2015.

[81] H. Zhou, R. Wang, and C. Wang, "A novel extended local-binary-pattern operator for texture analysis," *Information Sciences*, vol. 178, no. 22, pp. 4314–4325, 2008.

[82] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE transactions on image processing*, vol. 22, no. 5, pp. 1740–1752, 2013.

[83] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 533–544, 2010.

[84] T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI journal*, vol. 32, no. 5, pp. 784–794, 2010.

[85] M. H. Kabir, T. Jabid, and O. Chae, "A local directional pattern variance (ldpv) based face descriptor for human facial expression recognition," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on.* IEEE, 2010, pp. 526–532.

[86] 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*, vol. 58, no. 6, pp. 1233–1258, 1987.

[87] H.-B. Deng, L.-W. Jin, L.-X. Zhen, and J.-C. Huang, "A new facial expression recognition method based on local Gabor filter bank and pca plus lda," *International Journal of Information Technology*, vol. 11, no. 11, pp. 86–96, 2005.

[88] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[89] T. F. Cootes, G. J. Edwards, C. J. Taylor *et al.*, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[90] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. S. Huang, "A study of non-frontal-view facial expressions recognition," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.* IEEE, 2008, pp. 1–4.

[91] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[92] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "A novel approach to expression recognition from non-frontal face images," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 1901–1908.

[93] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-invariant facial expression recognition using variable-intensity templates," *International Journal of Computer Vision*, vol. 83, no. 2, pp. 178–194, 2009.

[94] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 306–313.

[95] R.-L. Vieriu, S. Tulyakov, S. Semeniuta, E. Sangineto, and N. Sebe, "Facial expression recognition under a wide range of head poses," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–7.

[96] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[97] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel, "Multi-view facial expression recognition using local appearance features," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3533–3536.

[98] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[99] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16. MIT, 2004, p. 153.

[100] S. Kumar, M. Bhuyan, and B. K. Chakraborty, "An efficient face model for facial expression recognition," in *Communication (NCC), 2016 Twenty Second National Conference on*. IEEE, 2016, pp. 1–6.

[101] S. Kumar, M. K. Bhuyan, and B. K. Chakraborty, "An efficient face model for facial expression recognition," in *2016 Twenty Second National Conference on Communication (NCC)*, March 2016, pp. 1–6.

[102] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2562–2569.

[103] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, "Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis," in *European Conference on Computer Vision*. Springer, 2014, pp. 151–166.

[104] Z. Zhu and Q. Ji, "Robust real-time face pose and facial expression recovery," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 681–688.

[105] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[106] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," in *Feature extraction*. Springer, 2006, pp. 315–324.

[107] O. Rudovic, I. Patras, and M. Pantic, "Regression-based multi-view facial expression recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 4121–4124.

[108] ——, "Coupled Gaussian process regression for pose-invariant facial expression recognition," in *European Conference on Computer Vision*. Springer, 2010, pp. 350–363.

[109] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2160–2167.

[110] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *European Conference on Computer Vision*. Springer, 2012, pp. 808–821.

[111] ——, "Multi-view discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[112] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[113] G. Zhong, W.-J. Li, D.-Y. Yeung, X. Hou, C.-L. Liu *et al.*, "Gaussian process latent random field." in *AAAI*, 2010.

[114] M. B. Christopher, "Pattern recognition and machine learning," *Company New York*, vol. 16, no. 4, 2006.

[115] X. Yu and X. Wang, "Uncorrelated discriminant locality preserving projections," *IEEE Signal Processing Letters*, vol. 15, pp. 361–364, 2008.

[116] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1027–1061, 2007.

[117] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[118] 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*. IEEE, 2000, pp. 46–53.

[119] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, 2010, pp. 1–4.

[120] J. Harrigan, R. Rosenthal, and K. Scherer, "New handbook of methods in nonverbal behavior research," *(Oxford University Press, 2008),* pp. 22.

[121] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2559–2573, 2013.

[122] S. Zhang, X. Zhao, and B. Lei, "Facial expression recognition using local fisher discriminant analysis," in *International Conference on Computer Science, Environment, Ecoinformatics, and Education*. Springer, 2011, pp. 443–448.

[123] S. Dongcheng, C. Fang, and D. Guangyi, "Facial expression recognition based on Gabor wavelet phase features," in *Image and Graphics (ICIG), 2013 Seventh International Conference on*. IEEE, 2013, pp. 520–523.

[124] W. Xue, "Facial expression recognition based on Gabor filter and svm," *Chinese Journal of Electronics*, vol. 15, no. 4A, p. 809, 2006.

[125] R. Azmi and S. Yegane, "Facial expression recognition in the presence of occlusion using local Gabor binary patterns," in *20th Iranian Conference on Electrical Engineering (ICEE2012)*. IEEE, 2012, pp. 742–747.

[126] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[127] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[128] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, 2011.

[129] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 285–339, 1991.

[130] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.

[131] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage, 2004.

[132] M. Kim and V. Pavlovic, "Hidden conditional ordinal random fields for sequence classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 51–65.

[133] K. D. Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. D. Brabanter, K. Pelckmans, B. D. Moor, J. Vandewalle, and J. A. K. Suykens, *LS-SVMlab toolbox users guide 1.7*, 2010.

[134] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "Emotion recognition from arbitrary view facial images," in *European Conference on Computer Vision*. Springer, 2010, pp. 490–503.

[135] I. Ari, A. Uyar, and L. Akarun, "Facial feature tracking and expression recognition for sign language," in *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*. IEEE, 2008, pp. 1–6.

[136] H.-D. Yang and S.-W. Lee, "Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4. IEEE, 2011, pp. 1726–1731.

[137] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 593–600.

[138] T. D. Nguyen and S. Ranganath, "Tracking facial features under occlusions and recognizing facial expressions in sign language," in *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, Sept 2008, pp. 1–7.

[139] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[140] S. Ong and S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 873–891, Jun 2005.

[141] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, Jun 1997, pp. 994–999.

[142] C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, 1999, pp. 116–122.

[143] H.-D. Yang and S.-W. Lee, "Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4, July 2011, pp. 1726–1731.

[144] S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 1521–1527.

[145] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of tangent distance and an image distortion model for appearance-based sign language recognition," in *Pattern Recognition*. Springer, 2005, pp. 401–408.

[146] U. Tariq, J. Yang, and T. S. Huang, "Multi-view facial expression recognition analysis with generic sparse coding feature," in *European Conference on Computer Vision*. Springer, 2012, pp. 578–588.

[147] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[148] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[149] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE transactions on image processing*, vol. 11, no. 3, pp. 293–305, 2002.

[150] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 2010, pp. 1–4.

[151] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.

[152] Z. Zheng, F. Yang, W. Tan, J. Jia, and J. Yang, "Gabor feature-based face recognition using supervised locality preserving projection," *Signal Processing*, vol. 87, no. 10, pp. 2473–2483, 2007.

[153] T.-K. Kim, J. Kittler, and R. Cipolla, "Learning discriminative canonical correlations for object recognition with image sets," in *European Conference on Computer Vision*. Springer, 2006, pp. 251–262.

[154] R. Urtasun and T. Darrell, "Discriminative Gaussian process latent variable model for classification," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 927–934.

[155] A. Shon, K. Grochow, A. Hertzmann, and R. P. Rao, "Learning shared latent structure for image synthesis and robotic imitation," in *Advances in Neural Information Processing Systems*, 2005, pp. 1233–1240.

[156] C. H. Ek and P. Lawrence, "Shared Gaussian process latent variable models," Ph.D. dissertation, PhD thesis, 2009.

[157] M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff, "The contribution of different facial regions to the recognition of conversational expressions," *Journal of vision*, vol. 8, no. 8, pp. 1–1, 2008.

[158] X. Zhu, J. D. Lafferty, and Z. Ghahramani, "Semi-supervised learning: From Gaussian fields to Gaussian processes," 2003.

[159] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in neural information processing systems*, 2004, pp. 1601–1608.

[160] N. D. Lawrence and J. Quiñonero-Candela, "Local distance preservation in the gp-lvm through back constraints," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 513–520.

[161] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[162] C. E. Rasmussen, "Gaussian processes for machine learning," *Cambridge, MA, USA:*, vol. 1, 2006.

[163] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[164] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 593–600.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati