Dynamic 3D Facial Expression Recognition

Payam Zarbakhsh

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical and Electronic Engineering

Eastern Mediterranean University
August 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy in Electrical and Electronic Engineering.

Prof. Dr. Hasan Demirel
Chair, Department of Electrical and Electronic Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Electrical and Electronic Engineering.

Prof. Dr. Hasan Demirel
Supervisor

Examining Committee

1. Prof. Dr. Hasan Demirel _____

2. Prof. Dr. Muhittin Gökmen _____

3. Prof. Dr. Fikret Gürgen _____

4. Assoc. Prof. Dr. Rasime Uyguroğlu _____

5. Asst. Prof. Dr. Kamil Yurtkan _____

ABSTRACT

In this study, dynamic 3D facial expression recognition is addressed by proposing novel landmark-based and appearance-based approaches. As a preliminary work, a set of geometric landmark-based features are extracted from 3D images, followed by sequential forward feature selection (SFFS) and a two-layered support vector machine (SVM), fuzzy SVM classifier to recognize six basic expressions. Experiments conducted on BU-3DFE data set proved that the proposed algorithm outperforms the conventional methods advocating the effectiveness of geometric landmark-based methods.

In the second phase, a novel method using time series analysis of landmark-based geometric deformations is proposed for dynamic 3D facial expression recognition. After head pose correction and normalization, a set of multimodal time series are constructed from the local temporal deformations by applying a sliding window averaging on a comprehensive set of geometric landmark-based deformations (point, distance and angle). This stage is interlocked with facial action unit analysis to identify the key points from facial landmarks. Then, neighborhood component analysis feature selection (NCFS) is utilized to discard redundant features. Finally, adaptive cost dynamic time warping (AC-DTW) is applied to classify six prototypic expressions. Experiments on BU-4DFE data set confirmed the effectiveness of the proposed algorithm.

In the third phase, an appearance-based dynamic 3D facial expression recognition is proposed using low-rank sparse codes and a novel spatiotemporal region of interest

(ROIs) pooling. 12 ROIs are defined using automatically detected and tracked landmarks in by applying a multi-point tracker. LBP-TOP feature descriptors are extracted from cuboids inside spatiotemporal regions of interests in both texture and depth sequences and are fused to form the feature matrix. Sparse codes are obtained using low-rank sparse coding. Finally, hidden-state conditional random fields are employed to classify six basic expressions. Experimental results on BU-4DFE data set verified that proposed method improves the accuracy of dynamic facial expression recognition in comparison to previously proposed approaches.

Keywords: Dynamic 3D facial expression recognition; Spatiotemporal analysis; Geometric landmark-based deformations; Time series analysis; Dynamic time warping; Facial landmark detection; Landmark tracking; Sparse Code; Region of interest.

ÖZ

Bu çalışmada, dinamik üç boyutlu yüz ifadesi tanıma, özgün öznitelik noktaları ve görünüm tabanlı yaklaşımlar önerilerek ele alınmıştır. Önerilen ön çalışmada, 3 boyutlu görüntülerden bir dizi geometrik öznitelik noktaları çıkarılarak, ardından altı temel ifadeyi tanımlamak için sıralı ileri öznitelik seçimi (SFFS) sonrasında iki katmanlı bir sınıflandırıcı kapsamında destek vektör makinesi (SVM) ve bulanık SVM sınıflandırıcı kullanılmıştır. BU-3DFE veri seti üzerinde yapılan deneyler, önerilen algoritmanın geometrik öznitelik noktaları tabanlı yöntemin etkinliğini geleneksel yöntemleri geride bırakarak ortaya koymaktadır.

İkinci aşamada, dinamik 3D yüz ifadesi tanıma için öznitelik noktaları tabanlı geometrik deformasyonların zaman serisi analizini kullanan yeni bir yöntem önerilmiştir. Kafa poz düzeltmesi ve normalizasyondan sonra, geniş bir geometrik öznitelik noktaları temelli deformasyon setine (nokta, mesafe ve açı) kayar ortalama pencere uygulanarak yerel zamansal deformasyonlardan dizi çok-kipli bir zaman serisi oluşturulmaktadır. Bu aşama, yüz üzerinde önemli noktaları belirlemek için yüz aksiyon birimi analizi ile gerçekleştirilir. Daha sonra, fazla özellikleri azaltmak için komşu bileşen analizi özellik (NCFS) seçimi kullanılır. Son olarak, uyarlanabilir maliyetli dinamik zaman atlaması (AC-DTW) altı prototipik ifadeyi sınıflandırmak için uygulanmıştır. BU-4DFE veri seti üzerinde yapılan deneyler önerilen algoritmanın etkinliğini doğrulamaktadır.

Üçüncü aşamada, düşük sıralı seyrek kodlar ve yeni bir zamanmekansal ilgi alanı (ROI) havuzu kullanılarak görünüm temelli bir dinamik 3D yüz ifadesi ifadesi

önerilmiştir. 12 ROI, otomatik olarak algılanan ve izlenen yüz işaretleri kullanılarak çok noktalı bir izleyici uygulanarak tanımlanmaktadır. LBP-TOP öznitelik tanımlayıcıları hem doku hem de derinlik dizilerindeki ilgi alanlarının zamanmekansal bölgelerinde bulunan küplerden çıkarılır ve öznitelik matrisini oluşturmak için birleştirilir. Seyrek kodlar düşük dereceli seyrek kodlama kullanılarak elde edilmiştir. Son olarak, gizli-durum koşullu rasgele alanlar altı temel ifadeyi sınıflandırmak için kullanılmıştır. BU-4DFE veri setindeki deneysel sonuçlar, önerilen yöntemin daha önce önerilen yaklaşımlara kıyasla dinamik yüz ifadesi tanıma doğruluğunu artırdığını doğrulamıştır.

Anahtar Kelimeler: Dinamik 3D yüz ifadesi tanıma; zamanmekansal analiz; Geometrik öznitelik noktaları tabanlı deformasyonlar; Zaman serisi analizi; Dinamik zaman atlaması; Yüz öznitelik noktaları tespiti; Öznitelik noktaları izleme; Seyrek Kod; İlgi bölgesi.

DEDICATION

To my valued grandfather

Mahmood Zarbakhsh

ACKNOWLEDGMENT

I would like to gratefully thank my supervisor, the Chair of the Electrical and Electronics Engineering Department, Prof. Dr. Hasan Demirel for his continuous support, availability, concern and extraordinary patience throughout my PhD study. Undertaking this PhD and related research has truthfully been a life-changing experience for me and without his immense knowledge and insightful comments I would not be successful.

I also would like to express my sincere gratitude to my follow-up jury committee members Assoc. Prof. Dr. Rasime Uyguroğlu and Asst. Prof. Dr. Kamil Yurtkan for their constructive advices and suggestions.

I am truly grateful to my parents and my sister for their supports and encouragements. There are no words with which I could properly express my appreciation of their unwavering support.

I thank with love my wife, Dr. Sanaz Fargangi who has stood by me since the beginning of this academic journey and inspired me to work to the best of my abilities.

I thank Dr. Ghazal Sheikhi, Dr. Rasul Dehghanzadeh, and Dr. Babak Mohammadi who have provided me with unique guidance and friendship during the arrangement of my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

d_{ij}	Distance between landmarks *i* and *j*
(x_i, y_i, z_i)	3D Cartesian coordinates of landmark *i*
$p(.)$	Probability function
E_r	Bayesian Error
m_i	Mean value of class *i*
σ_i	Standard deviation of class *i*
N_i	Number of samples in class *i*
$\varphi(s)$	Kernel function of SVM
m_k	Fuzzy membership of class *k*
ξ	Error measure in FSVM
y_i	Class label of i^{th} sample
α	Lagrangian parameter
β	Lagrangian parameter
θ	Pitch angle
φ	Yaw angle
ψ	Roll angle
F	Rigid affine transformation
Θ	Rotation matrix in rigid affine transformation
Γ	translation vector in rigid affine transformation
Δx_i^l	Point deformation of l^{th} landmark in i^{th} frame
$\Delta d_i^{l,m}$	Distance deformation of landmarks *l* and *m* in i^{th} frame
$\Delta \alpha_i^{l,m,k}$	Angle deformation of landmarks *l,m* and *k* in i^{th} frame

D	Number of modes of time series features
T	Length of time series features
S_T^D	Mean deformation multimodal time series feature
w	Size of mean sliding window
L	Linear transformation vector of NCFS
$p_{i,j}$	Probability of X_j being taken as the reference for X_i *in NCFS*
p_i	Probability of correct classification in NCFS
$\Lambda(L)$	LOO accuracy of 1NN in NCFS
λ	Regularization term
$d_{pr}(i, j)$	Element of preliminary distance matrix in AC-DTW
C	Cost function of AC-DTW
r	Control term to adjust many-to-one mapping in AC-DTW
g	Controller factor for effectiveness of cost function in AC-DTW
$P(i, j)$	Optimal warping path between i and j points
$a_{i,j}$	Element of A_s to control the number of reuses in AC-DTW
$b_{i,j}$	Element of A_R to control the number of reuses in AC-DTW
G	Gaussian convolution kernel
$\theta(x, y)$	Gradient orientation of pixel (x,y)
$m(x, y)$	Gradient magnitude of pixel (x,y)
$Y_{1:k}$	Observation vector of DE-MC
Pi	Mixture weight of DE-MC
p	Posterior probability
W_k	Weight of sample candidate in DE-MC
$S(F_i, F_1)$	Mean similarity measure of frame i with respect to the first frame
F_v	Feature matrix of sequence

f_v	LBP-TOP feature vector of cuboid v
(x_p, y_p, t_p)	Spatiotemporal coordinates of center pixel of cuboid
D	LRSC codebook
R	Representation matrix in LRSC
F	Feature descriptor
λ_1	Regularization parameter for sparsity
λ_2	Regularization parameter for low-rankness
s	Hidden state of HRCF
$p(y\|X,\theta)$	Conditional probability
L	Optimization function for HCRF
1NN	One Nearest Neighbor
2D	Two-Dimensional
2D FER	Two-Dimentsional Facial Expression Recognition
3D	Three-Dimensional
3D-DCT	Three Dimensional Discrete Cosine Transform
3D FED	Three-Dimensional Facial Expression Detection
3D FER	Three-Dimensional Facial Expression Recognition
4D	Four-Dimensional
4D FER	Four-Dimentsional Facial Expression Recognition
6E	Six Expression
10-CV	Ten Fold Cross Validation
AAM	Active Appearance Model
AC-DTW	Adaptive Cost Dynamic Time Warping
ASM	Active Shape Model
AUs	Action Units

AN	Anger
CFS	Correlation Feature Selection
CRF	Conditional Random Fields
DE-MC	Differential Evolution-Markov Chain
D-FED	Dynamic Facial Expression Detection
D-FER	Dynamic Facial Expression Recognition
DI	Disgust
DOG	Difference Of Gaussian
DTW	Dynamic Time Warping
FACS	Facial Action Coding System
FE	Fear
FED	Facial Expression Detection
FER	Facial Expression Recognition
FRNN	Fuzzy Rough Set Nearest Neighborhood
FSVM	Fuzzy Support Vector Machines
GA	Genetic Algorithm
HA	Happy
HCI	Human Computer Interaction
HCRF	Hidden-state Conditional Random Fields
HMM	Hidden Markov Model
HOG	Histogram of Orientated Gradient
HOG-TOP	Histogram of Orientated Gradient- Three Orthogonal Planes
IALM	Index Augmented Lagrange Multiplier
LBP	Local Binary Patterns
LBP-TOP	Local Binary Patterns - Three Orthogonal Planes

LDA	Linear Discriminant Analysis
LLC	Locality-Constrained Linear Coding
LOO	leave-One-Out
LRSC	Low Rank Sparse Coding
MKL	Multiple Kernel Learning
NCFS	Nearest Component Feature Selection
PCA	Principal Component Analysis
PDM	Point Distribution model
RBF-SVM	Radial Basis Function Support Vector Machine
RGB	Red Green Blue
ROI	Region of Interest
S	Subject
SA	Sadness
SEQ	Sequence
SFFS	Sequential Forward Feature Selection
SIFT	Scale Invariant Features Transform
SPP	Special pyramid pooling
SU	Surprise
SVM	Support Vector Machine
SVM-FSVM	Support Vector Machine-Fuzzy Support Vector Machine

Chapter 1

INTRODUCTION

1.1 Background

Image processing and pattern recognition have been a fast developing area of research embracing a diverse range of topics including image enhancement [1], face recognition [2], facial expression recognition (FER) [3], medical image processing [4], geology [5] etc. In fact, emerging advances in image registration equipment and storage devices have provided the access to two-dimensional (2D), three-dimensional (3D) and four-dimensional (4D) data sets in a wide range of domains. Emotion analysis from facial images and videos is one the recent topics with many applications in medical care, psychology, marketing, customer service industry, education and gaming [6]. In recent years, extensive studies have been conducted on facial expression recognition by researchers in computer vision, image processing and biometrics [7]–[9].

FER systems are designed to recognize different emotions expressed by movements of facial muscles from facial images or videos. From the dimension point of view, FER data sets can be categorized in three types including 2D, 3D and 4D data sets. Both 2D and 3D data sets provide spatial information of facial expression. In 2D data sets, texture information and 2D coordinates of facial landmarks are available while 3D images contain depth information and 3D coordinates of facial landmarks. In four-dimentsional facial expression recognition (4D FER) systems, also known as

dynamic 3D facial expression recognition systems, texture, depth and 3D coordinates of facial landmark are recorded as time sequences during emotion expression.

The aim of this thesis is to conduct a comprehensive study on dynamic 3D facial expression recognition by going beyond conventional approaches in feature extraction, feature selection, classification and related issues. For simplicity, we term dynamic 3D facial expression recognition as dynamic facial expression recognition (D-FER).

1.2 Problem Definition

There are different challenges in D-FER system design. Some of these issues correspond to the dynamic characteristic of D-FER system and the obligation to capture temporal information, while others are related to representation of spatial information. In the mainstream 2D/3D image processing systems, spatial information in texture/depth images is analyzed and the main challenge is how to convert the RGB images, gray-scale images and depth vertex images into a dense representation. Several appearance-based feature descriptors are developed for this purpose [10].

However, facial expression image processing and recognition are different from general image processing and recognition in the sense that expression-related information are mainly exist in the neighborhoods of some special spots (interest points) on face [11]. These specific spots are called facial landmarks or key points which are displaced by movements of facial muscles during the expression of emotion. In fact, the regions on face which are activated to exhibit emotions are categorized by experts into several action units (AUs) based on the location and displacement of different facial elements including eyebrows, eyes, nose, cheeks, lips

and chin [12]. The deformations in these facial elements result in displacement of facial landmarks and specific AUs contribute to each emotion. Consequently, landmark locations and their displacements have been exploited in facial expression analysis and recognition [13], [14]. This feature extraction method is known as geometric landmark-based approach in FER studies.

On the other hand, conventionally used feature descriptors which extract appearance information from small patches in texture and depth images are applied in FER studies. Furthermore, it has been confirmed that incorporating AUs information in appearance-based approaches improves the performance of FER systems [15]. Integrated FER systems rely on a combination of geometric and non-geometric information. Although these techniques have been recently initiated, they yield the promising results [16], [17]. Nevertheless, an efficient scheme that incorporates the concept of AU and facial regions of interest with general appearance-based methods is still questionable. There is a need for a comprehensive analysis on landmark-based approaches to discover the potential issues, drawbacks and strengths.

Another critical challenge in D-FER systems is capturing the dynamics of the expressions. Dynamic 3D facial expression recognition aims at detection of emotions from facial video sequences. Hence, unlike static FER systems functioning on spatial data extracted from facial still images, D-FER systems are designed to be employed on spatiotemporal data. Thus, capturing transitions in temporal domain is as crucial as capturing spatial features for these systems. Mainly, dynamic facial expression recognition is addressed as a spatiotemporal problem in image processing.

The third challenge in dynamic 3D facial expression recognition originates from the novelty of the problem. Facial expression recognition from 3D video sequences has appeared in the literature in 2008 [18]. Consequently, there is a large room for investigating advanced techniques in image processing, computer vision, machine learning, and pattern recognition in this domain. Addressing the issues in dynamic 3D facial expression recognition requires an extensive knowledge on the spatiotemporal characteristics of the problem as well as the properties of the tools which can be applied on it. Therefore, different phases of a D-FER system including preprocessing, automatic landmark detection and tracking, feature extraction, feature selection, feature coding/pooling and classification may be modified with proper techniques to improve the system performance.

1.3 Objectives

The aim of this study is to explore dynamic 3D facial expression recognition from different aspects. The list of the particular objectives to reach this inclusive aim is as follows.

1. To study the appearance-based feature descriptors and geometric landmark-based feature extraction approaches in D-FER systems.
2. To propose a novel landmark-based framework for capturing and analysis of spatiotemporal information.
3. To review the existing feature selection and classification approaches in pattern recognition and machine learning which are applicable in D-FER studies.
4. To incorporate alternative feature selection and classification methods for improved recognition performance.

5. To integrate landmark-based information in appearance-based approaches.
6. To adapt recent sparse coding and pooling methods into automatic dynamic 3D facial expression recognition.

In summary, these objectives open a new horizon in dynamic 3D facial expression recognition by exploring the subject of matter from different perspectives and suggesting novel alternatives for each phase of the whole system based on the recent trends in image processing, machine learning and pattern recognition.

1.4 Contributions

The main contributions of this study are given below.

1. Sequential forward feature selection (SFFS) and fuzzy support vector machine (FSVM) classification are applied in landmark-based 3D facial expression recognition to tackle the problems of high dimensionality, high-redundancy and multi-class characteristics.
2. Multimodal time series analysis is adapted for the first time in dynamic 3D facial expression recognition based on a comprehensive set of landmark-based geometric deformations.
3. Neighborhood component feature selection method is applied as a fast and effective feature selection method to address the excessive redundancy among high-dimensional landmark-based features.
4. Adaptive cost dynamic time warping (AC-DTW) is used for the first time as a time series classification to deal with inherent temporal dynamics facial expression video sequences.

5. Low-rank sparse coding is applied for the first time in dynamic 3D facial expression recognition to encode the texture and depth feature descriptors.
6. Landmark-based information is integrated into appearance-based descriptors by proposing a novel spatiotemporal region of interest (ROI)-based pooling.

1.5 Overview

The rest of this thesis is organized as follows. In the next chapter, Chapter 2, different components of FER systems are reviewed. The feature extraction, feature selection, and classification approaches applied in static and dynamic FER systems are presented and discussed to provide a clear view of the problem. In Chapter 3, a static facial expression recognition system based on SFFS and FSVM is proposed and evaluated. Features are extracted as the pairwise Euclidean distances between the landmarks. Chapter 4 addresses the geometric landmark-based approach in dynamic 3D facial expression recognition by adapting the multimodal time series analysis to the notion to point, distance and angle deformations extracted from the displacement of facial key points in an expression sequence. Neighborhood component feature selection and AC-DTW are then applied for feature selection and classification, respectively. Incorporating landmark-based information into appearance-based information is undertaken in Chapter 5. Spatiotemporal regions of interest are identified by a set of automatically detected and tracked landmarks. Texture and depth feature descriptors are extracted from these regions of interests and are coded using low-rank sparse coding. Then conventional spatial pyramid pooling (SPP) is replaced by the proposed ROI pooling. Pooled codes are classified by hidden state

conditional random field which is an effective classification method for dynamic problems. Finally, in Chapter 6, the thesis is concluded.

Chapter 2

LITERATURE REVIEW

2.1 The Overview of Facial Expression Recognition Systems

The role of emotions in human social communication is so critical that "recognition of emotion is known as a key component of intelligence" [19]. Over the last few years, as cameras and registration technologies have advanced in the acquisition of 3D images and videos, scholars are increasingly attracted to emotion analysis from facial images and video sequences. As a matter of fact, automatic facial expression recognition has become an emerging topic of interest for both scientists and engineers in intelligent systems because of its wide applications ranging from psychology to online games.

Automatic facial expression recognition systems process the visual information of facial recordings for emotion-related analysis and recognition [6]. There are three different input types for FER system, namely 2D [7], 3D [20], [21] and 4D [3]. 2D and three-dimensional facial expression recognition (3D FER) systems handle texture and depth still images of facial expression while in 4D systems facial video sequences of texture and depth are processed. More information about these systems and their similarities and differences are given in Section 2.2 and 2.3.

Regardless of this categorization based on the type of facial recordings, there are roughly three general components of facial expression recognition systems including data acquisition, representation and classification as shown in Fig. 2.1. Facial expression data acquisition is the concern of scholars who are preparing FER data for the research community. Different issues such as illumination, pose variation, location, and the type of visual data to be registered are the main elements in data acquisition. For facial expression recognition community, these issues are taken into account in preprocessing phase. The other matter is to represent the registered pixels in a compressed feature space. Geometric landmark-based methods, active shape model (ASM) [22], active appearance model (AAM) [23], facial action coding system (FACS) [24] and appearance-based descriptors are some examples of the well-known techniques used for facial expression representation. The third main concept is the facial expression classification. Examples of widely-used classifiers in this domain are neural network, support vector machine (SVM), hidden Markov model (HMM), and conditional random fields (CRF).

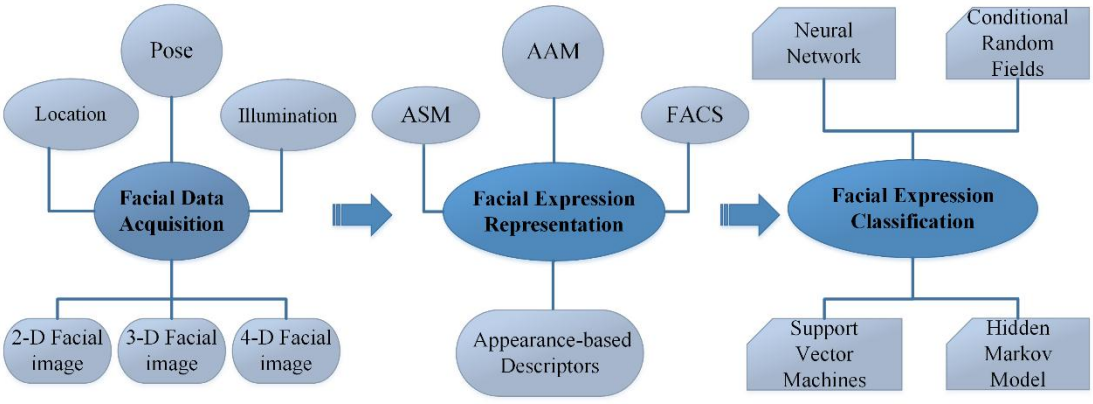


Figure 2.1: General components of facial expression recognition system.

Although the aforementioned concepts are the essential blocks for facial expression recognition, there are other processes necessary in practice. As stated before, the data

collected in acquisition phase can be affected by illumination, pose variation, clutter and other artifacts. Consequently, preprocessing should be applied on the facial expression images and videos. Moreover, the dimension of features computed to represent the expressive face is mainly high. High dimensional feature space is vulnerable to redundancy, which adds additional burden to the classifier since the training and test phases become computationally complex. Hence, feature selection and dimensionality reduction methods are extensively used in facial expression recognition systems aiming at selecting informative features and discriminative subspaces. In evaluation of the system performance, there are two main phases namely, train and test phase. In train phase, best feature subspaces are constructed from the train samples and through a learning algorithm, classifier models are trained. During the test phase, the unseen test samples are used to validate the facial expression recognition system.

Generally, a typical automatic facial expression recognition system includes four main blocks: preprocessing, feature extraction, feature selection and classification. Depending on the system architecture, preprocessing covers a wide range of functions including noise removal, illumination artifact cancelation, head pose correction, and landmark detection and tracking [25]–[27]. In feature extraction stage, image/video pixels are converted to representative feature descriptors. Feature selection and/or dimensionality reduction is the stage for addressing high-dimensionality of the extracted features. Classification is the algorithm that learns the characteristics of the features for each specific emotional state from the train data and it identifies the label for each test sample accordingly. An overview of a facial expression recognition system with the blocks involved in train and test phase is shown in Fig. 2.2.

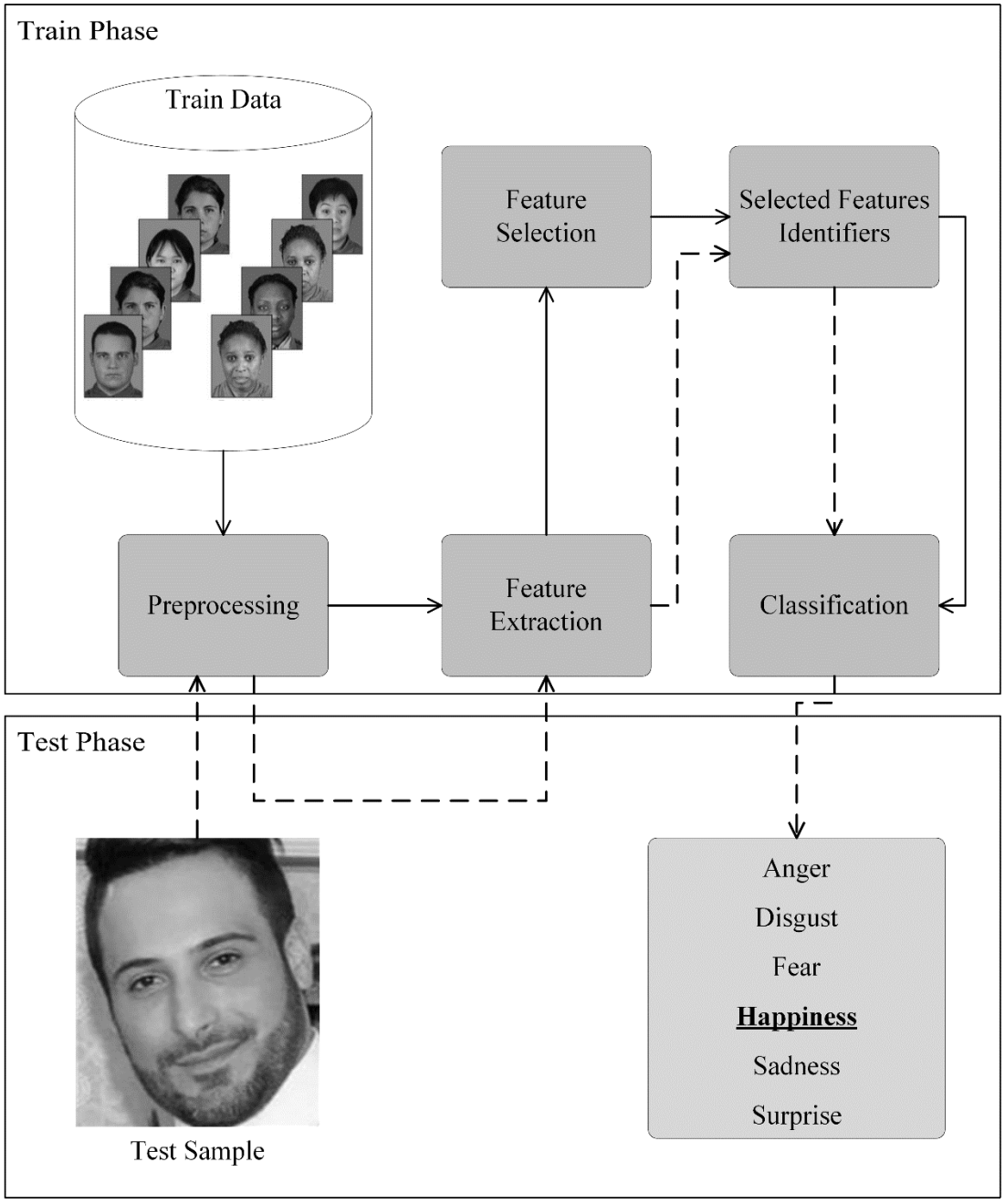


Figure 2.2: Overview of the training and test phases and the main blocks in an automatic facial expression recognition system.

For each test image or sequence, after preprocessing and feature extraction, informative features are considered. Then the label is predicted using the constructed model. In summary, the aim of FER system is to recognize the emotion from a new entry which is a facial expression image or video sequence. Facial expression recognition systems can be categorized into two main types.

Static and dynamic FER systems are developed to recognize facial still images and facial video sequences, respectively. The former relies on the spatial information of the facial expression while the latter processes spatiotemporal information. Consequently, data sets and feature descriptors are different for these two varieties of FER systems.

2.2 Static Facial Expression Recognition

In static facial expression recognition, the system explores facial still images for emotion analysis. In other words, spatial characteristics of the expressions are to apprehend for recognition task. Mainly, the images are recorded at the peak of the expression to represent maximum expression-specific deformation [28]. This phase of the expression is called the apex. The main difference between 2D and 3D FER systems is the availability of depth data in 3D systems. 2D facial images are either RGB colored images or gray-scale images capturing texture information as shown in Fig. 2.3. 3D images however, contain the depth vertices demonstrating the distances from the camera. Considering the AUs and the 3D displacement of the facial landmarks, 3D images are more informative for FER task. In fact, 3D information is valuable for improving system accuracy [21], [29].

a) b)

Figure 2.3: (a) RGB colored image and (b) gray-scale image.

As a matter of fact, although 2D and 3D FER systems are both known as static FER, there is a main issue related to 2D facial expression recognition as these systems are more susceptible to suffer from illumination changes [29], [30] and pose variations [31]. Extensive studies have been conducted on FER systems based on 2D images with significant performance [32]–[34]. But, sensitivity to illumination and pose variations are still the drawbacks. In recent decade, the progress in 3D acquisition techniques has provided a novel solution to address these concerns [29].

2.2.1 Static Facial Expression Recognition Data Sets

There are several publicly available 2D and 3D facial expression data sets. These data sets differ from some aspects such as the number of subjects, the quality of the images, the number of emotions expressed, availability for public use, providing posed versus spontaneous expressions, manually annotated landmarks, noise, clutter and artifact levels. Table 2.1 lists some of the public facial expression image data sets with their basic specifications. Depending on the aim of the FER research and the requirements, one may select among these data sets.

Table 2.1: The list of Publicly available Static FER data sets

Data Set	Expressions	#Subjects	#Images	Further Info
BU-3DFE	Anger, disgust, fear, happiness, sadness, surprise, neutral	100	2,500	83 annotated facial landmarks, 4 intensity levels
Bosphoros	Anger, disgust, fear, happiness, sadness, surprise, neutral	105	4652	includes intensity and asymmetry codes for each AU
AR Face	Smile, anger, scream	126	4,000	different illumination conditions, and occlusions (sun glasses and scarf)
Radboud Faces	Anger, disgust, fear, happiness, sadness, surprise, contempt, neutral	67	13,400	five camera angles, three gaze directions
EURECOM KFD	Neutral, smile, open mouth	52	936	6 annotated landmarks, recorded: left profile, right profile, occlusion eyes, occlusion mouth, occlusion paper, light on
AffectNet	Anger, disgust, fear, happiness, sadness, surprise, neutral, contempt	-	440000	collected from the Internet, including annotated: none, uncertain, non-face
ND-2006	Neutral, happiness, sadness, surprise, disgust, and other	888	13,450	-
FRGC v2	Anger, disgust, happiness, sadness, surprise, puffy	466	4,007	-
JAFFE	Anger, disgust, fear, happiness, sadness, surprise, neutral	10	213	-

2.2.2 Feature Extraction in Static Facial Expression Recognition

From feature extraction point of view, studies in FER domain pursue two main streams: Non-geometric approach and geometric (landmark-based) approach. Non-geometric approaches also known as appearance-based feature extraction methods rely on features extracted from texture and depth images. On the other hand, geometric methods include landmark-based feature extraction approaches relying on the relative deformations of facial key points [22], [23] as well as curvature features extracted from 3D vertices [35].

Moreover, in recent years, there have been some efforts to combine landmark-based features with appearance-based features by extracting feature descriptors from the zones in landmarks' neighborhood [17]. We termed these approaches as integrated feature extraction methods to avoid confusion. In this section, we review the commonly used feature extraction methods applied in static facial expression recognition.

Regardless of recording facial expression still images either as 2D or 3D, the common appearance-based approaches for feature extraction in FER systems are local spatial feature descriptors [36]–[39]. Local feature descriptors such as local binary patterns (LBP) [40], [41] , histogram of orientated gradient (HOG) [42], scale invariant features transform (SIFT) [40] and their variations [43], [44] are intended to capture regional topology and geometry of the emotion in small patches on facial images.

In computing local feature descriptors, the image is firstly partitioned into small patches. For each patch of the image, the pixel intensities are coded into a number

that characterizes the local spatial content. Fig. 2.4 shows a sample gray-scale image and the procedure of obtaining local binary pattern (LBP) descriptor. In Shan2009 [40], the authors have conducted a comprehensive set of experiments using LBP features and different classifiers to show that these local descriptors are efficient in capturing the properties of the human face for different emotion expressions. Their findings have also proved that LBP can perform successfully for low-quality videos with low-resolution.

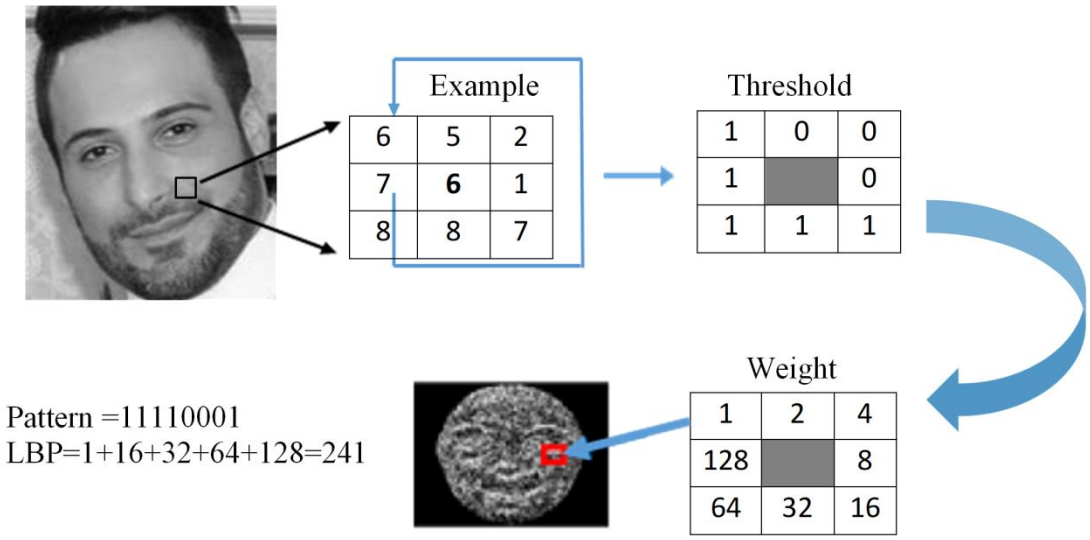


Figure 2.4: A sample gray-scale facial image and computation of local binary pattern feature descriptor.

A comparative study has been carried on by [41] to evaluate the performance of local ternary patterns (LTP), LBP, Gabor and HOG in facial expression recognition. They have shown that LBP feature descriptors provide superior results in comparison to others. In addition, a novel variation of conventional LBP has been proposed by [43]. Compound local binary pattern is a combination of the original LBP with the magnitude information of the difference between two gray values and it has employed for facial expression recognition effectively.

The competence of HOG feature descriptors in FER systems has been extensively studied in [42]. They have estimated the performance of FER systems using HOG feature descriptors considering different number of orientation bins and different cell sizes. It has been argued that with an appropriate parameter setting, HOG descriptor can function remarkably well in FER systems [42]. Median ternary pattern (MTP) is the name of a new feature descriptor that incorporates a coded 3-valued quantized gray-scale value and the median filtering benefits [44].It should be noted that appearance-based feature descriptors are basically extracted from both texture and depth images in 3D cases and then they are fused to represent geometrical and topological properties of the face in three dimensions while expressing an emotion.

The local descriptors are flexible feature extraction methods applicable in almost all image analysis and image processing tasks. In recent years, some other approaches for appearance-based local descriptors have been proposed and utilized in FER systems such as principal component analysis (PCA) [45] and wavelet-based local features [46].

Unlike appearance-based features that capture the local properties of the face in small image patches, landmark-based feature extraction approaches rely on the movement of the facial landmarks. In general, the coordinates of the points shaping facial segments, their distances and the angles between the lines connecting them are all known to signify deformation of the face in facial expression. Fig. 2.5 illustrates the basic landmarks recognized to represent the deformation of eyebrows, eyes, and mouth with related landmark-based features [47]. Fernandes et al. [48] have applied the point distribution model (PDM) to model the deformable expressive face using a set of points. The distances between these points have been considered as the

features. They have found that by applying a feature selection algorithm namely correlation feature selection (CFS), the performance of the facial expression recognition is improved significantly compared to the features selected manually by a human expert [48].

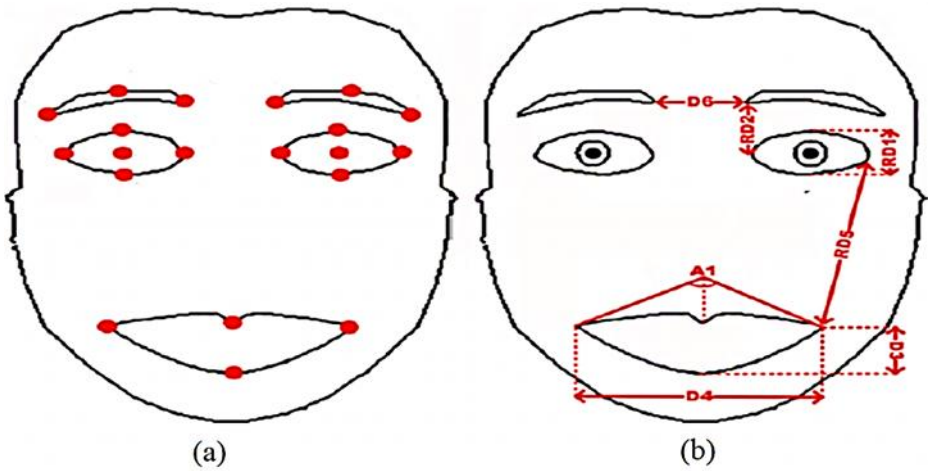


Figure 2.5: (a) Basic facial landmarks and (b) geometric landmark-based features[47].

Additionally, it has been shown that landmark positions are capable of signifying the deformation in face shape during an emotion expression [21]–[23], [48]–[50]. Basically, geometric landmark-based features are defined as the vectors of pairwise Euclidean distances between a set of facial landmarks located mainly around eyebrows, eyes, nose, and lips in several studies. It has been confirmed that distance-based features perform satisfactory in FER systems with proper feature selection and classification methods [13], [20], [21], [50]. The geometric approach proposed by [49] , is based on the Euclidean distance of the center of gravity of the facial image and automatically detected key points. Fiducial points are automatically annotated by ASM in advance. Geometric features are computed as the deformation between the neutral face of the subject and any of the expressive faces. These deformations contain discriminative information related to the expressed emotion [49].

ASM is another geometric approach being employed by [22] where the displacement of the facial feature points as well as the mean shape of the ASM for each expression are exploited to identify the emotions. AAM is also used to identify key points on the face. The deviations in the values of the annotated key feature points in relation to neutral face are then observed by a fuzzy logic system [22]. Another study by [35] have suggested the geometric descriptors to be driven from the 3D mesh of the face using triangular mesh models and principal curvature information.

Although both appearance-based descriptors and geometric landmark-based features have been employed effectively in FER domain, each of them has been criticized for some drawbacks. Appearance-based features are computationally costly and they reflect the general geometry, texture and topology of an image. Hence, the amount of information captured by these descriptors is very high and mainly superfluous for FER systems. On the other hand, landmark-based features are easy to compute and they principally carry the expression-related deformation. However, this category of features relies on the annotated facial points and they do not reflect the texture and fine deformations. As stated before, there are some studies that have combined the appearance-based features with geometric features to overcome these drawbacks. This approach includes a range of schemes including extracting feature descriptors from specific regions of the face (such as the patches around some landmarks) and the pre-identified regions of interest [51]–[53].In [53], Gabor wavelet features are extracted from the facial regions modeled by AAM. In [51], [52] the HOG feature descriptors is extracted from specific facial components defined according to facial AUs function during the expressions.

2.3 Dynamic Facial Expression Recognition

The dynamic 3D facial expression recognition which is also known as 4D FER includes inspecting 3D videos for facial expression recognition. D-FER systems are designed to process facial expression video sequences and as the information content of sequences exists both in spatial and temporal properties, conventional approaches adapted for static FER systems may fail in this domain. 4D FER is a new challenging field of study amongst image and video processing community. To provide some examples as the first published works in this field, we can mention the studies by [18], [54] published in 2008 and 2010. Since then, it has been a growing interest among the scholars to investigate dynamic facial expression recognition problem.

2.3.1 Dynamic Facial Expression Recognition Data Sets

The urge to study human emotions expressed in face has been the main motivation for FER community to collect and publish dynamic facial expression recognition data sets. Registering, processing, annotation and labeling of facial video sequences is much more complicated and time consuming than facial still images taken at the peak of the expression named apex. In fact, dynamic FER data sets contain video sequences captured during the whole expression procedure starting from neutral to onset, apex, and offset phases as shown in Fig. 2.6. These sample frames of happiness expression were taken from the texture (top row) and depth (bottom row) video sequences of binghamton university four-dimensional facial expression database (BU-4DFE) data set [55]. Although several data sets have introduced in recent decades for dynamic facial expression recognition studies, there are a few widely-studied ones based on their public availability, number of expressions, number of subjects, ground truth material and its reliability.

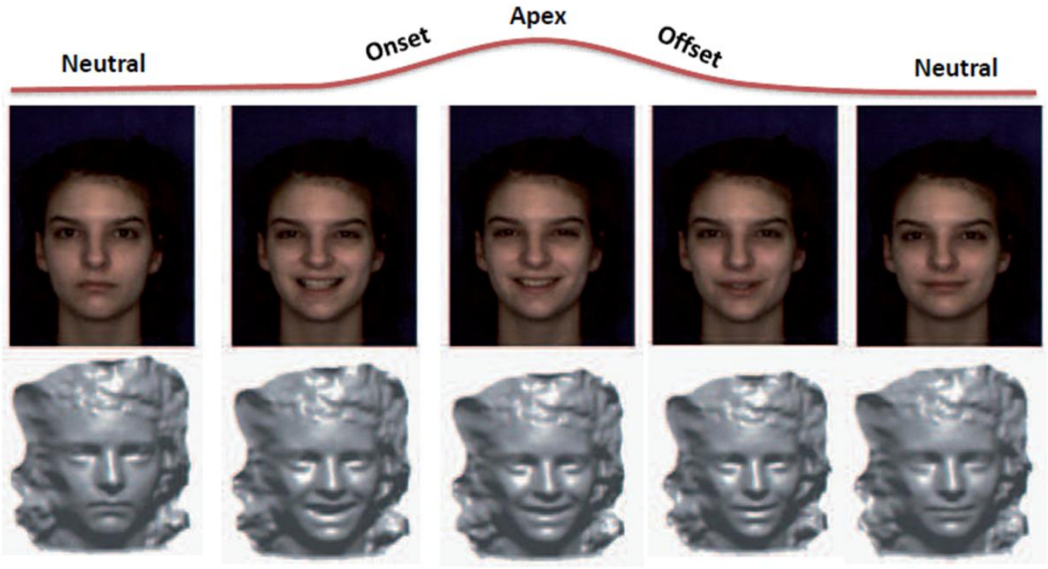


Figure 2.6: A sample 4D facial expression record: texture sequence (first row) and depth sequence (second row) [55].

During the last few years, dynamic facial expression data sets have been introduced and studied by researchers [56]–[63]. As a matter of fact, one of the reasons of growing publications in dynamic facial expression recognition is the availability of the data sets. Dynamic 3D facial expression recognition data sets or 4D FER data sets are collected as 3D video sequences. In other words, texture and depth frames are recorded as videos which means 3D spatial information as well as temporal transition information are available.

The list of popular dynamic facial expression recognition data sets with a summary of their main characteristics is presented in Table 2.2. This list may not include all the available data sets published and studied in recent decade but the widely-studied ones are considered. Depending on the undertaken research task, scholars can select among these data sets based on the number of subjects, number of sequences, expressed emotions, different intensity levels and the availability and reliability of the ground truth information.

Table 2.2: The list of publicly available D-FER data sets

Data Set	Expressions	#Subjects	#Sequences	Further Info
ADFES	Anger, contempt, disgust, embarrassment, fear, joy, neutral, pride, sadness, surprise	20	648	FACS coding, arousal and valence ratings,
BP-4D	Anger/upset, disgust, embarrassment, fear/nervous, happiness/amusement, pain, sadness, surprise/startle	41	328	FACS coding, spontaneous
BU-4DFE	Anger, disgust, fear, happiness, sadness, surprise	101	606	83 annotated landmarks on each frame
CK	AU sequences for anger, disgust, fear, joy, surprise, sadness	97	486	FACS coding
CK+	AU sequences for anger, contempt, disgust, fear, happy, sadness, surprise	123	593	FACS coding, Spontaneous smiles
MMI	AU sequences for anger, bored, disgust, fear, happy, sad, sleepy, surprise	75	2900	FACS coding, frontal and side viewpoints
Hi4D-ADSIP	Anger, disgust, fear, happiness, pain, sadness, surprise, other	80	3,360	3 intensity levels
UT Dallas	Anger, boredom, disbelief, disgust, fear, happiness, laughter, neutral, puzzlement, sadness, surprise	284	2,556	spontaneous, 9 viewpoints
DaFex	Anger, disgust, fear, happiness, neutral, sadness, surprise	8	1,008	3 intensity levels, audiovisual recordings,
STOIC	Anger, disgust, fear, happiness, neutral, pain, sadness, surprise	10	80	3 intensities levels

2.3.2 Feature Extraction in Dynamic Facial Expression Recognition

Feature extraction methods adapted for D-FER systems may differ from the conventional local feature descriptors previously described for static FER systems. Facial expression is a dynamic procedure containing transition. The time interval during which an emotion is expressed is remarkably informative. Transitions in this interval contain critical temporal information for facial expression recognition [57], [60], [64] .

In fact, static feature extraction approaches adapted for representing spatial information fail to capture and exploit temporal information in D-FER systems. However, the earlier published works in dynamic facial expression recognition have used conventional static descriptors for extracting features, and tried to model the dynamics by further processing the extracted features in preceding phases of the D-FER system[28], [65] . For instance, Bartlett et al. [65] have proposed a dynamic facial expression system based on Gabor filter bank. The images (frames of the video sequences) were converted into a Gabor magnitude representation. A bank of Gabor filters at 8 orientations and 9 spatial frequencies have applied to obtain the descriptors.

More recently, scholars have been proposing reformed variations of the conventional appearance-based descriptors or novel feature extraction approaches adapted specifically to the dynamic nature of D-FER systems [64], [66]. As stated in Section 2.2.2, feature extraction methods in the FER studies pursue two main streams: appearance-based approach and geometric landmark-based approach. The same applies on D-FER systems but the main challenge is representing the dynamics of the expression by either appearance-based or landmark-based features. Similar to static

facial expression recognition, non-geometric feature extraction approaches, mainly referred as appearance-based descriptors, rely on features extracted from texture and depth video sequences. On the other hand, geometric landmark-based and curvature-based features are extracted from the variations in the geometry of the landmark positions and the 3D mesh of the face respectively.

There are several conventionally used local feature descriptors in image processing which are modified to be applicable in D-FER systems. Descriptors such as local binary pattern in three orthogonal planes [59], histogram oriented gradients from Three Orthogonal Planes [67], expression lets [68], and spatiotemporal texture map [58] are examples of the reformed descriptors suggested and evaluated by the researchers for dynamic 3D facial expression recognition.

Local binary patterns-three orthogonal planes (LBP-TOP) and another variant of LBP, i.e. volume local binary pattern has been proposed by Zhao and Pietikainen [69] to address the issue of dynamic texture recognition in D-FER systems. Recently, Shao et al. [59] have successfully implemented a 3D dynamic FER on low-resolution videos using LBP-TOP descriptors. The local feature descriptors are extracted from small spatiotemporal cuboids in both texture and depth sequences. Then two codebooks are trained using locality constrained linear coding (LLC) and the feature descriptors of texture and depth sequences are converted into sparse codes. SPP is applied on the codes (concatenated texture and depth codes) to construct the feature vectors. This method results in comparable performance although low-resolution videos were considered [59].

An adaptation of well-known HOG descriptor named as histogram of orientated gradient-three orthogonal planes (HOG-TOP) has been proposed by Chen at al. for dynamic facial expression recognition by adding a temporal mode to conventional HOG [67]. This descriptor captures the dynamics of the texture to represent the properties of the face as its appearance changes during an expression. Motivated by LBP-TOP, HOG-TOP has been claimed to perform as effective as LBP-TOP in multimodal facial expression recognition [67] . It should be noted that although it has been claimed that HOG-TOP is a more compact representation compared to LBP-TOP, this feature descriptor has not been proved to individually function efficiently without fusion with other features. Fang and Zhao [70], [71] obtained the registered images using mesh matching between the shape models of two consecutive frames. Two methods of mesh matching namely, spin image and meshHOG are used to find correspondence vertices and then LBP-TOP features are extracted to represent static and dynamic characteristics. Radial basis function support vector machine (RBF-SVM) is applied for classification.

On the other hand, according to the characteristics of human facial expression, extensive studies have been conducted on landmark-based or curvature-based features. In [54], a vertex tracking approach is applied to adapt the 3D generic deformable models to all frames of the expression video sequence. The proposed system is complex and depends on manually marked landmarks**.** Dynamic range models constructed from 3D sequences have been employed in [18]. Nevertheless, both of the proposed systems relied on 83 annotated landmarks for the models.

A Fully automated approach has been suggested for 4D FER in [60] which basically uses geometric deformations. Radial curves have been extracted from the face mesh

to represent the characteristics of the 3D faces and expression-induced deformations have been quantified by Riemannian shape analysis (RSA). One limitation of this study is that it relies on accurate nose point detection which may fail for non-frontal view frames and in case of occlusion. Another one is that complex process is required to apply on high quality videos in order to achieve high performance. In [56], an integrated approach for dynamic FER which combines machine learning, parallel coordinates and human reasoning has been proposed. Facial points have been detected and tracked in video sequences. Features are extracted based on landmark movements and curvature changes (Gabor response) in AUs. Several machine learning approaches have been examined and acceptable results have been obtained.

A dynamic FER system have been suggested in [57] inspired by diffeomorphic motions of face muscles. Salient and common features among all candidates have been extracted for each expression and using training data, reference sequences named ‘atlas’ are constructed. While the performance of the system has been claimed to be superior to state-of-the-art, sensitivity to illumination variation is one of the drawbacks of this method. In addition, the system is computationally complicated since a dedicated atlas is to be constructed for each expression.

A fully-automated real-time approach for subject independent dynamic facial expression is proposed in [61] . The system relies on a set of automatically detected landmarks. A few landmarks are considered and then local descriptors are extracted in their neighborhood. In addition, geometric distances of landmarks are utilized as features. The feature extraction method in the proposed system can be assumed as an integrated one since it exploits landmark information as well as local feature

descriptors. Considering the relatively low complexity of the method and as it has been claimed its real time characteristics, the results are acceptable.

A fully landmark-based geometric feature extraction method has been used for dynamic facial expression recognition by [62] . This study utilizes a subset of 83 facial points provided by BU-4DFE data set. From the expression video sequence, the apex phase is detected and features are computed as the deviation of the pairwise distances between the facial points in the neutral frame and the apex frame. Although the accuracy based on 25 feature points is promising, this method ignores most of the transition information in the temporal domain by considering just neutral and apex frames.

In [63], a facial expression recognition system is proposed based on automatically detected landmarks. Several Landmarks were detected from each single frame and discrete cosine transform (3D-DCT) features are extracted. The reported accuracy of the study is limited compared to the state-of-the-art although too extent processing was applied for landmark detection, feature selection and classification. In [61], the pairwise distances between automatically detected 3D landmarks are fused with SIFT descriptors extracted in the neighborhood of the landmarks. Temporal HMM is utilized to model and recognized the expressions.

In [58], a joint dynamic facial expression recognition and expression intensity estimation is proposed. Geometric features which are basically the shape and the coordinates of facial landmarks are extracted using AAM. In temporal domain, only apex phase of expression has been considered. While it has been claimed that system

performs superior to other studies, again the dynamics of the expression were ignored.

In [72], free form deformation features are extracted from 3D sequences. Neighboring frames in onset and offset phases are considered for these motion-related features. The dynamics are modeled in the classification stage by HMM.

In addition, there are some attempts that rely on the key frame of the video sequence [66], [73], although identifying the corresponding frame is yet an issue and the temporal information in transitions is ignored. Yao et al. [66] suggest texture and geometric scattering features extracted from 2D and 3D key frames. For classification, multiple kernel learning (MKL) is applied on different combination of 2D and 3D operators. In Zhen2017 [73] , extract spatial facial deformation using Riemannian shape and amplified them by temporal filtering. Spectral clustering is applied to detect the key frame from the whole sequence. Both HMM and SVM classifiers are evaluated.

2.4 Feature Selection Methods in Facial Expression Recognition

In the field of human computer interaction (HCI), time and computational complexity are the factors limiting the real-time applications and implementation in electronic devices. For FER systems, specifically the dynamic facial expression recognition one of the main phases that carries a high computational and time complexity is the feature extraction. Moreover, with large feature matrices the algorithm used for classification adds burden to the system and it may degrade because of redundancy. Feature selection methods have been designed to decrease

the redundancy among the feature, select the most informative features and avoid the curse of dimensionality degrade the classifier algorithm.

It should be noted that both dimensionality reduction methods such as PCA and linear discriminant analysis (LDA) as well as attribute selection methods such as CFS and minimum-redundancy and maximum-relevance (mRMR) are referred as feature selection methods in this study. Both types are used to reduce the dimensionality of the original feature space, but PCA-like methods alter the structure of the original features and map the feature space into another space, while attribute selection methods preserve the properties of the original feature space by picking a subset out of it. Both approaches have been exploited in facial expression recognition.

On the other hand, the feature selection as a general machine learning and pattern recognition concept comprises a broad range of approaches and categories such as supervised versus unsupervised, iterative versus non-iterative, multivariate versus univariate, filters versus wrappers and embedded approaches. In fact, depending on the feature properties, the main issues classification, and limitation in time and complexity, one may decide on the feature selection method. However, the aim of this section is not to explore the detail of those approaches, but to briefly review the feature selection methods recently applied in facial expression recognition studies.

PCA is an unsupervised dimensionality reduction method while LDA is considered as a supervised simple classifier. These two methods have been widely used individually or in combination together in facial expression recognition studies. In the fully automated 4D FER approach suggested by [60] , LDA is applied to reduce

the dimensionality of the features extracted as geometric deformations. Kaur and Kaur [74] have exploited PCA for selecting the most informative Eigen expressive face of facial expression recognition. Their algorithm is based on speeded up robust features and K-nearest neighborhood classifier. PCA-based dimensionality reduction has also been utilized in a study by [75]. LBP feature descriptors are extracted from the image patches and PCA is applied to find the orthogonal basis vectors. After reducing the dimensionality of the feature space, Kohonen self-organizing map (KSOM) neural networks are applied for classification. Soyel et al. [13], have suggested an approach based on PCA and LDA to create the optimal feature projection subspace. The proposed system functions based on distance features and Fisher criterion for discarding redundant distances.

On the other hand, information theory-based feature selection approaches are among the most well-known approaches in feature selection. These approaches have been established using the entropy operator and include various methods such the ones based on simple entropy or the iterative mRMR method. A facial expression recognition system have been proposed by [21] using entropy-based feature selection method to select the informative distance features obtained from a set of facial key points. The entropy-based feature selection is applied for each expression and then followed by a two level SVM classifier. In fact, simple entropy-based feature ranking algorithms previously used in FER systems are unsupervised and they are criticized for adding redundant features to the subset.

Other approaches applied for feature selection in FER studies include multivariate filter methods such as CFS and optimization techniques such as genetic algorithm (GA). In [48] , facial landmark positions are detected by PDM and then geometric

features (empirical normalized distances) are extracted from the localized landmarks. CFS is applied to find a subset of informative, yet non-redundant features. The non-dominated sorting GA was also applied successfully in facial expression recognition to find the optimal feature subset [76]. One drawback of such feature selection approaches is the complexity level which makes it challenging for real-time facial expression systems.

Alternatively, by introduction of bag of words and pooling methods such as SPP in image processing and computer vision community, the concept of sparse coding has been initiated. Sparse coding is a new approach to covert the dense descriptor matrices into sparse matrices. Several algorithms has been proposed for sparse coding in image recognition including LLC [77], Laplacian sparse coding [78] and low-rank sparse coding [79]. Followed by a pooling mechanism, coding approaches have been successfully applied as an alternative to feature selection in image recognition and facial expression recognition. Specifically, in dynamic FER systems where the features are mainly spatiotemporal, these techniques can contribute to significant results.

In a study by [80] the proposed system of sparse representation using a PCA-based dictionary has been claimed to remarkably improve the performance of the facial expression recognition system compared to previous studies. Sparse Representation-based Classification was exploited in FER domain by [81] . The authors have suggested a system using the combination of Gabor texture features and local phase quantization descriptors. Adaboost method is then applied on Gabor texture features to select the useful ones. Sparse codes are then obtained on both types of features and classification is implemented on the fusion of the residuals. Ghimire et al. [16] have

proposed an automatic dynamic facial recognition system using a set of feature vectors computed as the displacement of the facial landmarks with respect to the first frame of the expression video sequence. The landmarks are detected and tracked via elastic bunch graph matching displacement estimation. For each expression, a prototype is constructed and multi-class AdaBoost with dynamic time warping (DTW) is used for feature selection. More precisely, the distance between a sample feature vector and the related expression prototype is used as a weak classifier to implement the supervised feature subset selection. Finally, SVM classifier is applied to recognize the expressions. Adaboost has been also used in [82] for selecting informative features among a set of Gabor features for facial expression recognition.

In summary, when deciding on the feature selection method for a FER or D-FER system, several aspects are to be considered. The complexity, computational and time burden are critical specifically in real-time and real-world applications. Since, maximal discriminative power and minimal redundancy among the features are important, supervised multivariate methods generally result in higher performance than unsupervised univariate methods. On the other hand, coding techniques have also been effectively applied in FER domain. In fact, we cannot jump into the conclusion that a specific approach fits the facial expression recognition the best. It depends on many factors in the system design including the characteristics of the extracted features, the data set, the problem and even the selected classifier.

2.5 Classification Methods in Facial Expression Recognition

Classification is the last process in any facial expression recognition system. This process generally includes a training phase with a learning algorithm and a test phase to evaluate the system performance. Similar to feature selection, classification is a

very comprehensive topic in pattern recognition and machine learning. Discussing the details of the classification approaches in accordance to their characteristics, strengths, and weaknesses is not in the scope of this thesis. Instead, we review some of the popular classification methods in facial expression recognition.

One of the commonly-used classifiers in FER studies is SVM. SVM is a binary classification approach with a learning algorithm that finds some support vectors from the train samples. The algorithm aims at constructing an optimal separating hyperplane with maximum distance between the closest samples of the two classes. SVM is a linear classifier in nature and thus using a kernel function, the nonlinear problems are mapped into a higher dimensional space with linear separability characteristics. In addition, as stated before, SVM is a binary classifier applicable in two-class problems. FER problems are all multi-class though and basically a collection of one-versus-all or one-versus-one linear SVMs are constructed by the scholars in facial expression recognition community.

A comprehensive study on LBP feature descriptors in FER systems have been conducted by [40]. It has been claimed that the best performing features are the boosted-LBP and as a classifier SVM outperforms LDA and linear programming. In geometric-based approaches for FER, SVMs are very popular [14], [44], [48]. In [48], RBF-SVM classification is applied for the proposed geometric landmark-based FER system. However, nonlinear kernel SVM such as RBF are costly from computational and learning time perspectives. The two-layered SVM structure suggested by [21] for 3D geometric-based FER performs effectively. Another SVM-based FER system using the length and slopes of line segments connecting facial points is proposed by [83] which has resulted in significant recognition accuracy.

SVM classifier is also applied in a study by [82] to recognize the facial expressions. Canavan et al. [84] has designed a dynamic facial activity analysis system by automatic landmark tracking, curvature features and SVM classifier. SVM classifier is also effectively used for posed and spontaneous 4D facial behavior analysis [85] . In Kumar2016 [62] , SVM is used in a dynamic 3D FER system based on Euclidian distance landmark-based features.

Hidden Markov Model is another widely-used classifier in facial expression and emotion analysis studies. The state-structure of HMMs makes them suitable for modeling the dynamics of the expression. In [54], based on 3D generic deformable models, a vertex tracking approach is applied to adapt the model to all frames of the sequence. Spatial and temporal features are then classified by HMM. Sandbatch et al. [86] have applied Gentleboost and HMM to distinguish expressions from video sequences containing onset, apex and offset phases.

A Fully automated approach has been suggested for 4D FER in [60] which basically uses geometric deformations. Radial curves have been extracted to represent 3D faces and expression-induced deformations have been quantified by RSA. In order to reduce the dimensionality of extracted features, LDA has been employed. Temporal HMM and random forest classifiers are applied to extract 3D motion and mean deformation respectively.

A fully-automated real-time approach for subject independent dynamic facial expression detection (D-FED) is proposed in [61] . A few landmarks are detected and local descriptors are extracted in their neighborhood. In addition, geometric distances of landmarks are utilized as features. In classification stage, a four state

HMM is used to model each of the four phases of the sequences (neutral, onset, apex and offset). More recently, a joint dynamic FER and expression intensity estimation have been successfully examined [58]. Geometric features which are basically the shape and coordinates of facial landmarks are extracted using AAM. In temporal domain, only apex phase of expression has been considered. For classification, HMMs have been employed.

Several other classifiers have been evaluated in FER and D-FER systems. In a comprehensive study by [56] , six different classifiers have been evaluated in D-FER including decision trees (J48), sequential minimum optimization for SVM , random forests, fuzzy rough set nearest neighborhood (FRNN), logistic regression, and vaguely quantified nearest neighborhood .The proposed integrated approach combines machine learning, parallel coordinates and human reasoning. Facial points have been detected and tracked in video sequences. Features are extracted based on landmark movements and curvature changes (Gabor response) in AUs. They have shown that FRNN and SVM outperform the other classifiers resulting in acceptable average performance for recognizing six basic expressions.

Shao et al. [59] have implemented a 3D dynamic FER on low-resolution videos. LBP-TOP features are extracted from small cuboids in texture and depth sequences. Then two codebooks are trained using LLC and feature descriptors are converted into sparse codes. SPP is applied on feature descriptors (combination of texture and depth features) to extract feature vectors. Finally, CRF are used for classification. Experimental results confirmed that the proposed approach provides comparable results although it used low-resolution videos as input.

LDA is also among the classification methods extensively studied by the researchers in facial expression studies [87]–[89]. Rosato et al. [87] have applied SVM on a set of generalized manifold and texture features extracted from automatically tracked landmarks for 3D facial expression recognition. In [89] , a PCA-LDA-based system is proposed for 3D facial expression recognition using tracked facial actions.

In addition, artificial neural networks have been applied for classification in facial expression recognition and emotion analysis. For instance, the system presented in by [90] is an automatic facial expression recognition system based on self-organizing feature maps . After face detection, the pupils are localized to correct the head rotation. SOM is used for feature extraction from the cropped and rotated faces. Finally, a multi-layer perceptron neural network (MLPNN) is adopted to classify neutral and six prototypic expressions.

Deep learning is another emerging trend in image processing and recognition. Specifically, convolutional neural networks (CNNs) and long-short-term memory network (LSTM) have been successfully applied in facial expression recognition [91]–[93]. Feasibility of these approaches relies on advances in GPU technologies made in recent years as well as availability of huge amount of data to train networks. Although these methods are successful in extracting useful information from loads of data, when the amount of data is limited, deep neural networks suffer from over-fitting [94]. In image and video processing, data collection is costly and thus the conventional methods are to be manipulated for enhanced performance. This limitation has been recently spotted by some researchers [92], [94]. It has been argued that although CNNs generally achieve high recognition rate with big data,

publicly available datasets for facial expression recognition do not contain sufficient data for deep architectures [92].

Regardless of these limitations, deep learning approaches have been exploited by researchers in both static and dynamic FER systems. Liang et al. [95] have proposed a BiLSTM-based dynamic facial expression recognition based on two networks to model spatial and temporal information separately and then fused the context. A CNN-LSTM system is recently presented in [96] which adaptively initializes CNN and LSTM for improved performance. Jung et al. [94] have integrated two deep learning models which extract temporal appearance and temporal geometry from video sequences and facial landmarks respectively.

2.6 Applications of Automatic Facial Expression Recognition

In human-machine interactions, when computers are to analyze the feedbacks and entries from the user, emotions cannot be ignored. FER is one of the emerging topics of interest in computer vision and image processing. There are a wide range of applications for automatic facial expression recognition including HCI, medical care, psychology, marketing, customer service, education and gaming.

The computers in HCI need to take the actions and provide the responses based on the feedbacks received from the user. While keyboard, mouse and touchscreens are recognized as the conventional mediums for converting user reactions into understandable signals for the machine, advances in technology results in more complicated ways of interactions such as voice and image. Facial expression recognition systems make it possible for the machine to present more precise and effective actions in HCI systems [97].

In addition, the introduction of automated assistive medical systems has encouraged the researchers to focus more on the detailed data collection from the patients. Emotional state of a patient is one of the valuable information. Emotion recognition can also aid the physiologists in diagnoses based on the subject's expressive behavior. In addition, facial expression recognition can provide further information for emotion studies that analyze human emotional behavior.

In marketing and customer service industry, the responses of the targets to advertisement, shop design, and product packaging can be explored by automatic emotional detection. Similarly, in computer-based education systems, the learner's engagement and attitude towards the concept and the lecturer can be investigated by automatic emotion analysis system. The policies, strategies and methodologies can be modified according to the emotional feedbacks collected in the system both in marketing and education industry.

Another application of automatic facial expression recognition is in the gaming [98]. Nowadays, multiplayer online games are very popular because they provide the players with the chance to interact and collaborate with other players. Facial expression recognition can replace text commands in these games and give the players a more real feeling. For example, the emotions of avatars can be controlled by such a system instead of conventional prompts [98].

Chapter 3

GEOMETRIC LANDMARK-BASED 3D FER

3.1 Introduction

The aim of this chapter is to investigate the landmark-based approach in 3D facial expression recognition. This chapter is considered as a preliminary step to study and understand the potential challenges in spatial domain before getting involved with temporal information in the following chapters. Conventional geometric distance features are extracted and the FER problem is addressed by proposing advanced feature selection and classification methods. In fact, previous studies in FER have been either based on simple univariate feature selection without considering feature interactions or using dimensionality reduction such as PCA which does not take into account the class discriminability of the subspaces. The first contribution of this part of the study is application of SFFS to find an optimum low-dimensional feature subspace for person-independent three-dimensional facial expression detection (3D FED). The second contribution is in classification phase where a novel two-layered SVM-FSVM is designed. The rest of this chapter is framed into four sections. In the Section 3.2, BU-3DFE data set is reviewed. In Section 3.3, proposed method is described. Experimental results are demonstrated in Section 3.4 and discussed in Section 4.5. The last section, Section 4.6 concludes the chapter. The outline of the proposed system and the system used as the reference for comparison are illustrated in Fig. 3.1.

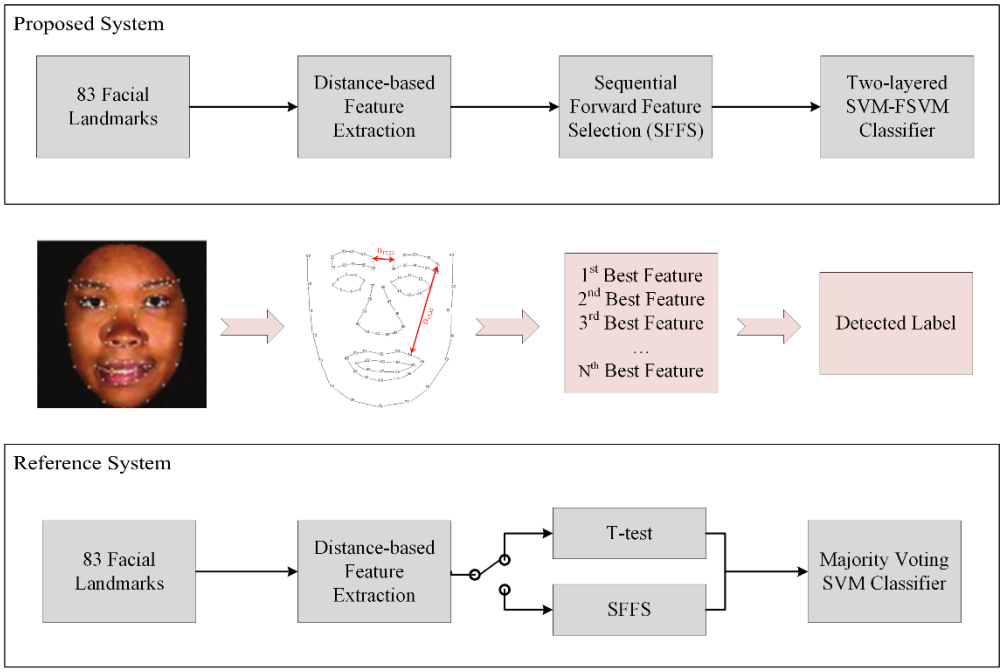


Figure 3.1: Overview of the proposed and the reference systems.

3.2 BU-3DFE Data Set

The data set used in the first phase of this study is 3D FED data set of Binghamton University known as BU-3DFE [99] . This data set is published in 2006 and it is one of the popular data sets in emotion analysis from 3D images. BU-3DFE images are recorded from 100 subjects including 44 male and 56 female subjects. Data set contains 25 texture and depth images for each of the subject. It comprises one neutral expression image and four different level intensities of emotion expression images for each of the 6 basic expressions namely anger, disgust, fear, happiness, sadness, and surprise. Data set also provides 3D coordinates of 83 points of the face model which are utilized for distance-based feature extraction. Four sample subjects are selected and illustrated in Fig. 3.2

.

Figure 3.2: Four sample subjects from BU-3DFE. First row: texture images and 83 landmarks of face model. Rows 3, 5 and 7: texture images, rows 2, 4, 6 and 8: depth images of different expressions in each column: (a) neutral (b) anger, (c) disgust, (d) fear, (e) happy, (f) sadness, and (g) surprise.

The pictures were captured at the apex phase of the expression. Provided landmark points are annotated on the faces as well. Fig. 3.2 (a) represents a typical neutral face model. Fig. 3.2 (b) to (g) show the six prototypic expressions as anger, disgust, fear,

happy, sadness and surprise respectively. Each emotion is expressed at four different intensity levels. Fig. 3.3 shows several sample subject expressing six emotions at four intensity levels. The left most image corresponds to the lowest level intensity while the right most one represents the highest level intensity. In this study, expressions of level four (the highest intensity) are considered.

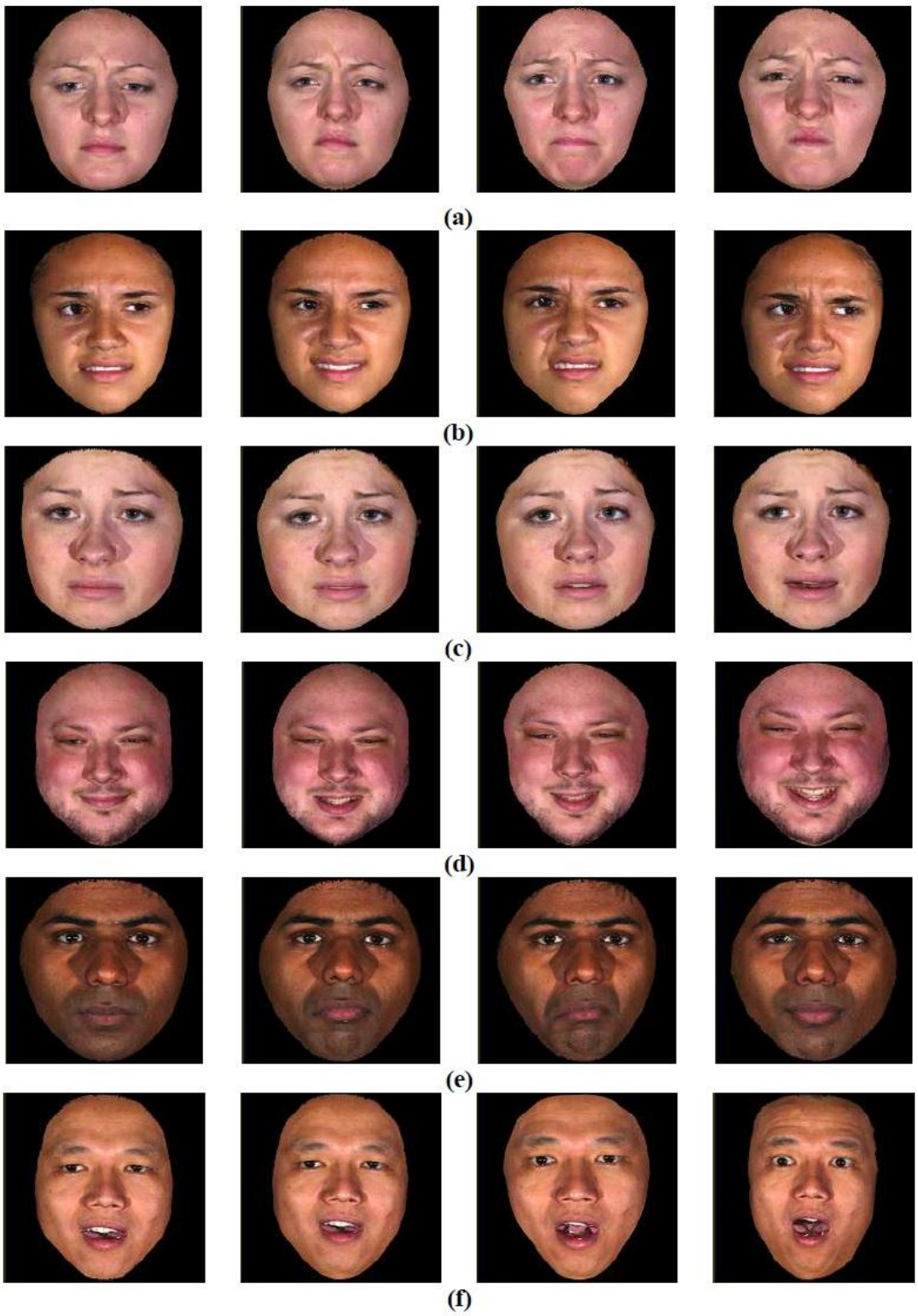


Figure 3.3: A sample subject from BU-3DFE with four intensity levels of expression (a) anger, (b) disgust, (c) fear, (d) happy, (e) sadness, and (f) surprise.

3.3 Proposed Method

As stated in section 2.2, geometric features have been utilized by scholars in facial expression recognition for decades. Distance-based geometric features which measure the pairwise distances of facial landmarks have been proved to perform efficiently in 2D and 3D FER systems. However, the large number of features induces the necessity to exploit a feature selection method. Here, SFFS is applied to select a subset of useful features. The proposed method has three phases: feature extraction, feature selection and classification.

3.3.1 Feature Extraction

Features are extracted as pairwise distances of 83 points of the face model in 3D space. Distances between all pairs of these 3D landmarks are calculated using Euclidean distance as shown in Eq. 3.1.

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{3.1}$$

where (x_i, y_i, z_i) and (x_j, y_j, z_j) are the 3D Cartesian coordinates of landmarks i and j respectively and d_{ij} is the distance feature. All distances are normalized by the distance between the inner corners of eyes to compensate the scale variations. As all combinations of the pairs of points are to be taken into account, the total number of features is equal to $\binom{83}{2} = 3403$. In fact, this is completely impractical to use all these features for classification problem. In addition, the results will be definitely poor because of redundancy. Hence, feature selection is applied as an initial stage on the train set.

3.3.2 Sequential Forward Feature Selection

SFFS is used as feature selection algorithms for each of the two-class classifiers separately. Conventional t-test is also applied as another feature selection method for comparison. In SFFS, algorithm starts from an empty set and adds features

sequentially to the set in order to minimize an error criterion. There is an internal 5-fold cross validation (5-CV) in which train data is divided into 5 partitions and based on the criterion calculated on internal test set, features are added to the selected subset in each iteration. Algorithm stops when either a predefined number of features are selected or the criterion stops improving. In this study, classification error of Naive Bayes classifier is used in SFFS as the optimization function to be minimized. SFFS is known as a multivariate feature selection method which means selected features are not highly correlated. In conventional univariate methods such as entropy-based, F-score and t-test, top features are mainly highly correlated and there is a huge amount of redundancy among features. Considering the displacement of the landmarks and their relative distances, distance-based geometric features are generally correlated in their nature. Using SFFS, a small subset of selected features can provide acceptable accuracy if an appropriate classifier is utilized. The algorithm of SFFS is as follows:

Sequential Forward Feature Selection Algorithm
1. Start with the empty subset $F = \emptyset$
2. Select the next best feature $f = arg\ min(E_r(x \cup F)), \quad x \in \{X-F\}$
3. Update feature subset $F = F \cup f$
4. If stop criterion is not satisfied, go to step 2

where X is the whole set of features and Er is the classification error of Naive Bayes classifier as follows. Assume a Naive Bayesian classifier which classifies sample

vectors *s* into classes C_is (*i=1,2,…6*). The Bayes error rate is the probability of incorrect prediction of the class label and it defines as follows.

$$E_r = \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \int p(C_i|s)p(s)_{s\in H_j} \tag{3.2}$$

where *p(.)* is the probability function, *M* is the number of classes and H_j is the area that classifier predicts the label of *s* incorrectly.

As mentioned before, a traditional supervised univariate feature selection namely student t-test is also implemented for comparison. In t-test, the score for feature i is calculated as in Eq. 3.3.

$$ttest_{Score_i} = \frac{|m_0 - m_1|}{\sqrt{\frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1}}} \tag{3.3}$$

where m_k, σ_k and N_k ($k = 0,1$) are the mean value, standard deviation and number of samples belong to class k respectively. The larger the score is, the more discriminative the feature is. This method ranks the features from the most discriminative to the least discriminative one and unlike SFFS it does not consider their interactions and redundancies.

3.3.3 Fuzzy-SVM Classification

Since the problem of facial expression recognition is a multi-class problem, a mechanism is required to convert predictions of two-class SVMs into multi-class labels of 6 expressions. Conventionally, multi-class problems are divided into several two-class classifiers based on either ‘one versus one’ or ‘one versus all’ scheme followed by a majority voting to obtain the predicted label. However, the problem of unclassifiable regions limits the accuracy of this method. In this study, both ‘one versus one’ and ‘one versus all’ schemes are used.

In fact, the classifier is designed in two layers and in the first and second level of which, one versus one and one versus all schemes are performed, respectively. Since there are 6 expressions in the data set, the number of all possible binary classifiers in level one is equal to 15 while in level two there are 6 classifiers. In the second layer conventional SVM is replaced by FSVM. FSVM has been introduced to pattern recognition community in recent decade [100]. Recently, its application in multi-class classification has been attracted the interest of many researchers specifically for small samples [101] . The mechanism of FSVM approach is described in the following.

The main idea of FSVM is to replace crisp classification into fuzzy classification. In other words, for a binary classification task, when a sample belongs to a class it cannot be a member of the other class, but in fuzzy classification a sample can belong to both classes with different membership values. Now, consider a set of labeled samples and their fuzzy membership values (s_1, y_1, m_1) , … , (s_k, y_k, m_k), (s_N, y_N, m_N).

where each d-dimensional sample s_k ($s_k \epsilon R^d$) is labeled as $y_k \epsilon (+1, -1)$ and its associated fuzzy membership is $\sigma \leq m_k \leq 1$ with a sufficiently small variable σ. Now, consider a mapping φ from R^d to the feature space Z so as $z = \varphi(s)$. As in conventional SVM where the target is to find the discriminating hyperplane $w.z+b=0$, we need to define an objective function but with taking into account the fuzzy membership. Note that the associated fuzzy membership of a sample defines its behavior towards the class labels. As a results when in SVM, the parameter ξ is error measure, in FSVM $m.\xi$ is the error. Hence, the optimal hyperplane can be identified by solving the following optimization problem.

$$\min \frac{1}{2} w.w + C \sum_{k=1}^{N} m_k \xi_k \tag{3.4}$$

Subject to

$$y_k(w.z_k + b) \geq 1 - \xi_k, \qquad i = 1, \dots, N, \qquad \xi_k \geq 0 \tag{3.5}$$

where *N* is the number of samples in training set and C is a constant value defined practically by the user. In order to solve the optimization problem, the Lagrangian *L* with parameters α and β is constructed as:

$$L(w, b, \xi, \alpha, \beta) = \min \frac{1}{2} w.w + C \sum_{k=1}^{N} m_k \xi_k - \sum_{k=1}^{N} \alpha_k (y_k(w.z_k + b) - 1 + \xi_k) - \sum_{k=1}^{N} \beta_k \xi_k \tag{3.6}$$

for the parameters *w*, *b* and ξ_k the conditions given in Eq. 3.7 must be satisfied.

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial w} = 0, \ \frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial b} = 0, \ \frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial \xi_k} = 0 \tag{3.7}$$

By calculating the related terms based on the above conditions, the Lagrangian in Eq.3.6 is altered to:

$$\min w(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{k=1}^{N} \sum_{n=1}^{N} \alpha_k \alpha_n y_k y_n K(s_k s_n) \tag{3.8}$$

Subject to:

$$\sum_{k=1}^{N} y_k \alpha_k = 0, \qquad 0 \leq \alpha_k \leq m_k C, \ k = 1, \dots, N \tag{3.9}$$

The fuzzy membership function is simply selected by first defining the overlap value and then finding the appropriate function that matches the main characteristic of the data. In this study, we have a balanced multiclass problem with equal number of samples per class. Hence, a simple rectangular fuzzy membership function with 50% overlap is used.

3.4 Experimental Results

For splitting data into test and train set, 10-fold cross validation (10-CV) is applied. All the experiments are repeated 10 times by randomly selecting 90% (90) of samples as the train set and 10% (10) as the test set. The task is subject-independent. In each fold, feature selection and classifier training are performed on train data and then evaluated using test data. Accuracies are reported as average over ten folds. The experiments are set as follows. Firstly, system based on t-test feature selection and conventional majority voting SVM is implemented as a reference one. Secondly, t-test is replaced by SFFS and again majority voting SVM is applied in order to provide a comparative perspective over t-test and SFFS. Finally, the proposed system using SFFS and SVM-FSVM is implemented. The architecture of the system established using majority voting SVM is shown in Fig. 3.4.

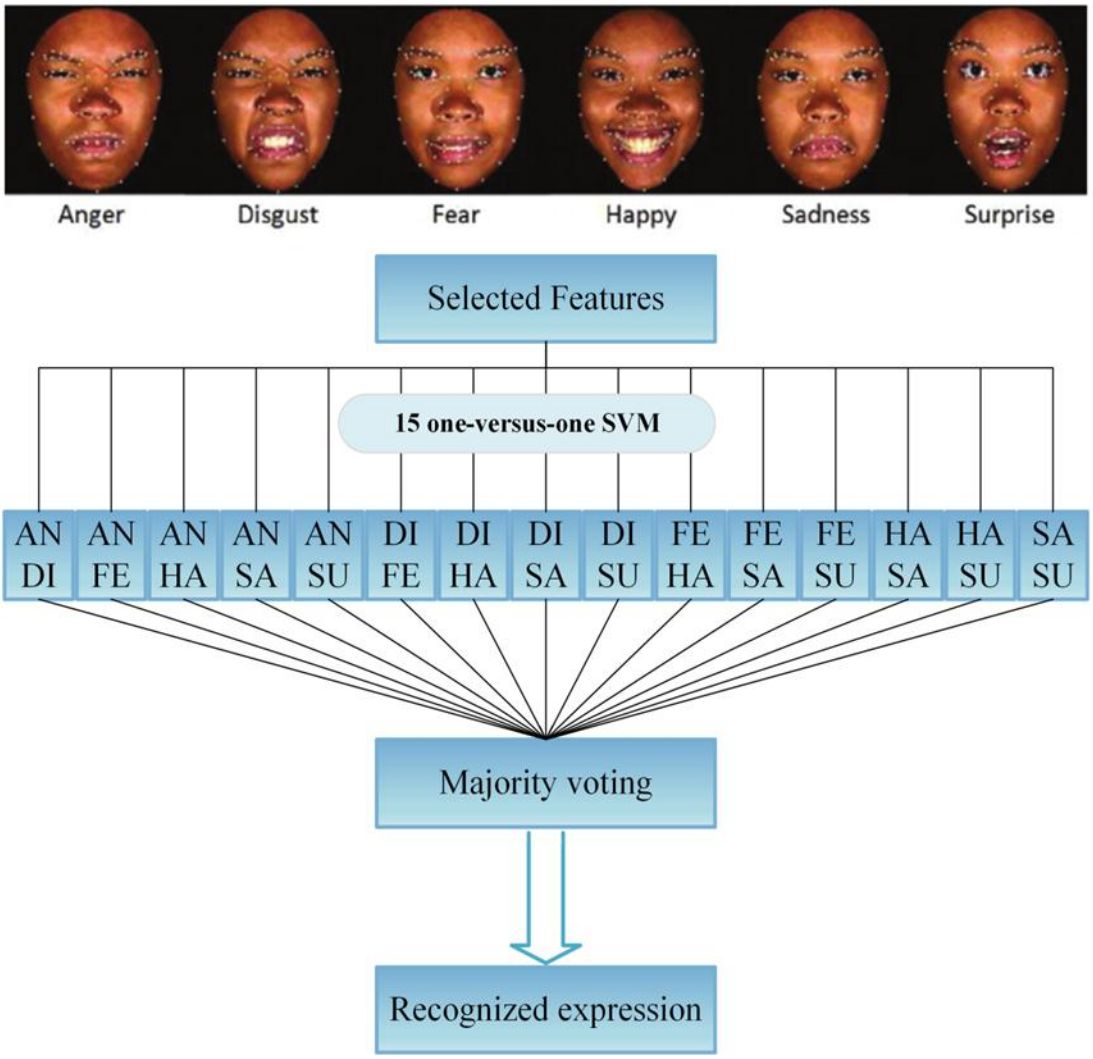


Figure 3.4: The architecture of the reference system (majority voting SVM).

Because the aim of the study is to find an optimum feature subspace, some initial evaluations have performed to find the optimum cardinality of feature set subsets for both t-test and SFFS. Based on average accuracies, 27 features provide maximal performance on t-test and adding more features does not improve average accuracy. In SFFS implementation, the size of the feature subset is identified by feature selection method. In other words, the iterative procedure is repeated until the decay in the error becomes negligible. The evaluations showed that maximum dimension of the best subsets is at the most 18. The reason for this difference is that the level of

redundancy among the features selected by t-test is high and this degrades the classification accuracy. Unlike t-test, SFFS functions as a multivariate feature selection that attains subsets of useful, yet on-redundant features. Hence, the accuracy of the system keeps improving for a larger number of features. In the reference systems based on conventional SVM, a one-layered majority voting classification system including 15 two-class SVM classifiers is constructed. Table 3.1 shows the average confusion matrix.

Table 3.1: Confusion matrix (t-test and majority voting SVM)

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**66.00**	3.00	4.00	1.00	0.00	26.00
Disgust	10.00	**69.00**	11.00	2.00	7.00	1.00
Fear	5.00	9.00	**62.00**	12.00	4.00	8.00
Happy	1.00	10.00	6.00	**82.00**	1.00	0.00
Sadness	0.00	1.00	9.00	2.00	**88.00**	0.00
Surprise	23.00	2.00	8.00	0.00	0.00	**67.00**
Overall	**72.33**					

In the next step, t-test is replaced by SFFS approach. As mentioned before, in SFFS, the iterative search algorithm finds a subset that minimized Bayesian error by taking into account feature interactions. Therefore, feature subset has more discriminate potential and the feature subspaces consist of less correlated attributes compared to the subspaces selected by t-test. This is also true for other conventional selection methods like entropy which does not consider feature interactions in selection procedure.

This property is illustrated in Fig. 3.5. Figure represents a visual perspective of class discriminability and feature correlation for two different cases. Two-dimensional feature subspaces of the best features for classification of fear vs. happiness and surprise vs. sadness are selected as examples. It is clear in the figure that classes are more discriminable when features are selected by SFFS.

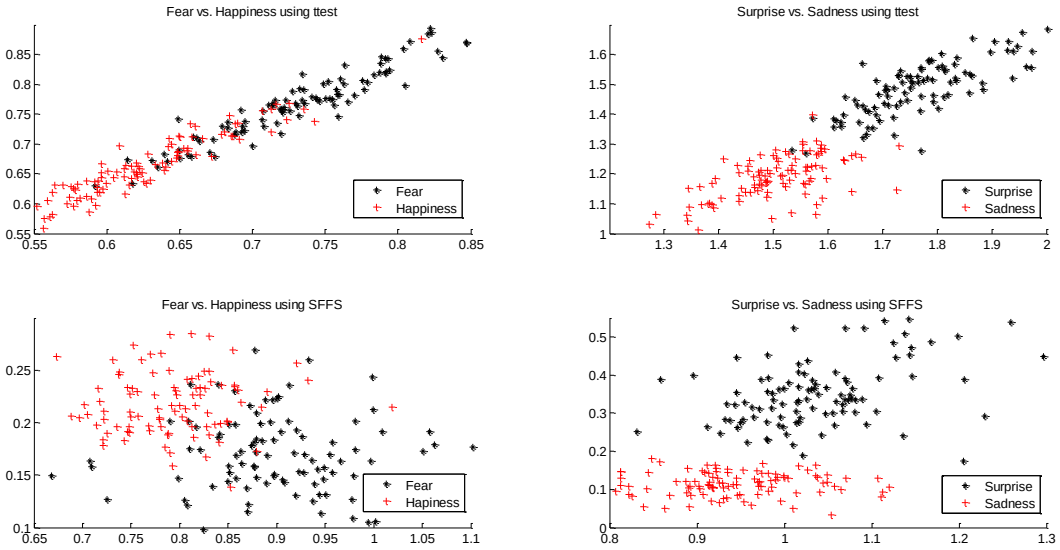


Figure 3.5: Example scatter plots in two dimensional optimal feature subspace (top: t-test, bottom: SFFS).

Another advantage of SFFS method is that there is no need to estimate the dimensionality of the optimum subset since the error reaches a plateau after a limited number of iterations. This procedure in one of the 10 folds is illustrated in Fig. 3.6. The curve of the changes in Bayesian error rate for classification of anger versus disgust, and fear versus surprise are plotted. It can be observed that dimensions best subsets are 15 and 7 respectively.

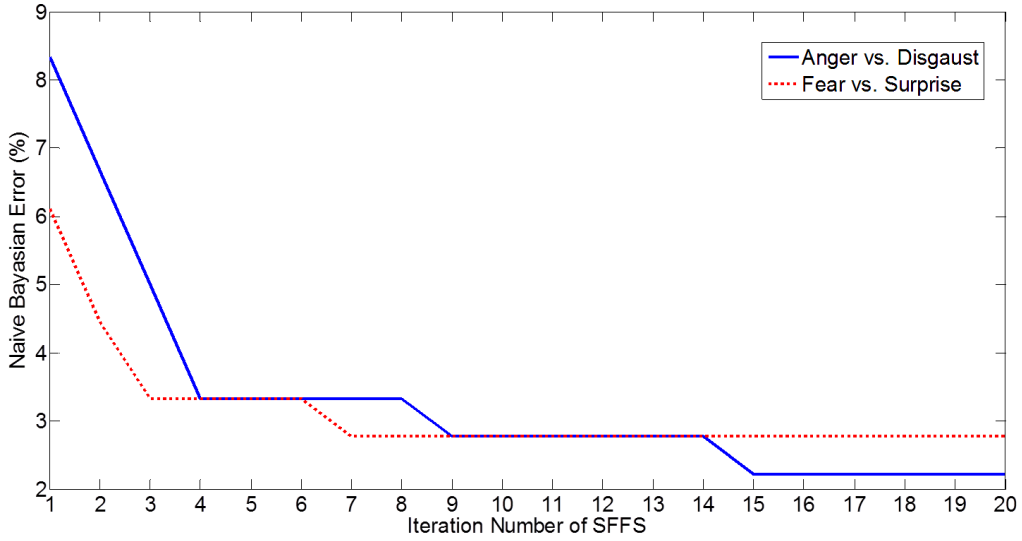


Figure 3.6: Two example error curves of SFFS feature selection procedure in one fold (dimensionality of the feature subsets is 15 and 7).

According to experimental results, dimension of the best subsets selected by SFFS ranges from 2 to 18 for each of the two-class classifiers in all folds. Average dimensionality of the feature subsets in 10-fold is reported in Table 3.2 for all 15 classification cases. In Table 3.3, experimental results obtained by these low-dimensional feature subsets and majority voting scheme is shown. It is notable that average accuracy has been improved by 6% compared to t-test feature selection. However, the accuracy rates of this experiments are poor compared to the state-of-the-art. A general perspective has been provided toward the capability of SFFS in selecting low-dimensional feature subset.

Table 3.2: Average dimension of feature subsets

AN/DI	**AN/FE**	**AN/HA**	**AN/SU**	**AN/SA**	**DI/FE**	**DI/HA**	**DI/SU**
15.7	14.9	4.4	14.6	17.4	16.7	13.8	8.0
DI/SA	**FE/HA**	**FE/SU**	**FE/SA**	**HA/SU**	**HA/SA**	**SU/SA**	**AVG.**
9.7	15.3	9.8	13.1	15.5	2.5	3.2	**11.5**

Table 3.3: Confusion matrix (SFFS and majority voting SVM)

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**72.00**	7.00	2.00	1.00	0.00	18.00
Disgust	10.00	**76.00**	8.00	1.00	5.00	0.00
Fear	2.00	10.00	**70.00**	9.00	2.00	7.00
Happy	1.00	1.00	12.00	**85.00**	1.00	0.00
Sadness	0.00	1.00	5.00	2.00	**92.00**	0.00
Surprise	15.00	5.00	3.00	2.00	0.00	**75.00**
Overall	**78.33**					

The proposed system for 3D facial expression recognition is based on distance features, SFFS and a two-layered SVM-FSVM classifier to achieve acceptable performance. More precisely, the first layer of the system consists of 15 one-versus-one SVMs each trained on a different subset of features selected by SFFS for that specific problem. After finding the optimal hyperplane, associated decision values of SVM are used as the input to the second layer. These decision values represent the attitudes of the train samples toward the optimal separating hyperplane for each expression versus any other expressions. In the second layer of the propose classification system, there are FSVM classifiers. The 15 decision values are fed to the second layer of 6 one-versus-all FSVM classifiers. In other words, input of the second stage classifier is $w.\varphi(x)+b$ values explained in section 3.3.3. In the second level, 6 FSVMs are trained for 6 expressions. Train labels are defined in a way that for each expression, output of the related FSVM is maximum (1) and output of the others are minimum (-1). Train label matrix for each of the 6 expressions is as follows. Fig. 3.7 provides a perspective of this system.

$$\begin{bmatrix} AN \\ DI \\ FE \\ HA \\ SU \\ SA \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

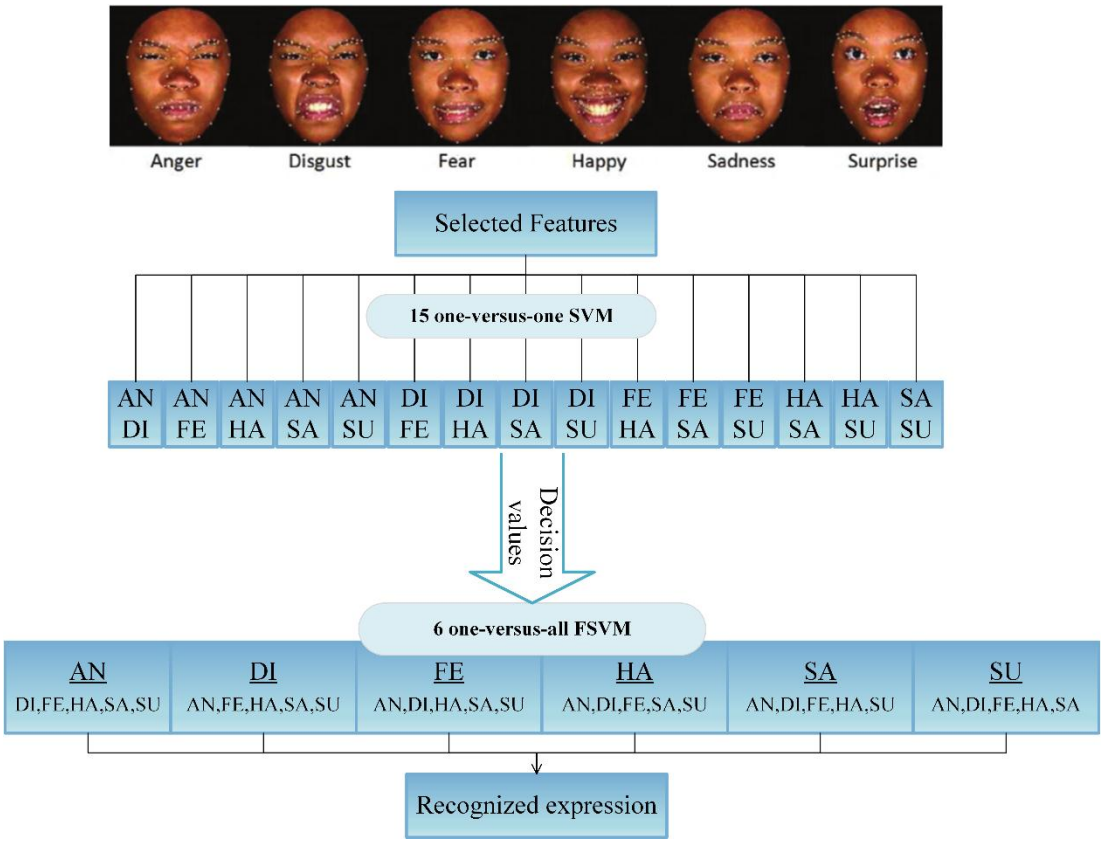


Figure 3.7: The architecture of the proposed system (SVM-FSVM).

After these 6 FSVMs are trained, accuracy of the system is estimated by test samples. In order to predict the label of a test vector, maximum argument is utilized. More precisely, the maximal argument of the fuzzy membership value, i.e. the *agrmax(m_k)* is the predicted expression. Table 3.4 shows the experimental results of proposed method in terms of confusion matrix. Average multi-class accuracy is improved significantly by applying the proposed two stage FSVM classifier. In order to highlight the effectiveness of SFFS method, the tests are also conducted on the whole set of features. Table 3.5 represents the recognition and confusion rates.

Average recognition accuracy of proposed SVM-FSVM classifier without applying feature selection is 81.83%. By comparing the results given in Table 3.4 and Table 3.5, it is clear that SFFS improves average accuracy by almost 6%. In fact, since SVM-based classifiers are vulnerable to redundancy, removing redundant features results in remarkable improvements.

Table 3.4: Confusion matrix (SFFS and proposed FSVM)

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**85.00**	3.00	3.00	1.00	0.00	8.00
Disgust	7.00	**85.00**	4.00	0.00	4.00	0.00
Fear	2.00	6.00	**82.00**	3.00	0.00	7.00
Happy	1.00	0.00	4.00	**94.00**	1.00	0.00
Sadness	0.00	1.00	3.00	1.00	**95.00**	0.00
Surprise	5.00	5.00	4.00	1.00	0.00	**85.00**
Overall	**87.67**					

Table 3.5: Confusion matrix (All features and proposed FSVM)

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**80.00**	5.00	5.00	1.00	0.00	9.00
Disgust	10.00	**78.00**	6.00	1.00	5.00	0.00
Fear	3.00	8.00	**78.00**	3.00	1.00	7.00
Happy	2.00	2.00	6.00	**87.00**	1.00	2.00
Sadness	1.00	1.00	6.00	3.00	**88.00**	1.00
Surprise	6.00	5.00	6.00	1.00	2.00	**80.00**
Overall	**81.83**					

3.5 Discussion

There are two mainstreams in FER research regardless of the details of the recognition system namely, geometric-based and appearance-based schemes. The most commonly addressed geometric approach is the landmark-based approach known well for its simplicity and robustness to illumination noise and acquisition artifacts. In this preliminary phase of the study, a novel 3D FER system is designed to recognize six prototypic expressions from still images. The experiments are structured in such a way that supports the claim of "using an efficient feature selection method and a well-designed classifier at the same time, landmark-based features may be successfully used in facial expression recognition".

Firstly, conventional t-test feature selection is compared with SFFS in feature selection stage. According to the experimental results, cardinality of subsets ranges from 2.5 to 17.4 on average for 10 folds. Total average is 11.5 which means that the selected feature subspaces are low-dimension. Using one layer of conventional majority voting SVM, SFFS improves the result of t-test from 72.33% to 78.33%. Then by applying the proposed two-layered SVM-FSVM classifier, average accuracy reaches to 87.67%. Considering that the work is performed in low-dimensional feature subspaces, this accuracy is comparable to previous studies as shown in Table 3.6. Table 3.6 confirms that proposed FER system achieves higher recognition accuracies in comparison to [35], [102] and [20] and [103]. Soyel et al. [76] have obtained a slightly better recognition rate by using an evolutionary algorithm for feature selection. These algorithms need tuning for the parameters and are computationally demanding. Lopes et al. [92] have achieved 90.96% accuracy by using a preprocessing approach and CNN. However, considering the time and

computational burden of deep-learning approaches this improvement is not unexpected.

Table 3.6: Comparison of proposed system with state-of-the-art

Author	Method	ACC%
Wang et al. [35]	Primitive Surface Feature Distribution + LDA	83.60
Tang et al. [102]	Geometric Distance and Slope Features + Majority voting SVM	87.10
Zarbakhsh et al. [20]	Geometric Distance Features +T-Test +Majority voting SVM	74.63
Oyedotun et al. [103]	RGB and depth map + DCNN	87.05
Lopes et al. [92]	Pre-processing + CNN	90.96
Soyel et al. [76]	Geometric Distance Features +NSGAII + PNN	88.30
Proposed Method	**Geometric Distance Features + SFFS + SVM-FSVM**	**87.67**

3.6 Conclusion

The aim of this phase of the work is implementing efficient facial expression detection (FED) system in an optimal feature subspace. An iterative feature subset selection algorithm based on SFFS is proposed which selects feature subspaces for 15 two-class (one-versus-one) classifiers individually. A two-layered SVM-FSVM is then designed for classification. Experiments conducted on BU-3DFE data set have proved that proposed method obtains acceptable results. The result of this work is worthwhile for low-complexity practical applications of FED. In addition, these findings provide the motivation for the second stage as by applying appropriate and efficient feature selection and classification methods, geometric landmark-based features are potentially valuable in facial expression recognition.

Chapter 4

GEOMETRIC LANDMARK-BASED DEFORMATIONS IN 4D FER

4.1 Introduction

Four dimentsional facial expression recognition (4D FER) systems are known as dynamic 3D facial expression recognition systems. Basically, emotions are expressed in the face during a time interval with different phases namely neutral, onset, apex and offset [104]. During these phases, the movements of facial muscles transform the spatial charm acteristics of some specific regions of the face known as AUs. Static FER systems which analyze texture and depth images are designed to process the spatial information of the facial images recorded at the peak of an expression known as apex. On the other hand, since the input to the dynamic FER system is the sequence of the facial expression frames, these systems mainly process the spatiotemporal information. For simplicity, we term the dynamic 3D facial expression recognition as D-FER.

As a matter of fact, the dynamics in the time interval during which an emotion is expressed are very informative. Transitions of feature descriptors and geometric deformations taken place in this interval contain critical temporal information for facial expression recognition [57], [60], [105] .In D-FER systems, texture, depth and even landmark coordinates are recorded as video/time sequences during all phases of

the emotion expression. However, capturing temporal information is a challenging task and several studies address this by identifying a key frame [66], [73] or by analysis the subsequences individually [54], [86] . In this work, a time series-based method is proposed which processes the full sequences of facial expression recognition. Multimodal time series are constructed by applying a temporal sliding window to capture the dynamics as mean geometric deformation of facial key points. In this work, we adapt a time series analysis method to construct multimodal time series from landmark based deformations extracted from all of the frames in videos sequences. The problem is addressed then by classification approaches adapted for time series. This study contributes to the literature in three aspects. Firstly, unlike conventional geometric schemes that rely on a limited set of landmark-based features [14] a comprehensive set of deformations including point, distance and angle are extracted from 3D coordinates of facial landmarks to acquire displacements in all directions. Secondly, the features in original high dimensional feature space are filtered out by an effective feature subset selection algorithm named neighborhood component feature selection [106]. Thirdly, the notion of multimodal time series analysis is adapted to process full sequences of facial expression recognition and consequently, AC-DTW classifier is used for the first time in D-FER systems.

The rest of this chapter is organized as follows. The 4D data set used in this thesis is described in Section 4.2. In Section 4.3, the proposed algorithm and its different phases are outlined. Section 4.4 represents the experimental results conducted on BU-4DFE dynamic 3D facial expression data set. The discussion of the results is given in Section 4.5. Finally, in Section 4.6, the chapter is concluded.

4.2 BU-4DFE Data Set

In order to evaluate the performance of the proposed D-FER system, a set of experiments are conducted on BU-4DFE [55] , a well-known dynamic 3D facial expression recognition data set. This data set is collected from 101 subjects including 58 female and 43 male subjects. The documents of the recorded data are grouped as females and males and for each subject the expression data is recorded in six different expressive conditions. For each expression, geometric model and texture colored images are provided as separate sequences presented as individual frames. Resolutions of the depth and texture video sequences are 35,000 vertices and 1040×1329 pixels per frame respectively. It should be noted that we term geometric models as 'depth images' in this thesis since they are related to the 3D face shape. Texture and depth videos are also available for each expression. This framework of the BU-4DFE data set is shown in Fig. 4.1.

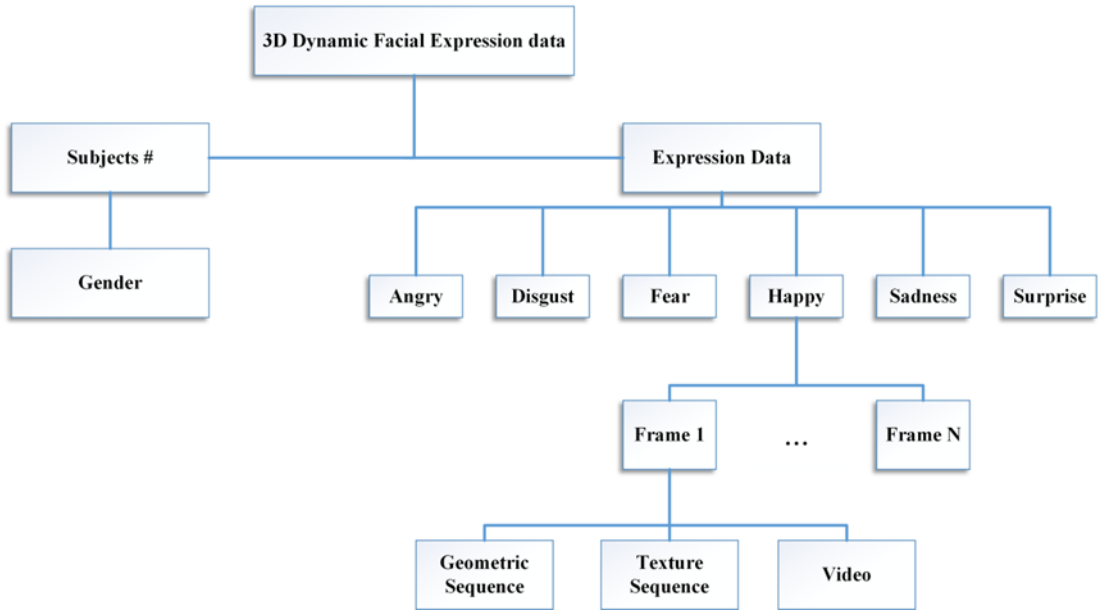


Figure 4.1: The Framework of BU-4DFE data set [55].

The subjects are with a variety of ethnic ancestries such as Asian, Black, Hispanic, and White. Each of the 101 subjects in data set has expressed 6 basic expressions including AN, DI, FE, HA, SA and SU. For each expression a sample recording is illustrated in Fig. 4.2. Samples are selected from both male and female groups and from different ethnicities. Since there are approximately 100 frames in each sequence, a subset of frames including 7 images are selected to represent the general process of the emotion expression.

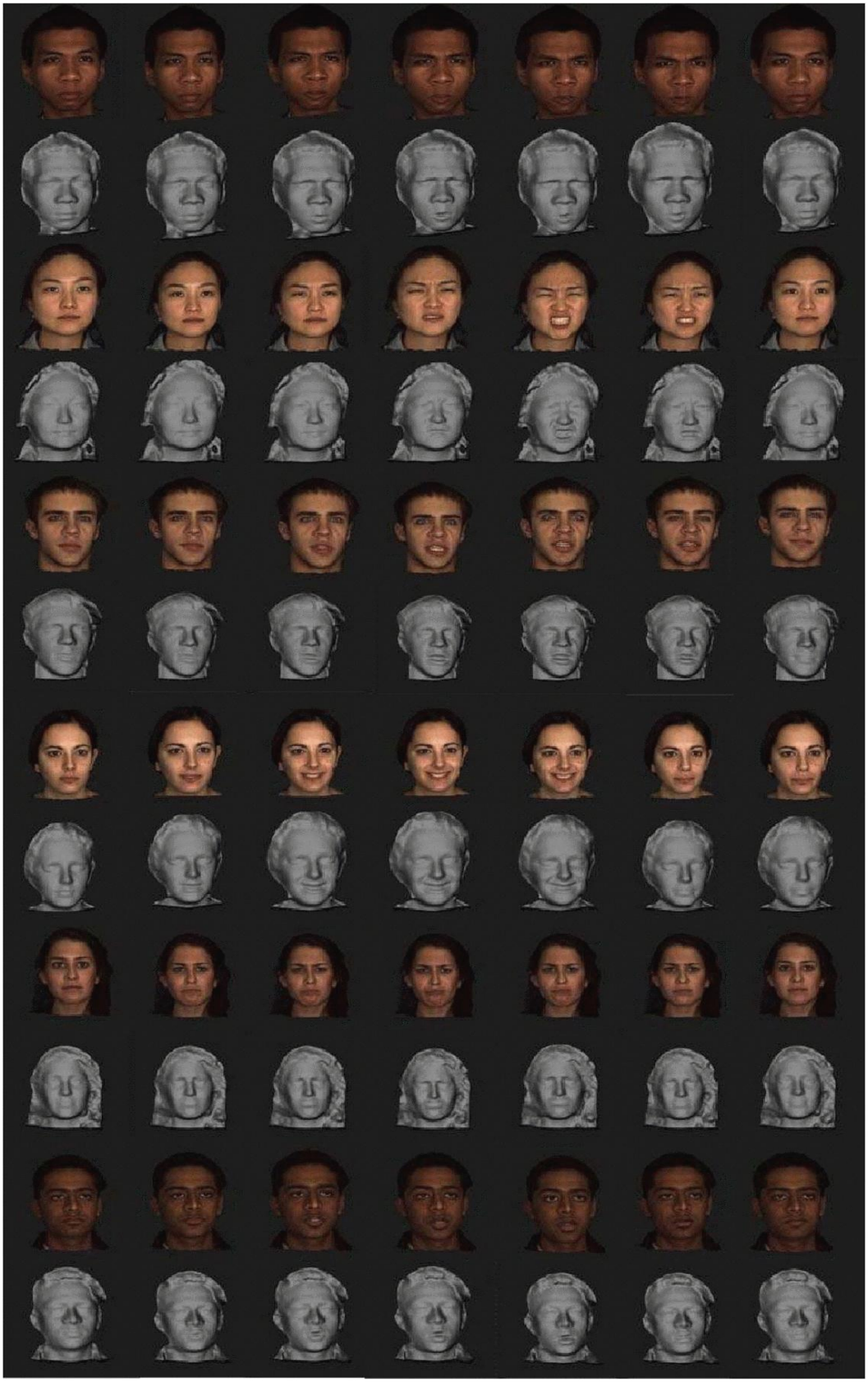


Figure 4.2: Sample frames of BU-4DFE texture and depth videos, from top to bottom: angry (male, black), disgust (female, East-Asia), fear (male, White), happy (female, White), sad (female, Latino), and surprise (male, India), respectively.

The rate of the recorded videos is 25 frames per second and thus the length of the sequences varies approximately between 3 to 4 seconds. For each expression in addition to texture and depth information, the 3-dimensional coordinates of 83 facial landmarks of face model are provided. These landmarks are located around critical facial regions including eyes, eyebrows, nose, mouth, chin and the face contour. In the first frame of each sequence, the landmarks are identified and then tracked in other frames using AAM. The depth (range model) sequences are aligned with the texture video and consequently, the detected 83 landmarks can also be projected into the depth sequences forming 3D feature vertices. These vertices and the 3D coordinates of the landmarks are also provided in the data set. Fig. 4.3 shows the tracked landmarks and the range models.

Figure 4.3: Tracked landmarks (top) and the range models (bottom) in BU-4DFE.

4.3 Proposed Method

The proposed method comprises four main stages: head pose correction and normalization, feature extraction, feature selection and classification. In the first stage, head pose correction and normalization are applied to adjust landmarks in all frames and among all subjects. Secondly, three types of geometric deformation

feature namely point, distance and angle feature are extracted and multimodal time series are constructed by sliding a mean deformation window over the full sequence. The next stage, feature selection based on nearest component feature selection (NCFS) aims at identifying a small subset of informative features and discarding redundant ones. Finally, AC-DTW is applied in classification stage to recognize the expressions. The architecture of proposed method is illustrated in Fig. 4.4 and the details of the stages are explained in the following.

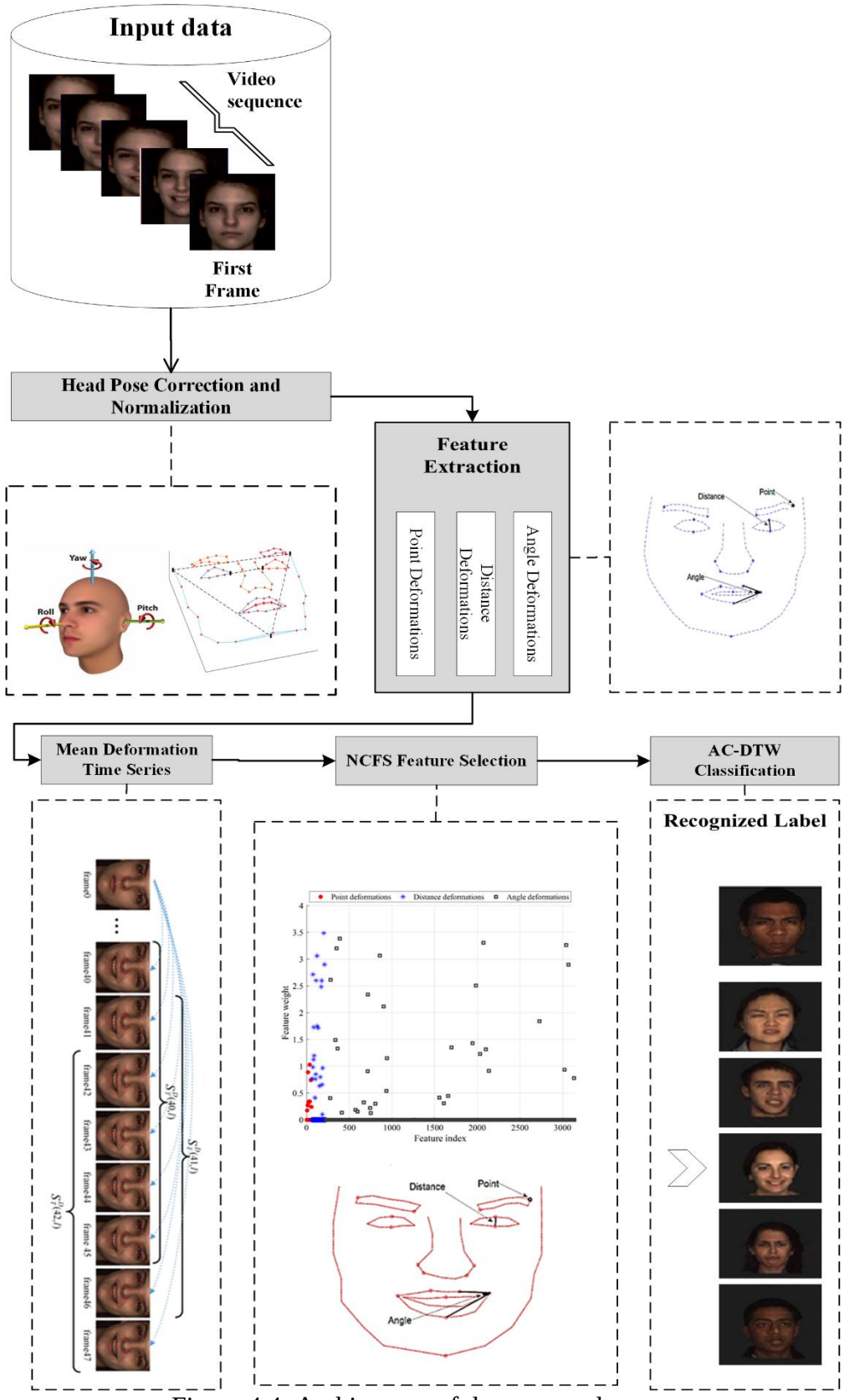


Figure 4.4: Architecture of the proposed system

4.3.1 Head Pose Correction and Normalization

In geometric D-FER systems which are designed based on facial landmarks, it is crucial to have a frontal view face. In other words, head pose can induce changes in the location of landmarks which may be mistaken as expression-related variations by the recognition system. Hence, it is of a great importance to normalize facial landmark coordinates according to head pose. Head movement is a known as an almost rigid transformation in 3-dimensional space. Derkach et al. [27] have argued that geometric landmark-based head pose estimation is as efficient as appearance estimate while it does not require a pre-training phase. They have also shown that this method outperforms dictionary-based methods.

In geometric landmark-based head pose estimation, given the coordinates of a number of facial landmarks, the head pose can be defined by three Euler angles around three axes. These angles are called pitch (nodding), yaw (shaking) and roll (tilting) as illustrated in Fig. 4.5 [25] Pitch angle (θ) is the head rotation around the horizontal *x* axis. Yaw angle (φ) measures the rotation around vertical *y* axis. Roll angle (ψ) is the rotation around the *z* axis perpendicular to frontal face plane. In order to estimate, two points are used. The line connecting inner eye corners in 3-dimensional space is projected into *xy* plane. Knowing that the equation of a line in *xy* plane $y = tan(\psi)x + b$, roll angle is obtained.

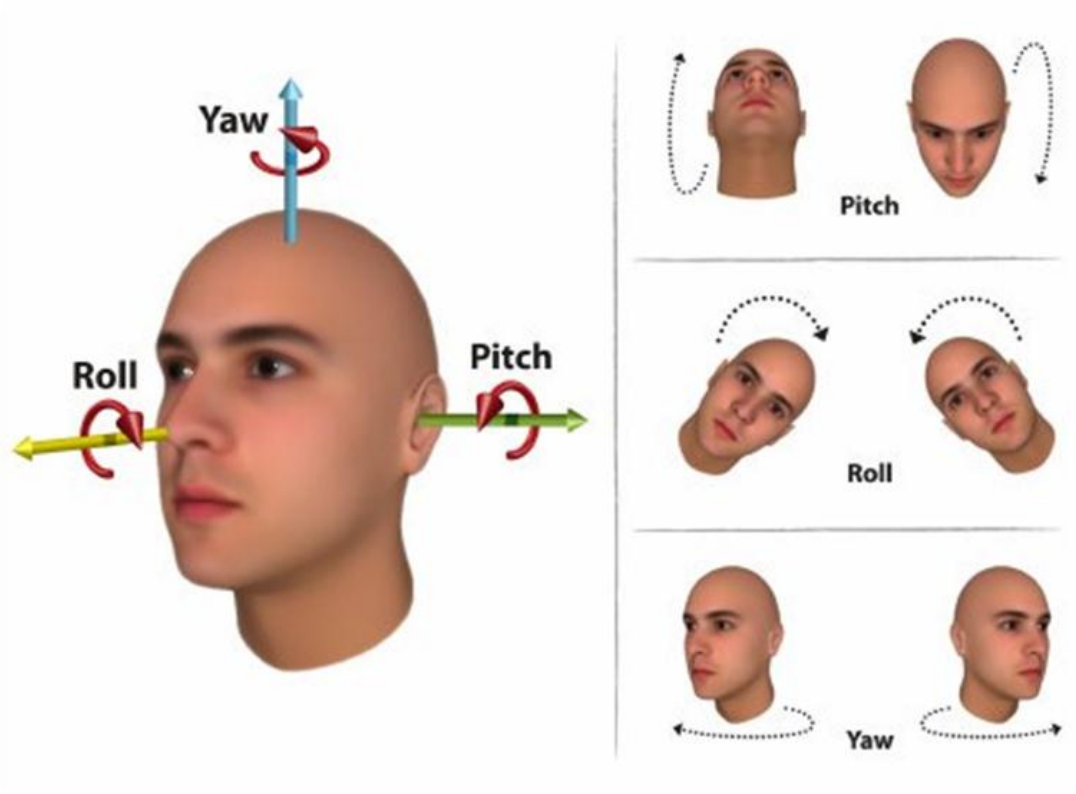


Figure 4.5: Head pose in terms of pitch, roll, and yaw angles describing 3D movement of head [25].

To estimate pitch and yaw, the equation of an approximate frontal face plane is needed. For this purpose, three points including forehead corners and chin tip are used. The normal vector of the plane crossing these three points are used to compute pitch and yaw rotation angles as follows.

$$\phi = \arctan\left(\frac{x_n}{z_n}\right) \qquad \theta = \arctan(\frac{y_n}{z_n}) \tag{4.1}$$

where $\vec{n} = [x_n, y_n, z_n]$ is the three dimensional normal vector of the frontal face plane.After estimation of the rotation angles *(θ, φ, ψ)* and given the reference point *(x_r, y_r, z_r)*, the landmark coordinates are corrected by applying a rigid affine transformation *F* as:

$$F_{\Theta,\Gamma} = [\Theta(\theta, \phi, \psi)|\Gamma(x_r, y_r, z_r)] \tag{4.2}$$

where the rigid transformation is defined by a rotation matrix Θ and translation vector Γ as shown in Eq. 4.2. Note that the midpoint of the inner eye corners is

assumed as the reference point in the 3D coordinate system to normalize the displacements among all frames. In each frame of the video sequence, the facial landmark point $l = [x, y, z]^T$ (l = l,...,19) is transformed to $l' = [x', y', z']^T$ as shown in Eq. 4.3.

$$l' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta & sin\theta \\ 0 & -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} cos\phi & 0 & -sin\phi \\ 0 & 1 & 0 \\ sin\phi & 0 & cos\phi \end{bmatrix} \begin{bmatrix} cos\psi & sin\psi & 0 \\ -sin\psi & cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} l + [x_r, y_r, z_r]^T \quad (4.3)$$

where θ ,φ and ψ are the estimated head rotation angles and (x_r, y_r, z_r) is the reference point coordinates. Note that this procedure is applied to the frames of all sequences in the data set. By correcting the head pose according to the reference coordinate system, point coordinates are normalized. The length of the line connecting inner eye corners is used as the reference length for scale normalization of distance deformations. Moreover, the projected length of this line on *x*, *y* and z is also used as a scale factor to compensate scale variance of point coordinates. Angle deformations are already scale and rotation invariant and do not need normalization. In Fig. 4.6 the facial landmarks in 3D space are illustrated. The reference line connecting inner eye corners and the frontal face plane are also annotated.

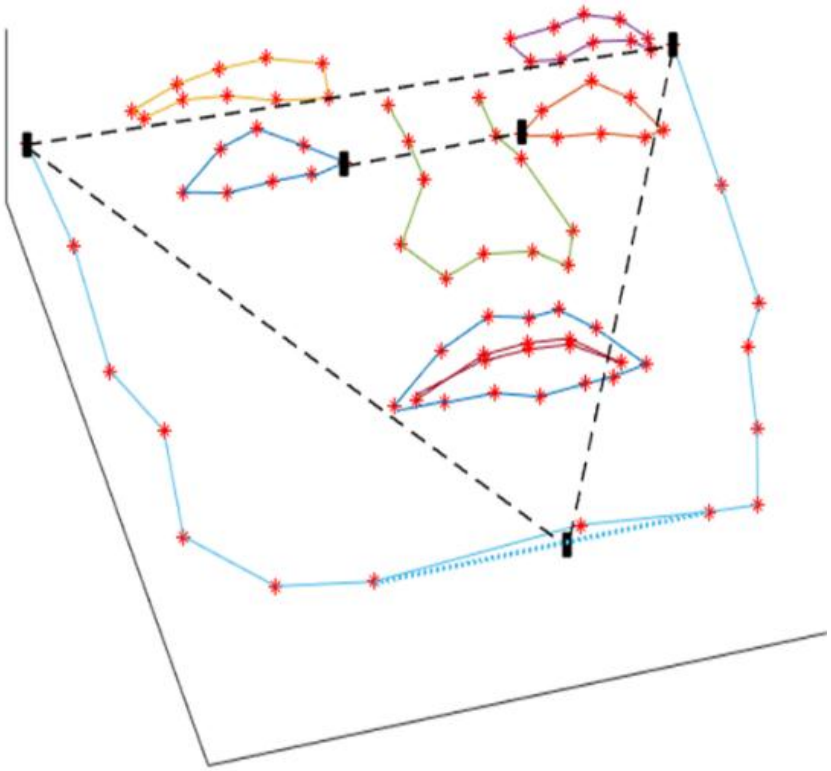


Figure 4.6: Facial landmarks in 3D space with reference line and frontal face plane.

4.3.2 Feature Extraction

As stated before, three types of geometric landmark based deformations are extracted from facial key points in this study. In BU-4DFE data set, there are 83 landmarks of the face model in 3D space numbered from 1 to 83. The two-dimensional perspective of these landmark points is illustrated in Fig. 4.7.

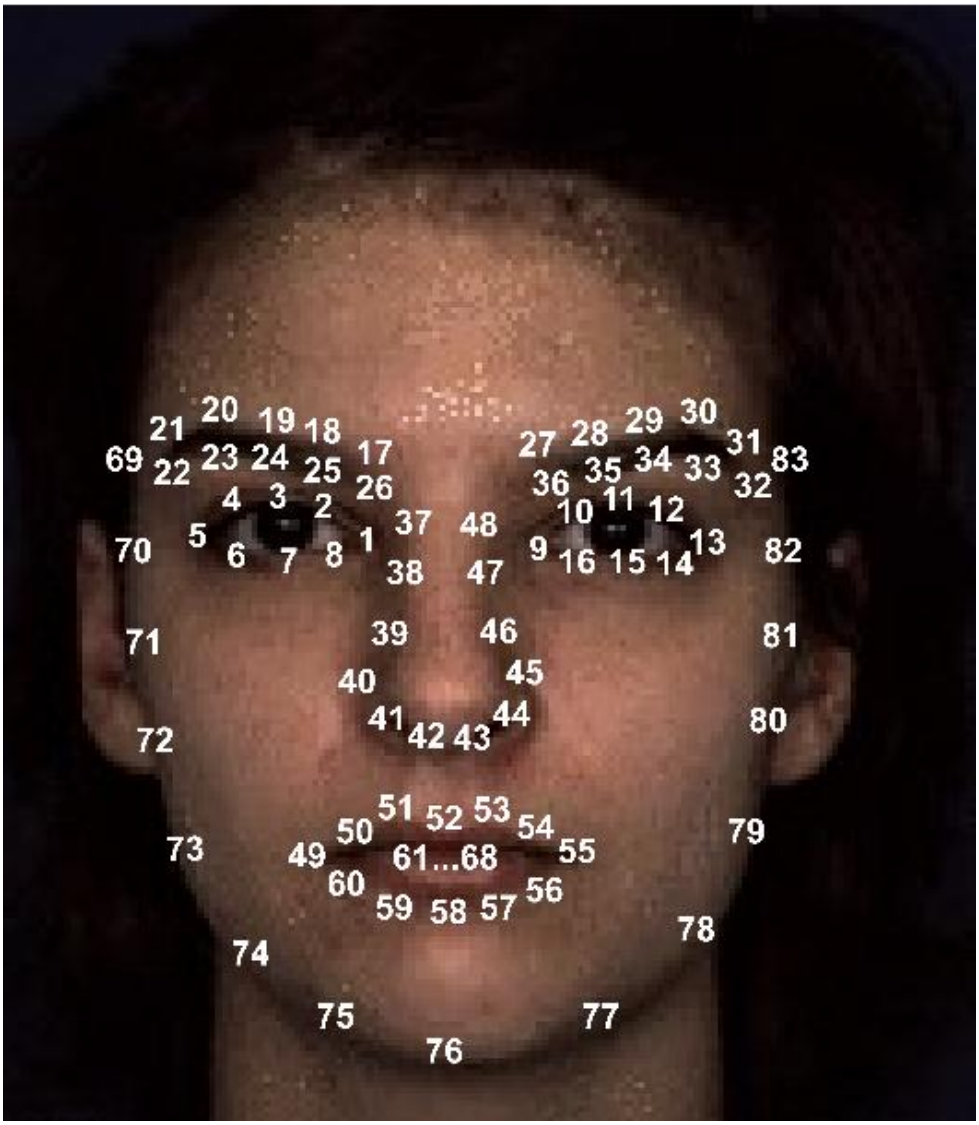


Figure 4.7: Facial landmarks in BU-4DFE data set.

As considering all of these landmarks for feature extraction is highly computationally costly, a subset of the landmarks are to be considered. We argue that a subset of these landmarks exhibit significant deformations during facial expression and the rest of them either do not deform during the expression or are redundant as they always deform similar to their neighboring landmarks. In order to identify these key points, we referred to FACS introduced by Ekman and Friesen [24] FACS defines 44 AUs based on facial muscle movements. Extensive research has been conducted to analyze how these AUs contribute to facial emotion expression both separately and in combination to each other [24], [26], [12].Basically, it has been argued that only a

small subset of these AUs involve in six basic facial expressions [107], [12], [26] . Tian et al. [12] have suggested an automatedface analysis system to recognize changes in facial expression into AUs. Their system has achieved significant recognition rates using 10 lower face and 6 upper face AUs. Wegrzyn et al. [107] have studied a facial map for emotions and identified the active AUs for each of the six basic expressions. A brief description of the AUs activating during six prototypic expressions is given in Table 4.1.

Table 4.1: Description of AUs contributing to six basic expressions

AU	Description	AU	Description
AU1	Inner brow raiser	AU15	Lip corner depressor
AU2	Outer brow raiser	AU16	Lower lip depressor
AU4	Brow lowerer	AU19	Tongue show
AU5	Upper lid raiser	AU20	Lip stretcher
AU6	Cheek raiser	AU22	Lip funneler
AU7	Lid tightener	AU23	Lip tightener
AU9	Nose wrinkeler	AU25	Lips part
AU10	Upper lip raiser	AU26	Jaw drop
AU12	Lip corner puller	AU27	Mouth stretch

In fact, these AUs displace a set of key points located around brows, eyelids, nostril wing, the nasal end of the nasolabial furrow, lips corners, philtrum, lips mover, and chin boss. Accordingly, the facial landmarks considered in this study are directly related to AUs which function during six prototypic expressions as shown in Table 4.2. It should be noted that several landmarks may be displaced by activation of one

specific AU. For instance upper lid raiser (AU5) deforms landmarks 2, 3, 4,10,11,12, but the considered landmarks in this study i.e 3, 11 have remarkable deformations compared to their neighboring landmarks. The reader can refer to [24] and [107] for more details.

Table 4.2: AUs contributing to six basic expressions and related landmarks numbered in Figure 4.6.

Expression	**Action Units**	**Landmarks**
Anger	{AU6,AU7,AU9,AU22,AU23,AU25}	{3,7,11,15,40,41,44,45,51,53,58,63}
Disgust	{AU4,AU6,AU9,AU10,AU16,AU19,AU25,AU26}	{31,36,7,15,44,55,58,63,76}
Fear	{AU1,AU2,AU5,AU20,AU25}	{3,11,21,26,31,36,49,55,58,63}
Happiness	{AU6,AU12,AU25}	{7,15,49,51,53,55,58,63}
Sadness	{AU1,AU4,AU15,AU25}	{21,26,31,36,49,55,58,63}
Surprise	{AU1,AU2,AU5,AU25,AU26,AU27}	{3,11,21,26,31,36,49,55,58,63,76}

The displacement of the 19 key points is described by three types of geometric deformations. Since the aim of this study is to capture both spatial and temporal information, related values are extracted from the individual frames and then time series features are constructed by applying a sliding window to obtain mean deformation. Geometric deformations are computed from the sequence of frames by taking the first frame as the reference one. We assumed that the first frame of each video sequence expresses a neutral face which is true to a great extent for the data set used in this work. For each type of geometric deformations, the deviation of the corresponding variable from the first frame is computed. Then, an averaging window is slid over the sequence to construct the time series. These steps are described in the following.

4.3.3 Point Deformation

As stated before, the time series-based features proposed in this work are obtained using mean geometric deformations. The sliding window averaging is applied on video sequences to compute three types of deformations. The first type of geometric deformations is defined as the displacement of the key points. Given a sequence of *N* frames numbered as *i=0,...,N-1*, the displacement of a the l^{th} landmark point's coordinates in the i^{th} frame relative to the first frame in 3-dimensional space is computed as:

$$\Delta x_i^l = x_i^l - x_o^l \qquad \Delta y_i^l = y_i^l - y_o^l \qquad \Delta z_i^l = z_i^l - z_o^l \tag{4.4}$$

where x_i^l, y_i^l and z_i^l are the three dimensional coordinates of the l^{th} landmark (*l*= l,...,19) in the i^{th} frame of the sequence. x_o^l, y_o^l and z_o^l are the landmark coordinates in the reference frame i.e. the first frame of the video sequence. Δ stands for the deformation along each coordinate. Fig. 4.8 shows how point deformations are obtained from a sample happy sequence. For each frame in the sequence, there are $3 * 19 = 57$ point deformation values. Similar scheme is used to extract distance and angle deformations from each sequence.

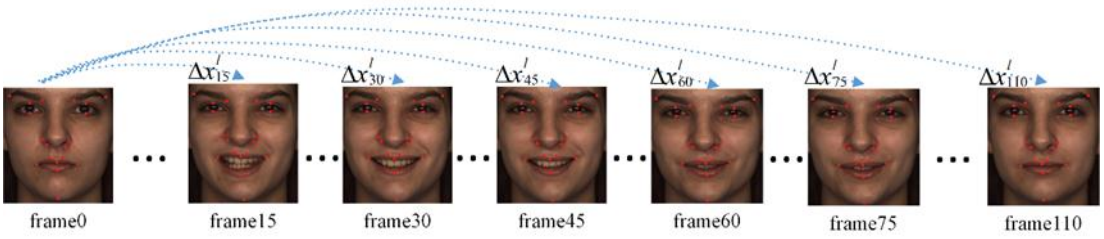


Figure 4.8: Computing point deformation from landmark coordinates in a sample happy sequence.

4.3.4 Distance Deformation

The second type of geometric deformations is defined as the changes in pairwise Euclidean distances between the key points. The number of pairwise distances

extracted from 19 key points explained in previous section is $\binom{19}{2} = 171$. Given a pair of landmarks *l* and *m* in the i^{th} frame, Euclidean distance between the landmarks is denoted as:

$$d_i^{l,m} = \sqrt{(x_i^l - x_i^m)^2 + (y_i^l - y_i^m)^2 + (z_i^l - z_i^m)^2} \tag{4.5}$$

The corresponding deformation values in the i^{th} frame are the deviations of the distance $d_i^{l,m}$ from the distance in the first frame ($d_0^{l,m}$) computed as:

$$\Delta d_i^{l,m} = |d_i^{l,m} - d_0^{l,m}| \tag{4.6}$$

4.3.5 Angle Deformation

These values are computed as the changes in the angle between the two sides of the triangle made by three key points. There are three angles in each triangle and thus the total number of angle deformations is equal to $3 * \binom{19}{3} = 2907$. Now, given a set of three key points: *l*,*m* and *k* in i^{th} frame as the vertices of a triangle, three vectors are firstly defined on each sides of the triangle. Given these vectors ($\vec{V}_i^{l,m}$;$\vec{V}_i^{l,k}$;$\vec{V}_i^{m,k}$), the three angles between each pairs of the vectors can be computed as:

$$\vec{V}_i^{l,m} = \left(x_l^i - x_m^i, y_l^i - y_m^i, z_l^i - z_m^i\right)$$

$$\vec{V}_i^{l,k} = \left(x_l^i - x_k^i, y_l^i - y_k^i, z_l^i - z_k^i\right)$$

$$\vec{V}_i^{m,k} = \left(x_m^i - x_k^i, y_m^i - y_k^i, z_m^i - z_k^i\right)$$

$$\alpha_i^l = \frac{\vec{V}_i^{l,m}\vec{V}_i^{l,k}}{\left|\vec{V}_i^{l,m}\right|\left|\vec{V}_i^{l,k}\right|} \qquad \alpha_i^m = \frac{-\vec{V}_i^{l,m}\vec{V}_i^{m,k}}{|\vec{V}_i^{l,m}||\vec{V}_i^{m,k}|} \qquad \alpha_i^k = \frac{\vec{V}_i^{l,k}\vec{V}_i^{m,k}}{|\vec{V}_i^{l,k}||\vec{V}_i^{m,k}|} \tag{4.7}$$

where $\vec{V}_i^{l,m}$ is vector initiating at l^{th} landmark and terminating at m^{th} landmark. Similarly, the two other vectors are computed. Note that angles between two vectors is the ratio of their inner products to the multiplication of their lengths. Angle

deformations are then obtained as the deviation of the α angles from corresponding values in the first frame computed as

$$\Delta\alpha_i^{l,m,k} = [\alpha_i^l - \alpha_0^l, \alpha_i^m - \alpha_0^m, \alpha_i^k - \alpha_0^k]^T \quad (4.8)$$

4.3.6 Multimodal Time Series Features

Multimodal time series features are the temporal representation of geometric deformations. Time series analysis has been never applied in FER systems. In this study, geometric deformation values are used to construct multimodal time series where the modes represent spatial information and the time represents the mean temporal information. Assuming a sequence of length *N*, the first frame of the sequence is the reference one numbered as 0. For each of the preceding frames numbered as 1 to $N - 1$, all geometric deformation values are calculated and concatenated to form a D-dimensional vector where D = 3135 (Note that the number of geometric deformations is equal to 57, 171 and 2907 for point, distance and angle deformations, respectively). Consequently, a frame of length *N* is represented by $N-1$ D-dimensional vectors. To obtain the multimodal time series features, a sliding window of size *w* with one frame shift is applied to the temporal axis of these geometric deformation sequences. Let S_T^D be the multimodal time series representation of the sequence of length *N*, where *T* is the length of the time series and *D* is the number of modes. The elements of S_T^D denoted as S_T^D *(t; l) (t = 1,…, T,* $T = N - w$ and *l = 1,…,D)* are computed as the local mean temporal deformation values across w frames. For example, the first mode of S_T^D is calculated from the deformation of the *x* coordinates of the first landmark point described in Section 4.3.3 as:

$$S_T^D(t,1) = \frac{1}{w}\sum_{i=t}^{t+w-1} \Delta x_i^1 \tag{4.9}$$

where w is the window size and t is the starting frame of the window on temporal axis. The other modes of the multimodal time series are calculated in a similar way from all geometric deformations. The sliding window size, w can be defined by evaluating the system performance using different values. Fig. 4.9 illustrates a portion of one sample happy sequence and sliding window of size 6. The proposed multimodal time series features defined in this study capture the local 4D spatiotemporal geometric deformations of the whole sequence.

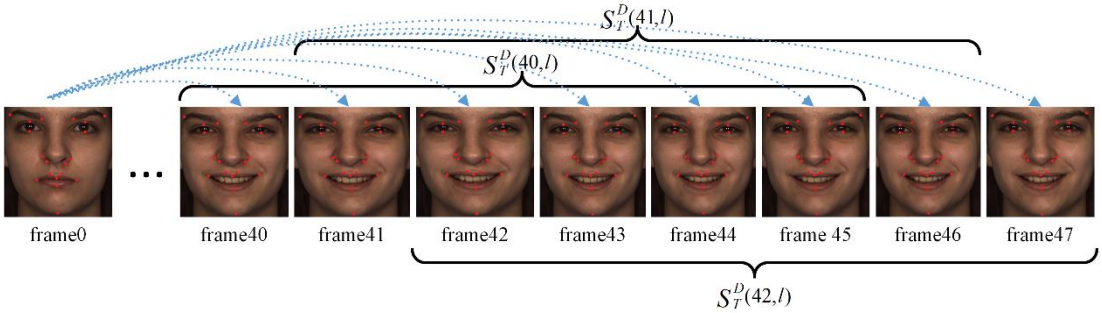


Figure 4.9: Computing local mean deformation in temporal domain for time series features in an example happy sequence with sliding window size = 6.

4.3.7 Feature Selection

Feature selection is applied in this study to reduce the dimensionality of the feature space or simply the modes of time series features. As features are extracted from facial landmarks, they are highly correlated and there is a high level of redundancy among them that reduces the performance of the classifier. Feature selection methods are not designed for time series and thus the maximum values of feature time series are considered as the input to neighborhood component feature selection. In fact, since the features represent the local temporal deviations from the first frame, the maximal values in temporal domain correspond to largest geometric deformations. The NCFS is applied in FER studies for the first time in this study. It is computationally efficient for high dimensional data sets such as the geometric features which we are challenging with in this work. In addition, it is a multivariate feature selection method that considers feature interactions and addresses the redundancy.

In other words, when two features are highly correlated, only one of them is assigned a nonzero weight. These features make it suitable for the proposed system. The NCFS algorithm performs as follows.

Let $X = \{(X_1, y_1), \dots, (X_n, y_n)\}$ be the set of samples in training data and their corresponding class labels i.e. expressions. Note that considering S_T^D as a time series feature for the sample sequence i, the $X_i \in R^D$ is computed as $X_i(:) = \max_t S_T^D(t, :)$ i.e. the maximum of the time series features. More precisely, the maximal value of each deformation curve is taken into consideration. Assume a distance function which is used to classify the samples using nearest neighbor classifier in D-dimensional space. By weighting the dimensions (features) using a linear transformation, the optimal subspace can be found where the informative/relevant features have larger weights and the redundant/irrelevant ones have zero weights. The NCFS finds a linear transformation vector L to transform the distance between two samples X_i and X_j:

$$d_{i,j}(L) = L^T[X_i - X_j] \tag{4.10}$$

Generally, NCFS defines an optimization problem that selects a d-dimensional subset from the original D-dimensional set of attribute (d $\ll$ D) to maximize the accuracy of the nearest neighbor classifier based on a leave-one-out scheme. Considering the fact that the nearest neighborhood classification accuracy is non-differentiable, a probability function is adopted to approximate the possibility of selecting a reference sample. The reference sample is the nearest sample and the probability of X_j being taken as the reference sample of X_i is given by Eq. 4.11.

$$p_{i,j} = \begin{cases} \frac{e^{-\frac{d_{i,j}(L)}{\sigma}}}{\sum_{k \neq i} e^{-\frac{d_{i,j}(L)}{\sigma}}} & i \neq j \\ 0 & i = j \end{cases} \tag{4.11}$$

where σ is the kernel function input parameter that controls the number of neighbors competing for reference point. $d_{i,j}$ (L) is the weighted distance given in Eq. 4.10. The sample X_i would be correctly classified if the reference sample has the same class label as it is how the one nearest neighbor (1NN) classifier works. Accordingly, the probability of correct classification of the one sequence is denoted as:

$$p_i = \sum_{\{j|y_j=y_i\}} p_{ij} \tag{4.12}$$

where y_i and y_j are the labels of samples X_i and X_j respectively and $p_{i,j}$ is the adopted probability explained in Eq. 4.11. In order to find the transformation vector L given in Eq. 4.10, an optimization problem is defined to maximize the accuracy of the 1NN classifier. Thus, the leave-one-out (LOO accuracy) of the nearest neighbor classifier is estimated as a function of the transformation vector as given below.

$$\Lambda(L) = \sum_i p_i - \lambda||L||^2 \tag{4.13}$$

where L is the transformation vector to weight the distances in D-dimensional space and λ is inserted as the regularization term.

Since the aim of NCFS is to maximize this objective function, its derivative with respect to L components is taken and used to update them in a learning procedure. As shown in Eq. 10, given the elements of L as weights of attributes, the ones with weights equal or close to zero are discarded. For details of the leaning algorithm, one may refer to Yang et al. [106]. After discarding the attributes with negligible weights, the number of modes of time series in reduced space is equal to d.

4.3.8 Classification

After reducing the dimensionality of feature time series, AC-DTW is applied to classify test samples [108] Conventional DTW searches for the optimal warping path to align two sequences of different lengths in order to minimize the distance. Noting that facial expression procedure starts with a neutral phase followed by onset, apex and offset phases, incorrect alignment of these phases degrades the classification performance. AC-DTW is a modified version of DTW that avoids over-stretching and over-compression of sequences with remarkable difference in their lengths which makes it suitable for D-FER systems. This algorithm has a cost function defined based on the number of points in one sequence mapped to one point in another one. The output is the distance between two times series and it is used for nearest neighbor classification. In this study, we extend the AC-DTW algorithm into a multivariate version to be applicable for classification of proposed multimodal time series features.

Considering two multimodal feature time series of different lengths denoted as $S_{T_1}^D$ and $R_{T_2}^D$, a preliminary distance matrix $D_{pr}= (d_{pr}(i,j))_{T_1 \times T_2}$ is defined with elements calculated as the pairwise distances between all the points on the two curves. Each element of the distance matrix is computed using all modes of the time series ($l = 1,\ldots,d$) as follows:

$$d_{pr}(i,j)=\sqrt{\sum_{l=1}^{d}[(S_{T_1}^D(i,l) - R_{T_2}^D(j,l)]} \tag{4.14}$$

where $S_{T_1}^D(i;\ l)$ and $R_{T_2}^D(j;\ l)$ are the elements of multimodal time series for two different sequences. In order to record the number of times each point is used in the warping path, two initially null matrices *AS* and *AR* are introduced with elements $a_{i,j}$ and $b_{i,j}$. The elements $a_{i,j}$ and $b_{i,j}$ respectively count the number of times each point of $S_{T_1}^D$ and $R_{T_2}^D$ are used in the warping path. Then, a control term (r) is added to adjust the tolerance to many-to-one mapping. The term r is denoted as $r = min\ (T_1, T_2)$ / max (T_1, T_2) to define a cost function. Note that the value of r is determined by the difference of the length of the time series. For two sequences of the same length,

the corresponding *r* value is equal to 1. As the length difference gets larger, the *r* value becomes smaller. This value is used to define a cost function as follows.

$$c(x) = g.r.x + 1 \tag{4.15}$$

where x takes its value from the elements of A_S and $A_R$and g is a factor to control the effect of cost function. Now, for simplicity, let's name the points of S_{T_1} and R_{T_2}as S_i and R_j respectively. A dynamic programming approach is applied to find the optimal warping path *P(i, j)* as given below.

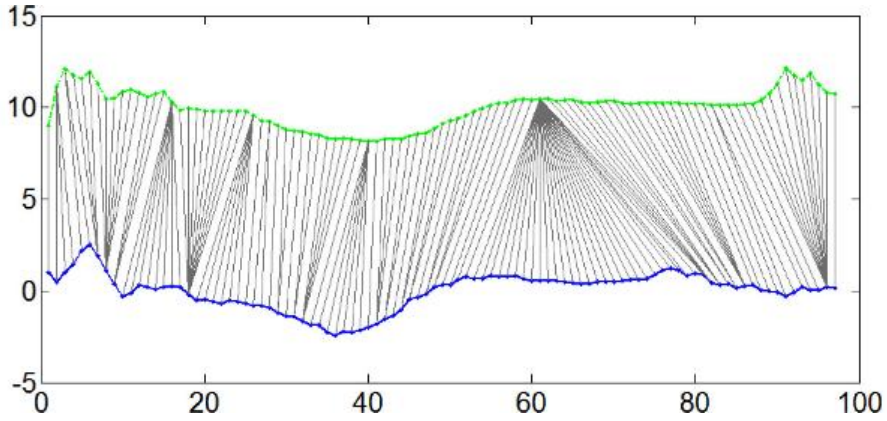
$$P(i,j) = min\begin{cases} c\left(b_{i-1,j}\right).d_{pr}(i,j) + P(i-1,j), & r_j \text{ is reused} \\ d_{pr}(i,j) + P(i-1,j-1), & \text{no reused point} \\ c(a_{i,j-1}).d_{pr}(i,j) + P(i,j-1), & s_i \text{ is reused} \end{cases}$$

(4.16) where *c* is the cost function that is applied when a point is reused from either of curves. Note that there is no cost when a point is not reused in the warping path. Otherwise, the elements of A_S and A_Ri.e. $a_{i,j}$ and $b_{i,j}$are used to determine the cost function. It should be also mentioned that in order to count the number of reuses, the elements of A_S and A_R are updated as:

$$\left(a_{i,j}, b_{i,j}\right) = \begin{cases} \left(1, b_{i-1,j} + 1\right), & r_j \text{ is reused} \\ (1,1), & \text{no reused point} \\ \left(a_{i,j-1} + 1, 1\right), & s_i \text{ is reused} \end{cases} \tag{4.17}$$

Finally, the AC-DTW distance between two multivariate time series is *P* (T_1*;* T_2). Fig 4.9 shows an example of optimal warping paths found by original DTW and AC-DTW [108] . It is obvious that AC-DTW aligns the sequences more appropriately without extreme extension or compression.

(a)



(b)

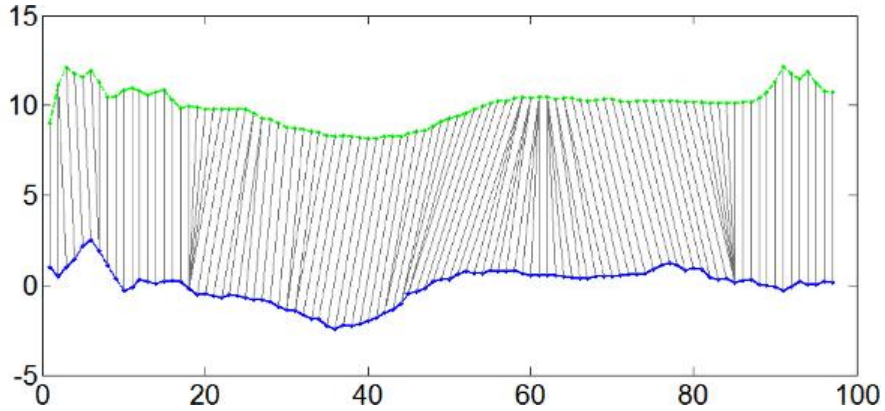


Figure 4.10: Optimal warping paths found by (a) DTW and (b) AC-DTW.

4.4 Experimental Results

In order to evaluate the performance of the proposed D-FER system, a set of experiments are conducted on BU-4DFE [55] a well-known dynamic 3D facial expression recognition data set. This data set is collected from 101 subjects including 58 female and 43 male subjects. The subjects are with a variety of ethnic ancestries such as Asian, Black, Hispanic, and White. Each of the 101 subjects in data set has expressed 6 basic expressions including AN, DI, FE, HA, SA and SU. Each expression is recorded as a video sequence of rate 25 frames per second. The length of the sequences varies approximately between 3 to 4 seconds. For each expression, texture and depth information are captured and the 3-dimensional coordinates of 83

facial landmarks of face model are provided. Resolutions of the depth and texture video sequences are 35,000 vertices and 1040×1329 pixels per frame respectively.

In order to be able to compare our results with previous studies, the experiments are conducted on both 100 and 60 subjects. Equal number of male and female subjects are randomly selected. All the experiments are implemented in *Matlab2018Rb*. Head pose correction and normalization phase is performed on all of frames of the data as a preliminary step. Frontal face plane is constructed by three points including forehead corners (landmarks {69; 83} and the chin middle point defined as the midpoint of landmarks {75; 77} (refer to Fig. 4.7). The reference point for rigid transformation applied in head pose correction phase is the midpoint of the line connecting inner eye corners. It should be noted that the length of this line is utilized for normalization of distance values. Given the line connecting inner eye corners and frontal face plane, pitch, yaw and roll angles are estimated. The head pose is then corrected by transforming coordinates of the landmarks.

In feature extraction stage, three types of landmark based geometric deformations including point, distance and angle are extracted form 19 key points around eyes, eyebrows, nose, lips and chin. Referring to Fig. 4.7 which shows the landmarks in BU-4DFE data set and Table 4.2 which represents the landmarks significantly representative for deformations in basic expression, the considered landmarks for this stage of the study are: {3; 7; 11; 15; 21; 26; 31; 36; 40; 41; 44; 45; 49; 51; 53; 55; 58; 63; 76}. After computing deformation values, a local temporal mean operator as a sliding window of length w with one frame shift is applied to obtain multimodal time series. The number of frames in each window (w) is defined through a set of

experiments evaluating system performance for different window sizes as 1,4,6 and 15.

Having all of the obtained time series features, subject independent (10-CV) is applied to partition the data into train and test set. In the experiments conducted on 60 subjects, 54 subjects are considered for train and the remaining 6 subjects for test in each fold. Similarly, there are 90 train and 10 test subjects for experiments on 100 subject. NCFS is applied on the train data and resulting attribute weights are used reduce the modes of time series features in both train and test sets. It should be noted that maximum values of the deformations along temporal axis are taken as the input to feature selection phase. NCFS is implemented by Statistics and Machine Learning ToolboxTM in Matlab2018b using neighborhood component for classification (*fscnca*) function. Fig. 4.11 shows an example of attributes' weights obtained by NCFC in one of the 10 folds for $w = 6$. Three regions are identified by different markers to spot point, distance and angle deformations respectively. Note that there are many redundant features with negligible weights which can be discarded.

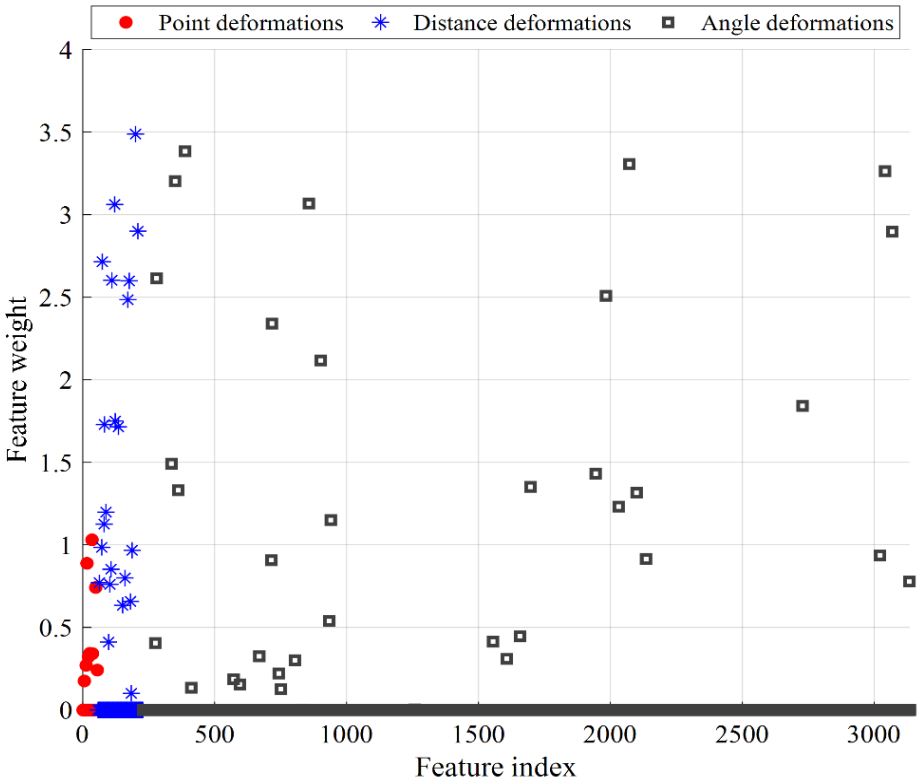


Figure 4.11: Feature weights obtained by NCFS.

A visual representation of the selected attributes is given in Fig. 4.12. The dimensionality of the reduced space is remarkably lower that original space. More precisely, the average size of the selected subsets across all folds is 67 features.

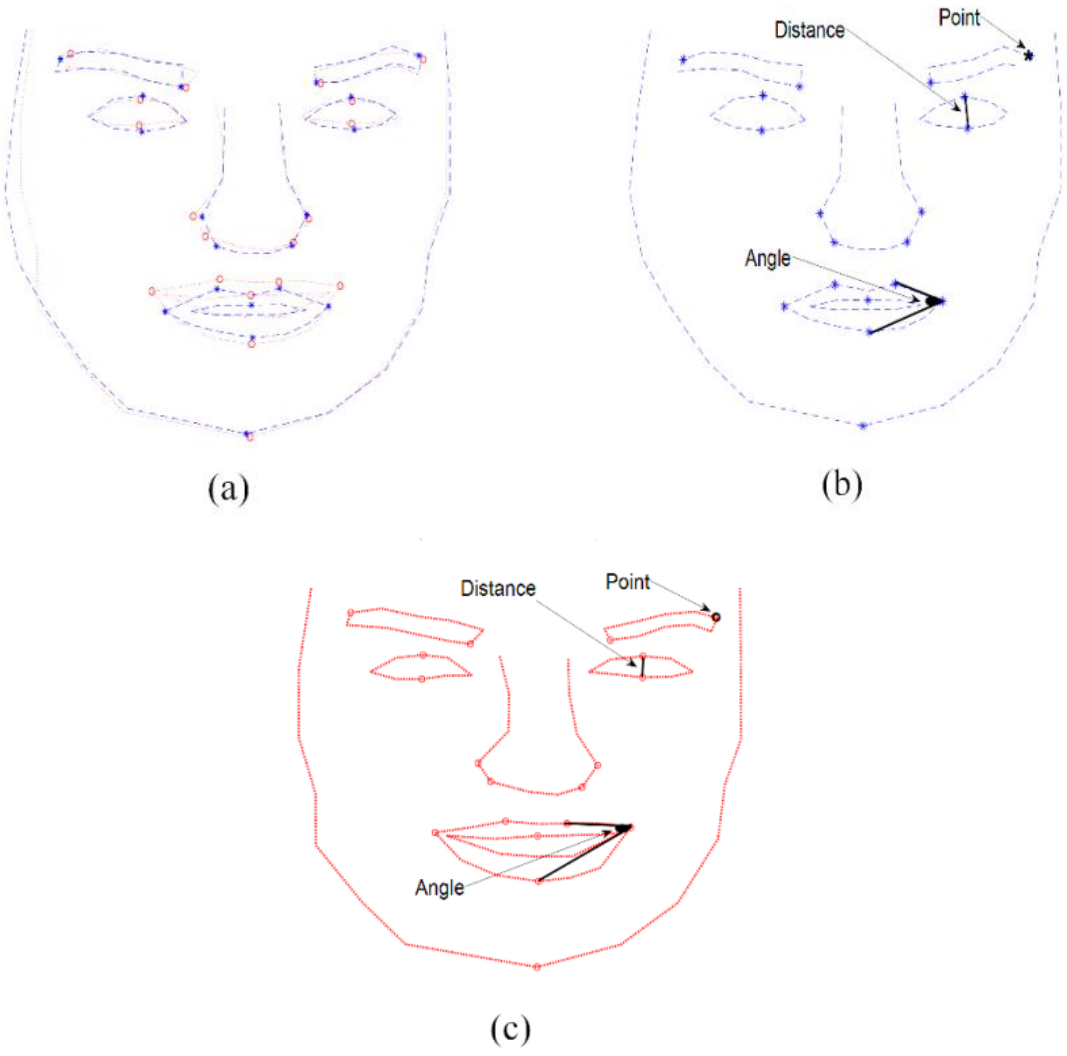


Figure 4.12: Visual representation of selected features of point, distance and angle deformation for a sample happy expression, (a) landmark locations in reference frame (blue) and an expressive frame (red), best landmark-based deformations in (b) reference frame and (c) expressive frame.

The classification phase of the proposed method is training-free. For each of the test samples represented as a multivariate time series in reduced space, multivariate AC-DTW distance from all train samples is computed. As AC-DTW optimizes the stretching and compression rate, the distances are correctly estimated regardless of the variations in the length of the sequences. The nearest neighbor classifier is applied on the computed distances to recognize the label. In order to have a comparative perspective, conventional DTW is also implemented. All the

aforementioned phases are evaluated using different window sizes. Table 4.3 presents the average recognition rate of six basic expressions for different values of (*w)* on 60 and 100 subjects separately. Both AC-DTW and DTW classification results are presented in Table 4.3.

Table 4.3: Average recognition rate of AC-DTW and DTW on six basic expressions for different sliding window sizes.

#Subjects	Window size	Average recognition rate (%)	
		AC-DTW	DTW
S=60	w = 1	91.39	78.61
	w = 4	91.94	79.44
	w = 6	**92.50**	**80.00**
	w = 15	89.44	77.78
S=100	w = 1	82.00	69.17
	w = 4	82.67	68.83
	w = 6	**83.50**	**69.83**
	w = 15	81.33	68.50

Window size, *w*, is an important parameter where highest accuracy is achieved with *w*=6 for both AC-DTW and DTW classification schemes. In fact, an optimal window size requires a tradeoff between preserving the details of the temporal deformations and smoothing. A very small window size is not capable of smoothing out the artifacts while a very large window size over-smoothes the aligned curves. In addition, AC-DTW significantly outperforms conventional DTW according to its capability to align the curves by preserving a tradeoff between the number of points mapped to a single point and the minimum distance obtained from the warping path. More precisely, AC-DTW finds a more reliable distance value as it aligns the curves

more efficiently. This difference between DTW and AC-DTW is clearly shown in Fig. 4.13.

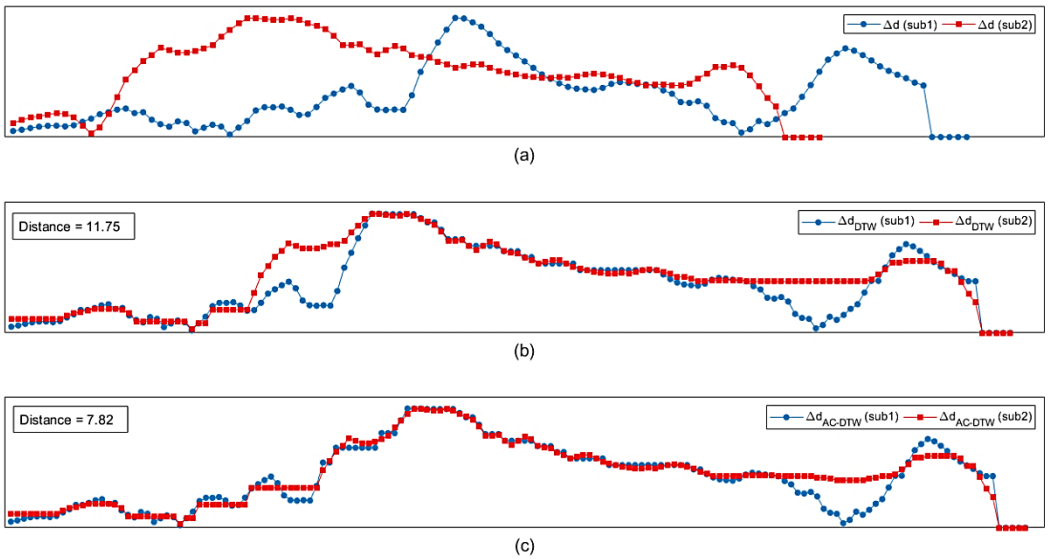


Figure 4.13: Distance calculation for anger sequences of two subjects, (a) original distance deformation curves, (b) curves aligned by DTW and (c) curves aligned by AC-DTW

In the figure, two deformation curves (w=6) of anger expression are superimposed. It should be noted that the curves are one of the selected modes of the extracted time series features. These curves belong to two different subjects named as sub1 and sub2. Fig. 4.13 (a) illustrates the original curves. The horizontal axis is the time and it can be seen that the length of the sequences is significantly different. In order to compute the distance between the curves, DTW and AC-DTW are applied to find the warping path. In the warping path, the curves are aligned by mapping several points to one point aiming at minimizing the distance. Fig. 4.13 (b) shows the curves aligned by DTW and Fig. 4.13 (c) illustrates the curves aligned by AC-DTW. Note that the many-to-one mapping is illustrated by repeating the single point on the aligned curve resulting in a horizontal section. This way, the aligned curves would be

of the same length. The distances obtained from the minimum warping path are equal to 11.75 and 7.82 for DTW and AC-DTW respectively. The shapes of the aligned curves intuitively confirm that AC-DTW is superior to DTW. In addition, considering that both curves are extracted from the same expression sequence i.e. anger the smaller the distance is the more accurate the classification is.

In order to have a clearer image of system performance for each of the six basic expressions, the confusion matrix of recognition results is represented. Table 4 and Table 5 show the confusion matrix of recognition results for 60 subjects and 100 subjects respectively. Average recognition accuracy is 92:50% for 60 subjects which means 333 sequences out of 360 sequences are recognized correctly by the system. For the case of 100 subjects, 501 out of 600 sequences are correctly distinguished by the system. The drop in recognition rate when comparing 60-subject and 100-subjects can be explained by the corrupted sequences bias in the data set.

Table 4.4: Confusion matrix for 60 subjects and window size w = 6.

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**95.00**	3.33	1.67	0.00	0.00	0.00
Disgust	5.00	**88.33**	5.00	0.00	1.67	0.00
Fear	1.67	5.00	**88.33**	1.67	3.33	0.00
Happy	0.00	0.00	1.67	**96.67**	1.67	0.00
Sadness	5.00	0.00	3.33	0.00	**90.00**	1.67
Surprise	0.00	0.00	3.33	0.00	0.00	**96.67**
Overall	**92.50**					

Table 4.5: Confusion matrix for 100 subjects and window size w = 6.

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**86.00**	6.00	4.00	0.00	3.00	1.00
Disgust	7.00	**78.00**	9.00	1.00	4.00	1.00
Fear	4.00	6.00	**82.00**	3.00	5.00	0.00
Happy	0.00	3.00	8.00	**86.00**	2.00	1.00
Sadness	9.00	2.00	4.00	1.00	**82.00**	2.00
Surprise	1.00	2.00	9.00	0.00	1.00	**87.00**
Overall	**83.50**					

As a matter of fact, two sets of additional experiments are conducted starting with only point deformations to provide a clear perspective on the relative contribution of each type of deformations. Then, distance deformations are added to the system and the whole process is repeated by merging point and distance deformations. In each experiment, all phases of the proposed method including head pose correction, related normalization, feature selection, and classification are implemented. It should be noted that the experiments in this section are performed on 60 subject and the window size is set to w = 6. Table 4.6 shows the contingency table of the point deformation-based system. The performance is remarkably lower than that of the proposed algorithm. In fact, the average accuracy has dropped to 65.28% with lowest recognition rate on disgust expression. Then, distance deformations are concatenated with point deformations and the experiments are replicated. The results are represented in Table 4.7 confirming the significant contribution of distance deformations in the performance of the system. More specifically, average accuracy is improved to 90.00% with the highest improvement rate (31.66%) on disgust expression recognition.

Table 4.6: Confusion matrix of point-deformation-based system (60 subjects and window size w = 6).

Expression	**Recognition Accuracy (%)**					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**68.33**	11.67	6.67	8.33	1.67	3.33
Disgust	15.00	**56.67**	16.67	5.00	5.00	1.67
Fear	5.00	11.67	**65.00**	3.33	10.00	5.00
Happy	6.67	6.67	8.33	**66.67**	11.67	0.00
Sadness	18.33	5.00	8.33	5.00	**61.67**	1.67
Surprise	5.00	0.00	13.33	5.00	3.33	**73.33**
Overall	**65.28**					

The aforementioned systems are compared to the proposed system encompassing all three types of deformations i.e. point, distance and angle. The results of this section approve that the distances of the facial landmarks play a crucial discriminative role in expression recognition. In other words, for expressions that have a high rate of confusion, displacement of the landmark reflected by point features is not discriminative enough. Distance features represent the relative movement of the facial landmarks and thus enhance the recognition rate. On the other hand, the weights of selected features illustrated in Figure 4.10 still support the influence of point deformation. Moreover, the average accuracy increases from 90.00% to 92.50% by adding angle deformations which advocates the effectiveness of angle deformations.

Table 4.7: Confusion matrix of point and distance-deformation-based system (60 subjects and window size w = 6).

Expression	Recognition Accuracy (%)					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**91.67**	3.33	3.33	0.00	1.67	0.00
Disgust	3.33	**88.33**	6.67	0.00	0.00	1.67
Fear	1.67	5.00	**86.67**	3.33	3.33	0.00
Happy	1.67	1.67	1.67	**90.00**	5.00	0.00
Sadness	5.00	0.00	5.00	0.00	**88.33**	1.67
Surprise	0.00	1.67	3.33	0.00	1.67	**93.33**
Overall	**90.00**					

Lastly, in order to have a comparative perspective of proposed system performance, recognition rates of recent D-FER studies conducted on BU-4DFE data set are represented in Table 4.8. For each study, method, classification and experimental settings are mentioned in the table. For instance, subject (S = 60); six expression (6E); 10-CV; window (Win = 6) means the experiments have been conducted on 60 subjects to detect 6 basic expression and 10-CV is used. Win = 6 means the subsequences of length 6 frames are processed and classified. For the full seq. cases, the whole sequence is processed and classified similar to current work. Key-frame means particular frame is identified in the sequence and is used for recognition. In general, we can argue that for 100 subjects the proposed algorithm performs significantly better than other methods. For 60 subjects, the recognition rate of our work is superior to the results reported in [66] [3] [86] [61][63] [109] and [110] The limitation of key-frame-based methods proposed by Zhen at al. [73] is that clustering based identification of key-frames makes the real-time implementation impractical compared to our method which simply calculates the distance between two vectors. Hence, our method is suitable for real-time applications. Although they achieve high performance by using a high amount of spatial data (50 sample vertices of each of the 200 radial curves) in case of low

quality videos and mass scale data, temporal information might be used to compensate the lack of sufficient spatial data. Moreover, their algorithm for key-frame detection is based on the assumption that all sequences contain onset, apex and offset phases. In real-time applications however, a video can contain either of these phases and not necessarily all of them. Our proposed system does not assume this and as it aligns the curves, missing phases have less effect on the performance. Sun et. al. [54] which achieves higher performance considers all surface points for vertex flow modeling. This requires high quality video which may not be suitable for practical purposes with low quality and low-resolution data. The study by Amor et al. [60] results in higher recognition performance but has limitation due to its sensitivity to precise nose tip detection. In fact, the algorithm requires frontal view faces and since the whole procedure relies on nose tip detection, head rotation would reduce the performance. Our proposed method on the other hand, does not rely on frontal view faces since it uses a set of landmarks to estimate and correct head pose. It is worth mentioning that proposed approach marginally outperforms the deep learning method proposed by Li et al. [110]. Their Dynamic Geometrical Image Network (DGIN) has resulted in 92.25% average accuracy. However, deep learning approaches are computationally costly, require large data sets and elaborate preprocessing stage [92]. It can be argued that proposed low-complexity approach is effective even when it is compared to complicated deep-learning approaches.

Table 4.8: Comparison of proposed method with previous studies on BU-4DFE data set.

Research Work	Method	Classifier	ES*	ACC*
Reale et al.[111]	Spatiotemporal Volume + Nebula Feature	SVM-RBF	100S, 6E, LOO, Win=15	76.10 %
Fang et al.[71]	MeshHOG + LBP-TOP	SVM-RBF	100S,6E,10-CV,-	75.82 %
Fang et al.[71]	Spin Image + LBP-TOP	SVM-RBF	100S,6E,10-CV,-	74.63 %
Proposed Method	**Multimodal Time Series Geometric Deformation + NCFS**	**AC-DTW**	**100S,6E,10-CV, Full seq.**	**83.50 %**
Yao et al. [66]	Texture and Geometric Scattering	MKL	60S,6E,10-CV, Key frame	90.12 %
Zhen et al.[73]	Spatial Facial Deformation + Temporal Filtering	HMM	60S,6E,10-CV, Key frame	95.13 %
Zarbakhsh et al.[3]	LBP-TOP + Spatiotemporal Region of Interest	HCRF	60S,6E,10-CV,-	86.67 %
Sandbach et al. [86]	3D motion-based Features	GentleBoost + HMM	60S,6E,6-CV, Variable Win	64.60 %
Amor et al. [60]	Geometric 3D Motion Extraction	LDA-HMM	60S,6E,10-CV, Win=6	93.21 %
Sun et al. [54]	Transformational Vertex Flow	HMM	60S,6E,10-CV, Win=6	94.37 %
Berretti et al [61]	Pairwise Distance of 3D Landmarks and SIFT	HMM	60S,6E,10-CV, Win=6	72.25 %
Xue et al. [63]	3D-DCT + mRMR	LDA and kNN	60S,6E,10-CV, Full seq.	78.80 %
Berretti et al [61]	Pairwise Distance of 3D Landmarks and SIFT	HMM	60S,6E,10-CV, Full seq.	79.40 %
Zhen & Huang [109]	Muscular Movement Model + Genetic Algorithm	SVM + HMM	60S,6E,10-CV, Full seq.	87.06 %
Li et al. [110]	Geometric Images (DPI, NCI, SII)	Deep Learning	60S,6E,10-CV, Full seq.	92.22 %
Proposed Method	**Multimodal Time Series Geometric Deformation +NCFS**	**AC-DTW**	**60S,6E,10-CV, Full seq.**	**92.50 %**

4.5 Discussion

The proposed dynamic 3D facial expression recognition system is tested on BU-4DFE data set. Experiments are conducted using 60 and 100 subjects in order to be able to compare the recognition rates with the state-of-the-art. The suggested approach based on geometric landmark-based local deformations and AC-DTW have resulted in 83.50% and 92.50% average recognition rate for 100 and 60 subjects respectively. This rates confirm the effectiveness of the proposed system. On 60 subjects, the highest recall rate (96.67%) is achieved for happy and surprise expression while the lowest rate is obtained for fear and disgust expressions (88:33%). Referring to Table 4.8, these numbers are satisfactory compared to previous studies.

4.6 Conclusion

In this study, a new approach in dynamic 3D facial expression recognition is proposed based on time series analysis of landmark-based geometric deformations. After head pose correction and normalization of landmark positions, a comprehensive set of geometric deformations are computed to form multimodal time series features. Referring to the activation patterns of facial AUs in six basic expressions, a set of 19 landmarks out of 83-annotated landmarks in BU-4DFE data set are considered. From these landmarks, point, distance and angle deformations relative to the reference frame are computed in all the consecutive frames of the expression sequences. Multimodal time series features are constructed by computing the temporal local mean of the deformation. This step is performed by applying a sliding window mean operator on all these point, distance and angle values from the first frame. To tackle the high dimensionality of the feature space i.e. the modes of time series features, NCFS feature subset selection is applied. NCFS is an effective

and simple supervised embedded feature subset selection method suitable for high dimensional data with a large number of irrelevant features. As a result, it selects a small subset of informative geometric deformations for classification. Based on selected features, the modes of original multimodal time series features are reduced. Finally, a recently introduced variant of DTW known as AC-DTW is utilized to classify these time series.

The main limitation of the current work is that it relies on the correct facial landmark locations. In the cases where the coordinates of landmarks are not available, an automated landmark detection method is required. However, considering the successful studies conducted on automatic facial landmark detection in recent years, landmark detection is not a serious issue. On the other hand, relying on landmarks for dynamic feature extraction makes the system less sensitive to noise, and lighting artifact unlike the systems relying on feature descriptors extracted from texture and depth. As a potential future work, an automated landmark detector/tracker may be attached to the system to implement an automated low-complexity system suitable for real time applications.

Chapter 5

APPEARANCE-BASED FEATURE DESCRIPTOR IN 4D FER

5.1 Introduction

As described in the previous chapter, facial landmarks play a crucial role in facial expression recognition research. Since the first two stages of this study, geometric approaches in static and dynamic FER have been explored respectively. Novel image processing and machine learning methods are suggested and successfully tested. However, geometric approaches ignore the valuable information in the texture and depth images or videos. In this stage, we developed a non-geometric 4D FER system also known as dynamic 3D FER or D-FER system. Generally, D-FER systems consist of several blocks including preprocessing, feature extraction and classification. In preprocessing stage, face bounding box detection, alignment, intensity normalization and imaging variation compensation are applied. In feature extraction phase, several methods have been proposed based on facial landmarks or AUs such as AU-based geometric measures [56], diffeomorphic motion features [57], and AAM landmark coordinates [58] [112] . The main limitation of such feature extraction methods is that they rely on manually annotated facial landmarks. In other words, an automated system would be required to firstly detect landmarks and then it should be examined whether the system performs acceptably with automatically detected landmarks.

On the other hand, generalized versions of well-known feature descriptors have been successfully applied in D-FER systems including SIFT [113], HOG [114] ,Gabor wavelet [65] and LBP-TOP [59] .There are two main challenges in this category of studies. Firstly, there is a high load of computation since the number of cuboids in each sequence is large. Secondly, descriptors provide too much detailed information which is not possible to be fed into the classifier. In order to tackle this problem, it is required to summarize spatial information. In recent years, sparse coding algorithms have become popular in image and video processing[79], [115]. By converting feature descriptor matrices into sparse representation, SPP can be used to construct a compact representation of an image. However, in facial expression recognition, there are specific spatiotemporal regions which contain most of the discriminative information for classification of expressions. These regions of interests are better candidates for feature pooling than conventional pyramid pooling. In this regard ROIs can be defined based on facial AUs or facial landmarks, which are more representative for expression than appearance. This approach results more subject-independence of the system.

In this chapter, a dynamic feature extraction system based on low-rank sparse coding and ROI pooling is proposed. The first contribution of this part of the work is that we extend the notion of SPP into ROI pooling. It should be noted that ROIs are defined based on automatically detected landmarks. In fact, landmarks are acquired in the first frame of each sequence and then they are tracked in other frames in order to reduce time and computational complexity. Spatiotemporal regions of interest are determined using identified landmarks. The proposed system combines the idea of cuboid-based feature descriptors with AU information by pooling sparse codes from

spatial ROIs. To the best of our knowledge, low-rank sparse coding is applied in a FER for the first time. The rest of this chapter is as follows. Proposed methodology is described in Section 5.2. Results of conducted experiments are represented in Section 5.3. In Section 5.4 the results are discussed. Finally, this stage of the study is concluded in Section 5.5.

5.2 Proposed Method

The proposed system is fully automated and comprises several phases. The first two phases are landmark detection and tracking using particle filters. Then, spatiotemporal ROIs are extracted using positions of landmarks. LBP-TOP Feature descriptors are then computed in cuboids. Subsequently, sparse coding phase is implemented based on low rank sparse coding (LRSC). Finally, hidden-state CRF are employed for classification of expressions. In Fig. 5.1, system architecture is illustrated in a block diagram. As shown in the figure, in landmark detection phase, the first frame is processed to obtain candidate points and detect 22 landmarks. These points are tracked in the other frames afterwards. As shown in the figure, spatiotemporal regions of interests are taken out from video sequences. Feature vector is attained by concatenating texture and depth features. In sparse coding phase, feature matrix is transformed to a sparse code matrix and pooling is applied in each ROI. Finally, the classifier is trained to estimate the probabilities of each expression for each test sequence. These steps are explained in detail in the following.

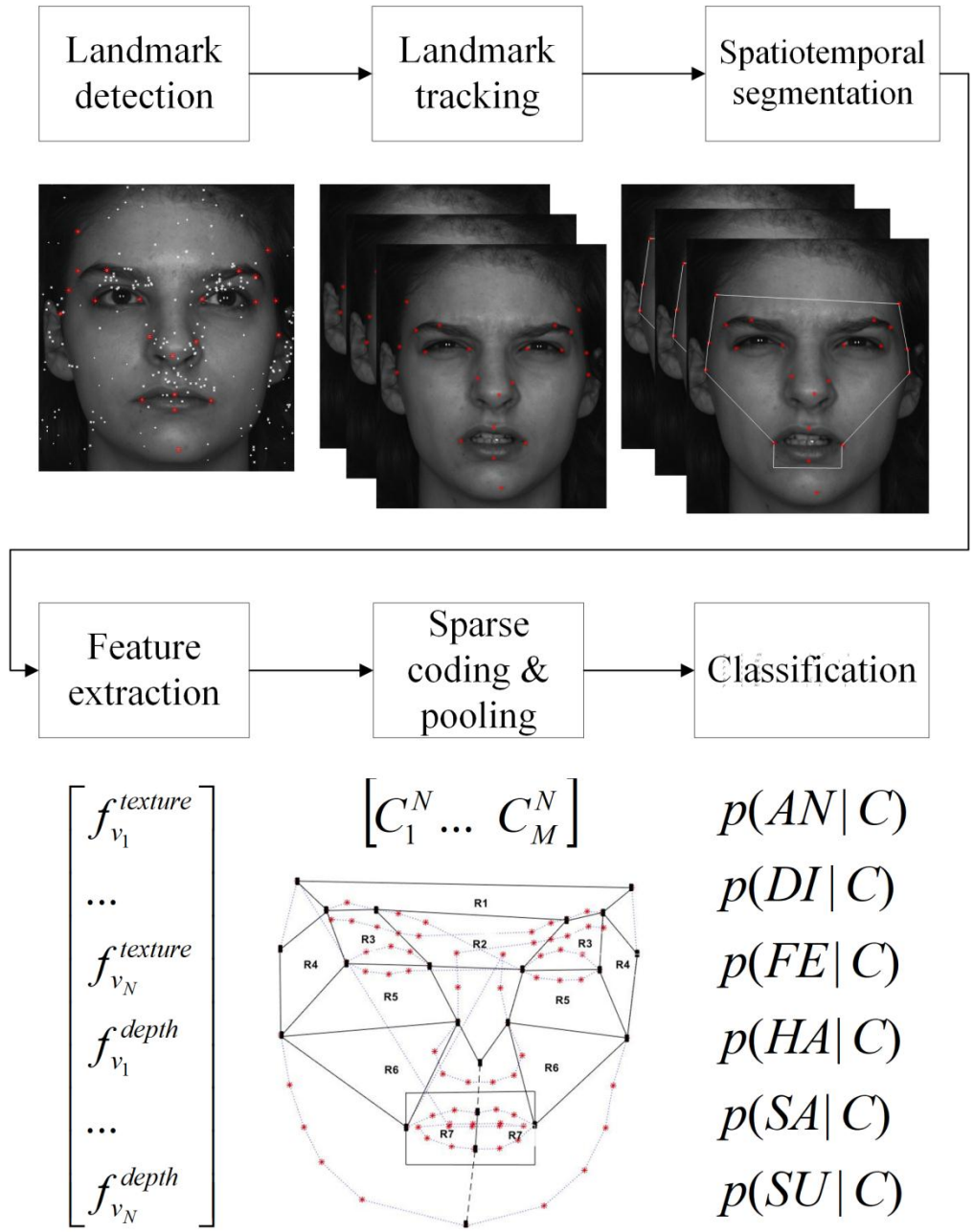


Figure 5.1: Architecture of the proposed method.

5.2.1 Landmark Detection

In this research, In facial expression recognition, landmark detection schemes have attracted the interest of many researchers [116], [117]. As mentioned previously, landmarks are firstly detected in the first frame of each texture sequence. It should be noted that by texture sequence we mean the RGB video sequence. In this stage, we employed the algorithm used in Lowe, 2004 [118] to detect candidate landmarks.

The landmarks are selected based on facial AUs as their corresponding spatial ROIs are representative for facial expressions. After detection of face bounding box, candidate points are identified using scale space extrema method. The scale space extrema can be detected using the Gaussian kernel function convolved with the input image. The description function *L(x, y)* of input image in different scale space is expressed as:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{5.1}$$

Where *L(x,y,σ)* is the spatial scale image, *I(x, y)* indicates input image of facial region, and *G(x,y,σ)* is the Gaussian convolution kernel function defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{5.2}$$

For each pixel located at *(x,y)* in an input image *I*, the difference of Gaussian (DoG) function in scale *σ* is computed as follows.

$$DoG(x, y, \sigma) = [G(x, y, k\sigma) - G(x, y, \sigma)] * I(x, y) \tag{5.3}$$

where *k* is a constant multiplier to change the scale of smoothness. In this study, 5 different scales are used resulting in 5 DoG images. In order to detect candidate point for landmark detection, scale space extrema are identified by comparing each pixel with its 8 surrounding pixels on the same scale, 9 closest pixels on one scale up and 9 on one scale down. If the pixel is a minimum or maximum among all point under comparison, it is considered as an extremum. Fig. 5.2 represents the procedure of obtaining DoG images and the extremum. The possible candidate point is shown in red and the 26 closest pixels used for comparison on the three levels are shown in blue.

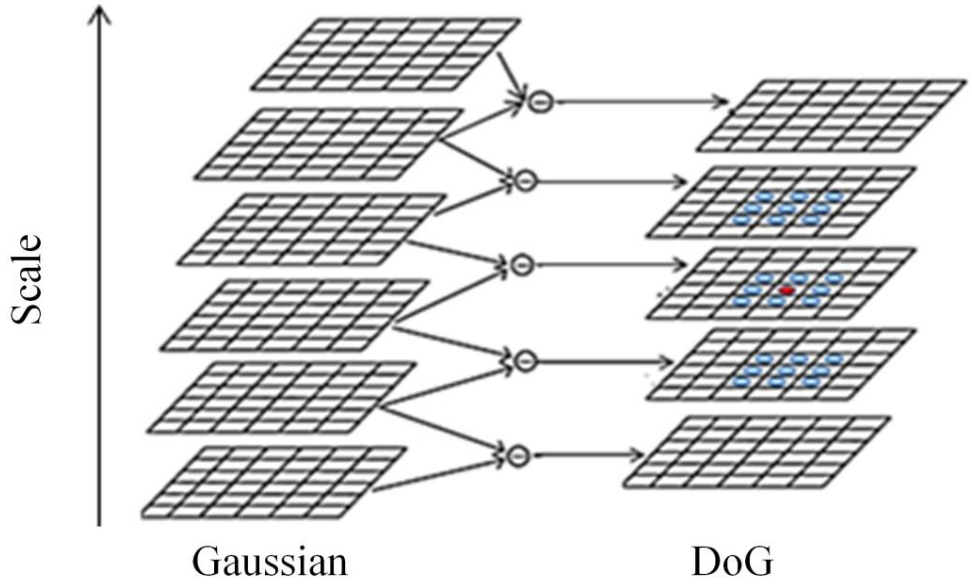


Figure 5.2: Obtaining DoG images and finding candidate points.

After candidate point selection, gradient magnitude and orientation histogram are used to extract features from candidate points as well as ground truth landmarks [119]. The gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ feature descriptors are computed as:

$$m(x,y) = \sqrt{[L(x+1,y) - L(x-1,y)]^2 + [L(x,y+1) - L(x,y-1)]^2} \qquad (5.4)$$

$$\theta(x,y) = \tan^{-1}\left[\frac{L(x+1,y)-L(x-1,y)}{L(x,y+1)-L(x,y-1)}\right] \qquad (5.5)$$

where L is the image at scale σ. To detect the landmarks from the interest candidate points, a set of landmark detectors with the feature description from the gradient orientation histogram of the input images are constructed. The descriptor is constructed from a vector containing the values of all the orientation histogram entries. At the center of each landmark, a neighborhood window is selected and divided into 16 sub regions of 4 × 4. Using Eq. 5.4 and 5.5, the directions and amplitudes of all pixels in the sub regions are obtained, and then accumulated into

orientation histograms summarizing the contents over the 4 × 4 sub regions. Using the orientation histogram, the eight direction distributions in the ranges of (0,π/4,π/2,3π/4,π,5π/4,3π/2,7π/4) is calculated with the length corresponding to the sum of the gradient magnitudes near that direction within the region. The amplitude and Gaussian function are also applied on the eight direction distributions to create the direction histogram of sub regions. The feature description of each landmark point is obtained by connecting the direction descriptions of all sub regions. The total number of the direction descriptions is 16 since we have 4 × 4 sub regions of the landmark descriptor. So the length of a landmark point detector is 128 = 16 × 8. These feature vectors are applied to Adaboost classifier which learns the locations of 22 landmarks. In the training stage, the first three frames of each sequence and their corresponding ground truth file is used. 22 desired landmarks are selected from 83 points provided in BU-4DFE data set and the detected nose tip. Each landmark and its 8 neighbors are assumed as positives, while 10 randomly selected points inside the face region are taken as negatives. In testing phase, candidate points are marked in the first frame and Adaboost which learnt the model, detects the landmarks among them.

5.2.2 Landmark Tracking

It is not computationally reasonable to detect landmarks in all frames. Instead, visual tracking methods such as mean shift, Kalman filter and particle filter are used. Mean shift is a simple tracker with an iterative algorithm aiming at minimizing the distance between the histogram of the target model and that of the candidate point [120] In mean shift tracker, motion information of the object is ignored. Kalman filter is a minimum variance estimator with Gaussian assumption [121]. For facial landmark

tracking, as the head moves nonlinearly, Kaman filter fails to track the points accurately. Unimodality and tracking delay are the other drawbacks of this approach.

In recent years, particle filters have attracted the attention of many researchers in visual point tracking. Particle filters are Bayesian filters without Gaussian assumption which propagate the particle estimation according to probability densities. Their capability fit into any state-space model even in presence of nonlinearity makes them efficient trackers in facial landmark detection. Particle filters are simple, robust and flexible but for multi-point tracking, conventional approach may break down to one point after some iterations [119] . It is due to the high dimensionality of the state variables and can be effectively addressed by resampling which discards the particles with small weights [122] . This modified version of particle filter is known as differential Evolution-Markov chain (DE-MC).

In this study, multiple tracker version of DE-MC particle filters is applied to track detected landmarks in other frames of the sequence [119] . In fact, each landmark point is modeled by a particle which is part of the parametric mixture model. For a given candidate point (observation), density probability function of the location of the landmark (target) is estimated by weighted particles. Details of this approach are demonstrated in [119]. As there are 22 landmarks in our model, the corresponding 22-component model over the state X_k is:

$$p(X_k|Y_{1:k}) = \sum_{j=1}^{22} Pi_{j,k} p_j(X_k|Y_{1:k}) \quad (5.6)$$

Where $p_j(X_k|Y_{1:k})$ is the posterior probability of the j^{th} landmark, $Y_{1:k}$ is the observation vector $Y_{1:k} = \{Y_1, Y_2, \dots, Y_k\}$ and Pi is the mixture weight. The motion model of facial landmarks is considered in learning phase to improve the

performance. Motion model is the probability of the X_k given the previous state X_{k-1} ($p(X_k|X_{k-1})$. By replacing $Y_{1:k}$ with $Y_{1:k-1}$ in Eq. 5.6, the predictive distribution becomes a function of previous states. Using total probability law:

$$p_j(X_k|Y_{1:k-1}) = \int p_j(X_k|X_{k-1})\, p_j(X_{k-1}|Y_{1:k-1}) dX_{k-1} \tag{5.7}$$

The measurement model is $p(Y_k|X_k)$ which the probability of the k^{th} observation is. Now, Eq. 5.6 can be restated as:

$$p(X_k|Y_{1:k}) = \lambda_k \sum_{j=1}^{22} Pi_{j,k} p_j(X_k|Y_k).$$

$$\int p_j(X_k|X_{k-1})\, p_j(X_{k-1}|Y_{1:k-1}) dX_{k-1} \tag{5.8}$$

Where λ_k is a constant value. In learning phase, positions of the landmarks are estimated using train data. Principally, particles are sampled from landmarks to estimate the correct distribution function. In tracking mode, for the k^{th} frame candidate particles are sampled from the appropriate distribution ($\hat{X}_k$) based on their weights computed via 22 estimated likelihood distributions. The weight of each sampled candidate is updated according to the following formula.

$$w_k = w_{k-1} \frac{p(Y_k|\hat{X}_k) p(\hat{X}_k|\hat{X}_{k-1})}{p(\hat{X}_k|X_{0:k-1}, Y_{1:k})} \tag{5.9}$$

Given automatically detected landmarks in the first frame, landmarks in consequent frames are detected by maximizing color-based observation likelihood.

5.2.3 Spatiotemporal Segmentation

After finding the position of landmarks in all frames, spatial and temporal ROIs are detected. In spatial domain, our suggested regions are characterized to adapt to AUs in FER literature as well as human reasoning about expressive regions. It is an established fact that facial AUs are mainly around eyes, eyebrows, lip, cheeks and lower forehead. In recent few years, finding the appropriate regions to compute

feature descriptors and how to combine descriptors have been studied in facial expression area[123], [124]. Zhalehpour et al. [125] have introduced the terms relevant and irrelevant sub-blocks in facial expression. In Fang et al. [56] feature descriptors are extracted from regions of cheeks, inner brows, eye outer corners and forehead. It has been suggested that local patches around landmarks are more efficient for feature extraction patches in FER than holistic patches [124]. Motivated by these studies, we defined 12 spatial ROIs in human face which. These regions are constructed based on 22 landmarks picked from 83 points of face model and nose tip. Our assumption is texture and depth information in these regions are fundamentally representative for expressions. Fig. 5.3 shows proposed landmarks and corresponding ROIs.

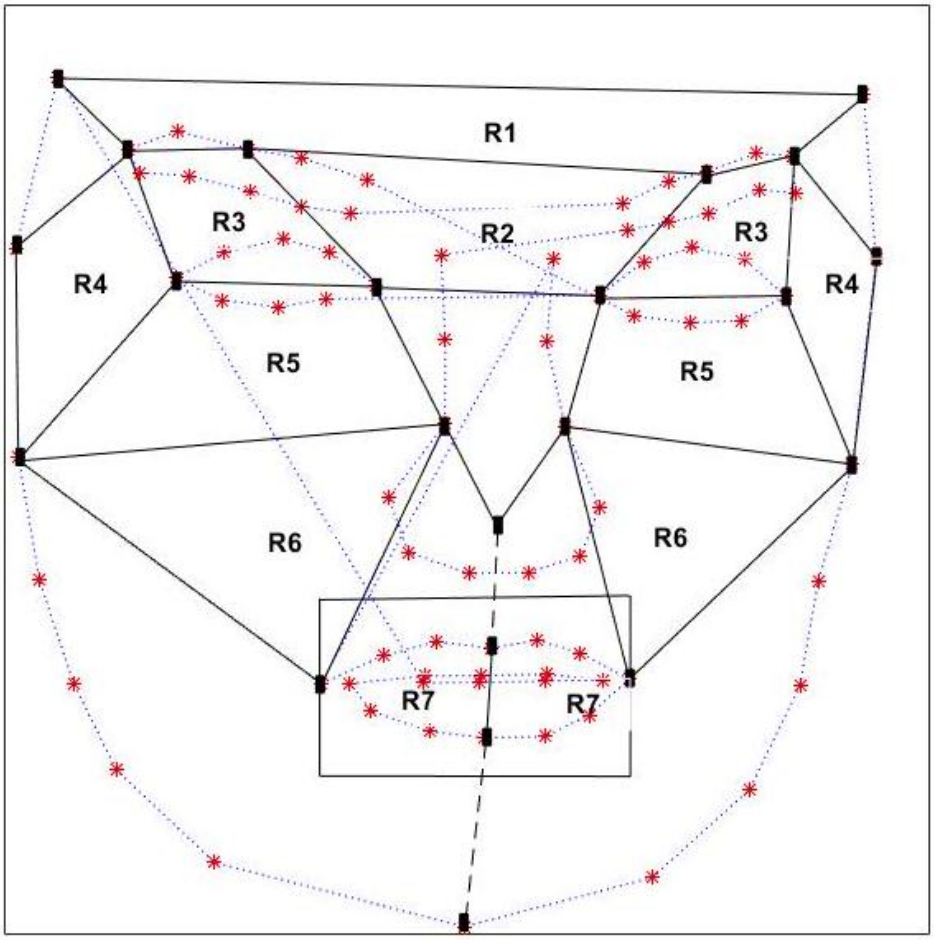


Figure 5.3: Facial landmarks and proposed ROIs (* 83 points of face model, ▪ 22 landmarks).

In addition to spatial ROIs, temporal segments of sequences are to be identified. Facial expressions comprise 4 phases including neutral, onset, apex and offset as shown in Fig. 5.4 for disgust expression. In the studies based on deformations and geometric features, apex part is taken for FER [112], [126] .However, temporal dynamics of expressions appear in onset phase. It has been proved that onset phase which contains transitions can be used efficiently for facial expression recognition [59]. In this study, onset phase is segmented from the whole sequence using a simple similarity measure based on mean Euclidean distance.

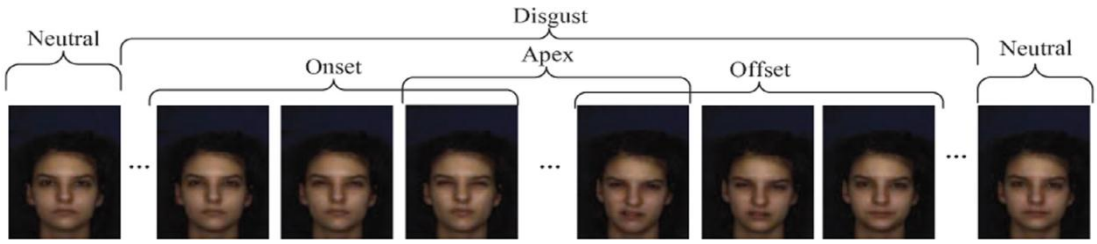


Figure 5.4: Different phases of disgust expression.

The first frame is taken as the reference and assumed to denote a neutral face (this assumption is true for majority of sequences in BU-4DFE data set). For each subsequent frame F_i, mean similarity measure in respect to first frame (F_1) is calculated using the coordinates of detected landmarks (x_j, y_j) as follows.

$$S(F_i, F_1) = 1 - \frac{\frac{1}{22}\sum_{j=1}^{22}\sqrt{(x_{i,j}-x_{1,j})^2+(y_{i,j}-y_{1,j})^2}}{\max_j\left(\sqrt{(x_{i,j}-x_{1,j})^2+(y_{i,j}-y_{1,j})^2}\right)} \tag{5.10}$$

After calculation of *S* over all frames of the sequence, a similarity curve is obtained. The frames in the largest decreasing line segment of the curve are supposed as the onset phase.

5.2.4 LBP-TOP Feature Extraction from Texture and Depth Videos

LBP are among the most effective spatial feature descriptors. As explained in Chapter 2, LBP is extracted from small patches by thersholding the pixels surrounding a central pixel. The size of the neighbor pixels depends on the number of surrounding points defined for descriptor extraction. The number of points can vary based on the radius of the circle enclosing the center point. In Fig. 5.5, the structure of the patches for 8, 16 and 24 points are illustrated and a sample 8-point LBP code is obtained to clarify the computation of this descriptor. The dimensionality of the LBP features is identified by the number of specific codes assigned to each pattern. Since LBP is rotation invariant, the dimensionality is predefined and can be learnt from LBP tables. For 8-point LBP descriptor, the dimensionality of the feature

descriptor vector is equal to 59. More precisely, 59 discriminant rotation-invariant LBP are possible.

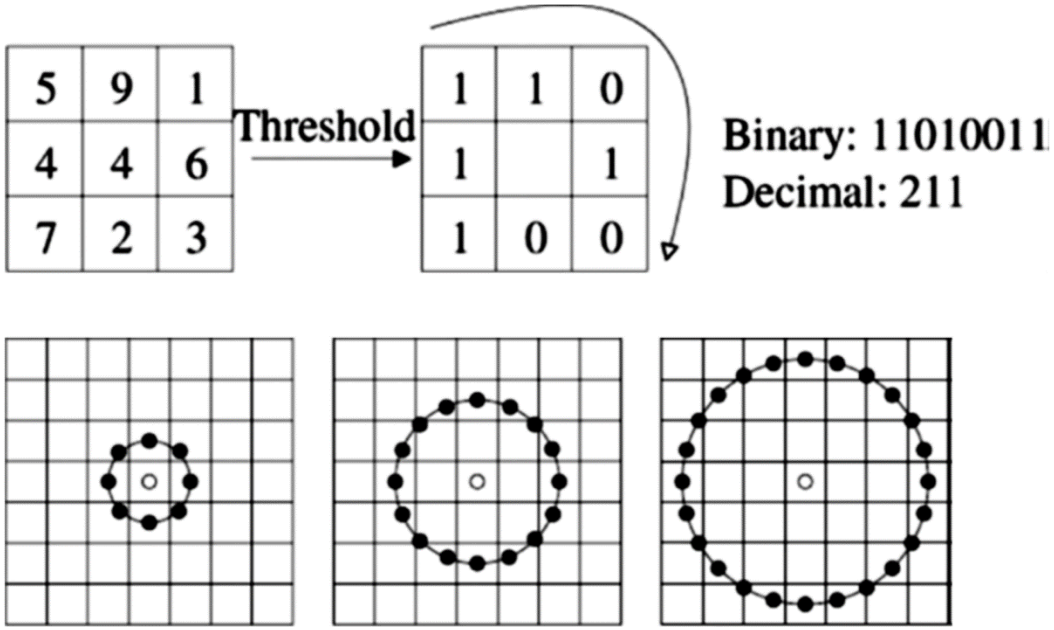


Figure 5.5: A sample 8-point LBP code (top) and the perspective of 8, 16 and 24 neighboring points (bottom).

In dynamic 3D facial expression recognition, time axis also exists. To adapt the LBP into dynamic image processing, variants of LBP feature descriptor such as volume local binary patterns (VLBP) and LBP-TOP [69] have been suggested. LBP-TOP features have been successfully utilized in D-FER systems [59], [69]. It reduces the computational time and the irrelevancy of the descriptors for FER classifier. The fundamental computation approach of LBP-TOP is similar to conventional LBP. However, features are extracted from three orthogonal planes namely, XY, XT and YT. Hence, instead of image patches, the blocks are named as cuboids. The schematic of 8-point LBP-TOP feature extraction is shown in Fig 5.6. It can be seen from the figure that in each plane of the cuboid, the LBP code is extracted and the histograms are computed on all the cuboids.

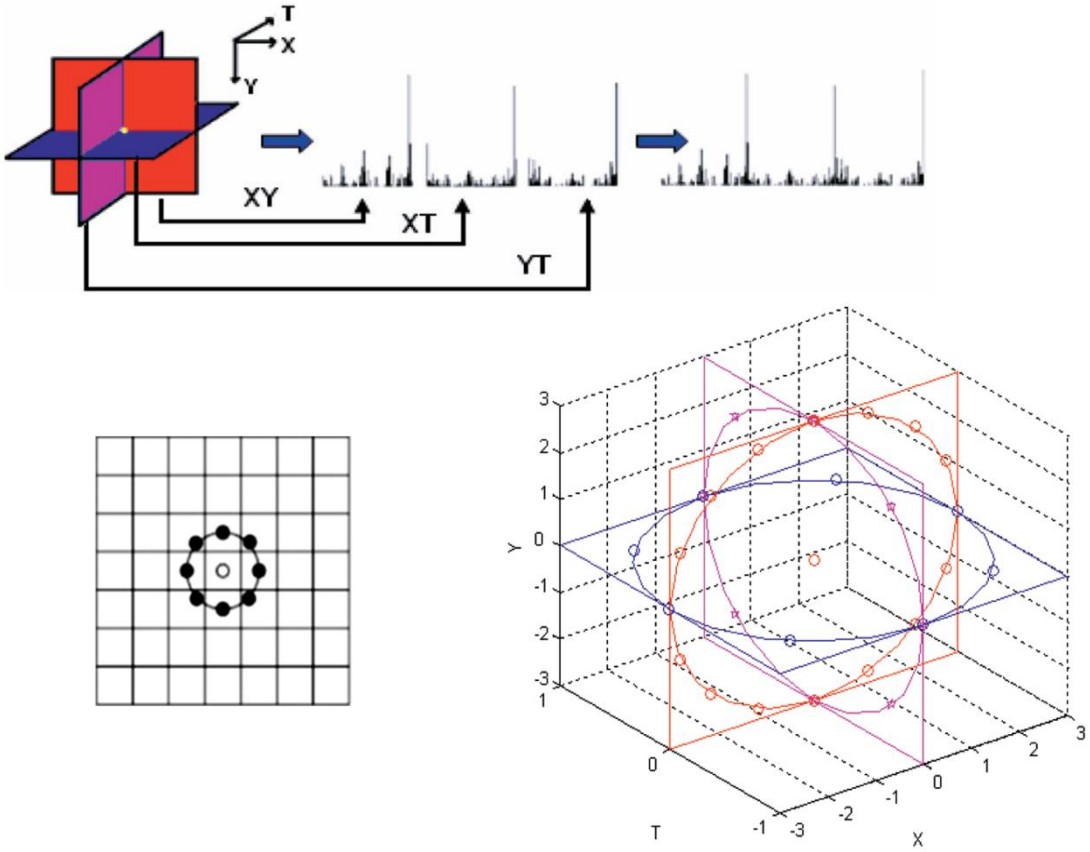


Figure 5.6: A schematic of 8-point LBP-TOP feature descriptor.

In feature extraction stage, the difference between our work and previous works is that feature descriptors are extracted from cuboids located in ROIs and not the whole face region. A simple preprocessing stage is applied in advance to compensate size variations and head rotation. Inner eye corners and nose point are used to align face and resize all frames. All pixels inside spatiotemporal ROIs (onset phase frames and 12 spatial regions) are considered in feature extraction phase. As LBP-TOP features are computed using 8 neighboring points, the dimension of histogram for each plane is 59 [69] It should be noted that, from this stage on, texture and depth videos are sequences of gray-scale images. The feature vector extracted from each cuboid is a 59*3 vector. Since feature descriptors are computed in texture and depth videos separately the descriptor matrix of a sequence is of size $N_s*\tau*(59*3*2)$ where N_s is the number of patches in spatial domain and τ is the number of cuboids in temporal

axis. For simplicity we assume $N=N_s * \tau$ and define the feature matrix of the sequence as:

$$F_v = \{(f_{v1}^{texture}, \dots, f_{vN}^{texture}), (f_{v1}^{depth}, \dots, f_{vN}^{depth})\} \quad (5.11)$$

Where $f_v \in R^{(59*3*2)}$ and it is defined in cuboid p as

$$f_v = \{LBP(x_p, y_p), LBP(x_p, t_p), LBP(y_p, t_p)\} \quad (5.12)$$

In the above formula, (x_p, y_p, t_p) are the spatiotemporal coordinates of the central pixel of the cuboid p. This phase of feature fusion is shown in Fig. 5.7. For detailed explanation of LBP-TOP feature descriptors one can refer to [69].

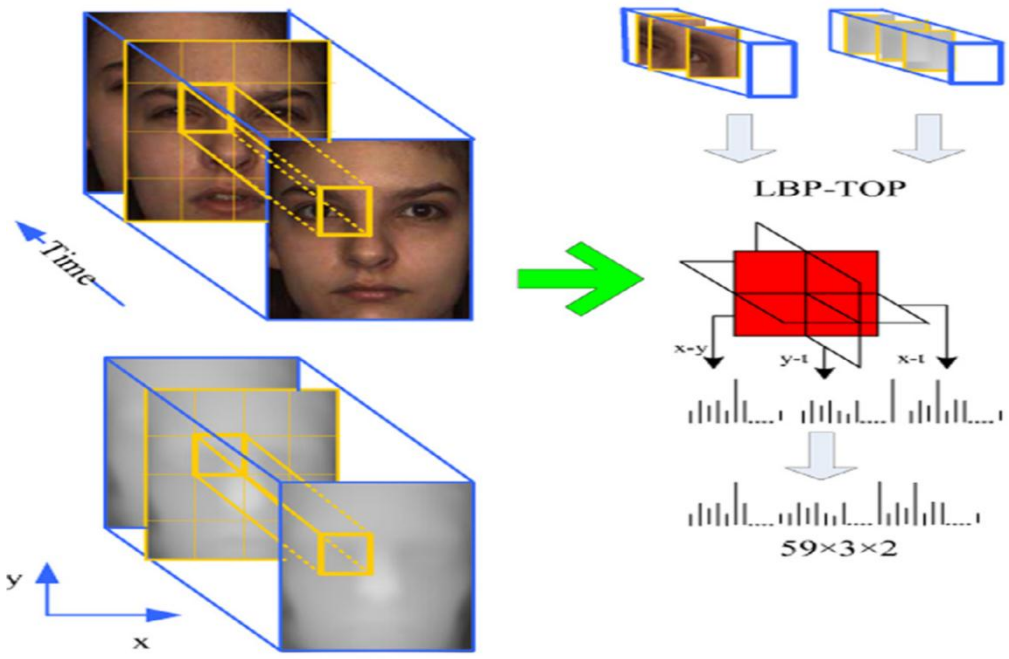


Figure 5.7: Fusion of texture and depth LBP-TOP feature descriptors.

5.2.5 Low-Rank Sparse Coding

Splitting tensile was Feature descriptors are dense and difficult for linear classifiers to be correctly classified. Sparse coding is a new approach to covert descriptor matrices into sparse matrices. Several algorithms have been proposed for sparse

coding in image classification including LLC [59], Laplacian sparse coding [127] and LRSC [128] .Codebook training for sparse coding is defined as an optimization problem with different constrains such as locality, sparsity and low-rankness. Most recently, low-rank sparse coding has attracted immense attention as it encourages both sparsity and spatial consistency. In other words, as the feature descriptors extracted from small neighboring regions have relevancies among them, the resulted sparse code should be low-rank [128] .By taking into account the underlying relevancies in LRSC, the codebook training phase converges rapidly, as the sparse low-rank optimization problem can be solved by a closed form update formula.

Assume that *F* is the feature descriptor, *D* is the codebook and *R* is the representation matrix. In LRSC, the optimization problem is defined as:

$$\min_Z \frac{1}{2}\|F - DR\|_n^2 + \lambda_1\|R\|_* + \lambda_2\|Z\|_{1,1} \quad (5.13)$$

λ_1 and λ_2 are regularization parameters to control sparsity and low-rankness. Comprehensive description of LRSC is presented in [79]. It has been proved that by transforming the form of the optimization problem, it can be solved by conventional Index Augmented Lagrange Multiplier. IALM is basically used for matrix rank minimization and makes it possible to obtain a closed update formula which results in time and computation simplicity compared with other approaches such as LLC [128]. Note that in closed form there is no need to set regularization parameters. Instead a used defined stop criteria (ε) is required to check the amount of changes in *B* after each iteration. Getting a codebook *D*, a feature descriptors matrix $F_v \in R^{N*(59*3*2)}$ is converted into sparse codes $C \in R^{N*M}$ as represented in Fig. 5.8.

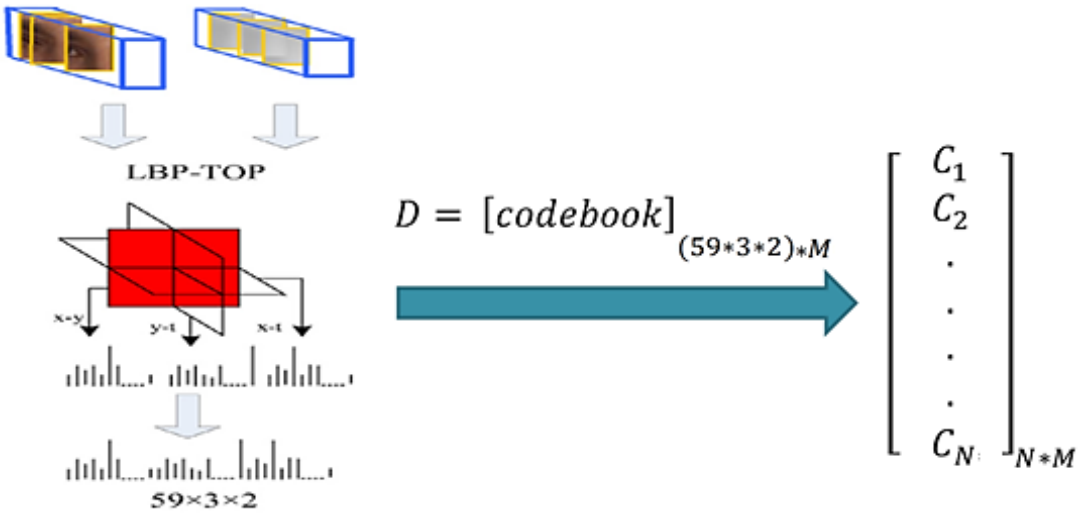


Figure 5.8: Low-rank sparse coding of LBP-TOP features.

5.2.6 Region of Interest Pooling

Basically in image processing systems, multiscale SPP is applied after coding to pool the codes in three local scales. However, in facial image processing systems such as FER systems, the local regions can be defined based on the face areas deformed during the expression. In this thesis, the notion of pooling is extended to spatiotemporal ROI pooling. In other words, maximum pooling in ROIs is then applied to sparse codes representing each sequence as a matrix of size $12*\tau*M$ where M is the length of the sparse vector, 12 is the number of ROIs and τ is the number of cuboids on temporal axis which depends (but not equal) on the number of frames in onset phase. SPP versus proposed ROI pooling is presented in Fig. 5.9.

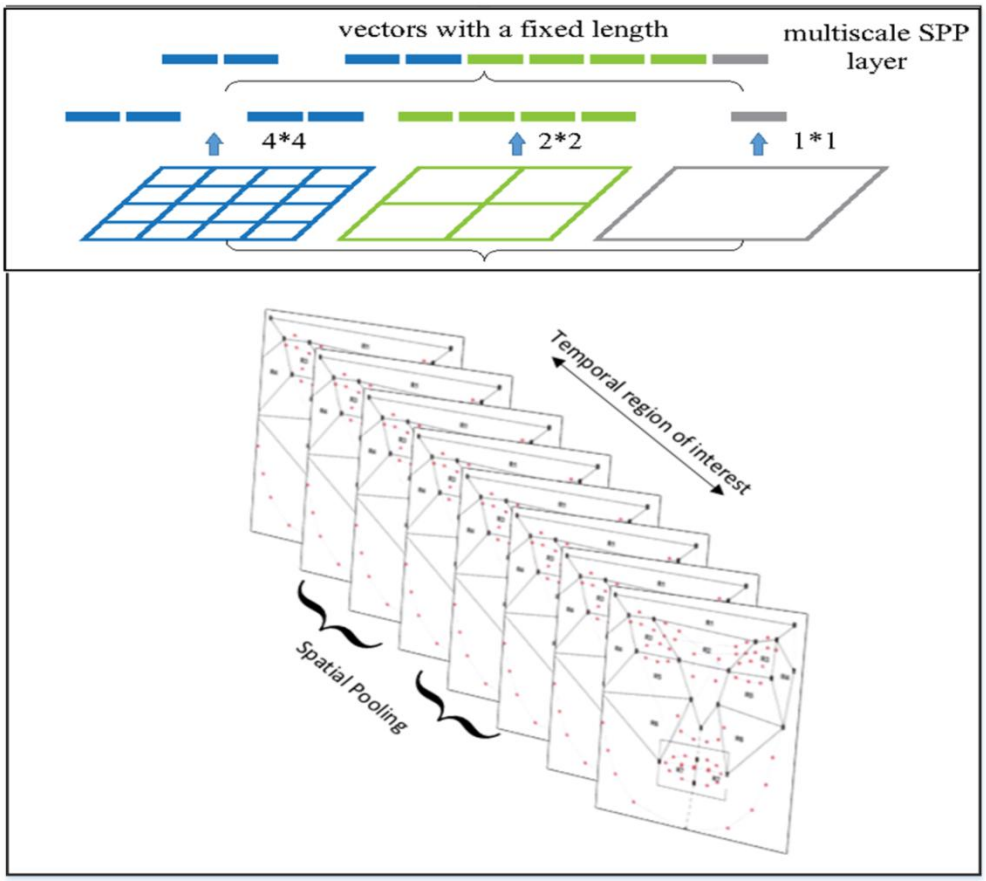


Figure 5.9: Multiscale SPP (top) versus ROI pooling (bottom).

5.2.7 Hidden-state Conditional Random Field Classifier

Conditional random field (CRF) have been extensively employed in multi-class image and video classification problems [129] Variations of CRF have been proposed and examined in image classification particularly in gesture and body motion recognition such as hidden-state conditional random fields (HCRF) [130] and latent dynamic conditional random field (LDCRF). Recently, there has been an increased interest in applying CRF for FER [59] While CRF assigns one label to each frame in a sequence, HCRF predicts one label for the whole video. In fact, the latter has more capability in modeling dynamics of temporal sequences. Moreover, by letting the model have hidden states which are trained simultaneously on multiple expressions, recognition accuracy improves significantly [130] .In this study, HCRFs

are utilized for performing multiclass classification problem of facial expression recognition. Assume that $X = \{x_1, x_2, \dots, x_m\}$ is a feature matrix of m local descriptors and y is the class label. HCRF estimates the conditional probability of a class label given an observation as:

$$p(y|X,\theta) = \sum_s p(y,s|X,\theta) = \frac{\sum_s e^{\psi(y,s,x;\theta)}}{\sum_{y' \in Y, s \in S} e^{\psi(y',s,x;\theta)}} \quad (5.14)$$

where $S \in R^m$ and specific underlying characteristics of each class is captures by a hidden state s. In potential function ψ, consistency between a y, X and a configuration of s is measured by parameter θ. Learning process is an optimization problem to find corresponding θ value which maximizes the following function. Basically, gradient decent algorithm is used for optimization.

$$L(\theta) = \sum_{i=1}^{n} \log p(y_i|x_i,\theta) - \frac{1}{2\sigma^2}\|\theta\|^2 \quad (5.15)$$

The first term is the log likelihood function and the second one is the logarithm of a Gaussian probability function.

5.3 Experimental Results

Proposed algorithm for dynamic 3D FER is examined on one of the most popular data sets in this field, BU-4DFE data set. This data set contains expression video sequences (RGB and 3D mesh) of 101 subjects each expressed 6 basic facial expressions including AN, DI, FE, HA, SA and SU. Each RGB and 3D mesh sequence (texture and depth) contains approximately 100 frames recorded at 25 frames per second. Resolution of texture and 3D model frames are about 1040×1329 pixels and 35,000 vertices respectively. Basically, researchers apply their experiments either on 360 video sequences of 60 subjects or all 600 video sequences from all 100 subjects. In Shao et al. (2015) [59] however, 95 subjects are selected for the experiments. Here, we conducted our experiments on 60, 100 and 95 subjects in

order to be able to make a reasonable comparison. Motivated by [59], low-resolution images of size 240*160 are used in all the experiments in this phase of the study.

Subject independent 10-CV is performed on the data sets. The ratio of test to train subjects are 6/54, 9/86 and 10/90 for 60-subject, 95-subject and 100-subject experiments respectively. Face bounding box in all frame of texture image is detected by Viola-Jones software. In landmark detection and tracking, besides 21 points picked from face model, nose tip is required which is simply extracted from depth frames assuming that images are frontal-view. The nose tip point is used to determine the top border of region 7 (R7) by connecting a line between this point and upper lip central point. The middle point of this line defines the topline for R7. Similarly, the line connecting chin landmark and lower lip center is used to identify the lower border of R7. More precisely, one third of this line length is assumed. These trends are to make sure R7 always encompasses lips. Head movement is normalized using inner eye corners and nose tip. More precisely, the line connecting inner eye corners is aligned horizontally and nose tip is positioned at the same location as the first frame.

After landmark detection and tracking, onset phase frames are segmented from the sequence using similarity curve. Fig. 5.3 illustrates similarity curves of one subject for different expressions. It can be seen that onset phase corresponds to the longest transition in the similarity measure when it decreases from maximum to minimum value. An example temporal segment for fear expression is annotated in the figure by vertical dashed lines. It should be noted that the dimensions of cuboids are selected empirically as 10*10 spatial patches on 3 frames on temporal axis. 8-point LBP-TOP features are extracted from a rectangular area including only the cuboids inside

spatiotemporal ROIs in a sequence. Codebook training parameters for LRSC are selected as $M = 1024$ (the size of the codebook) and $\varepsilon = 10^{-3}$ (stop parameter).

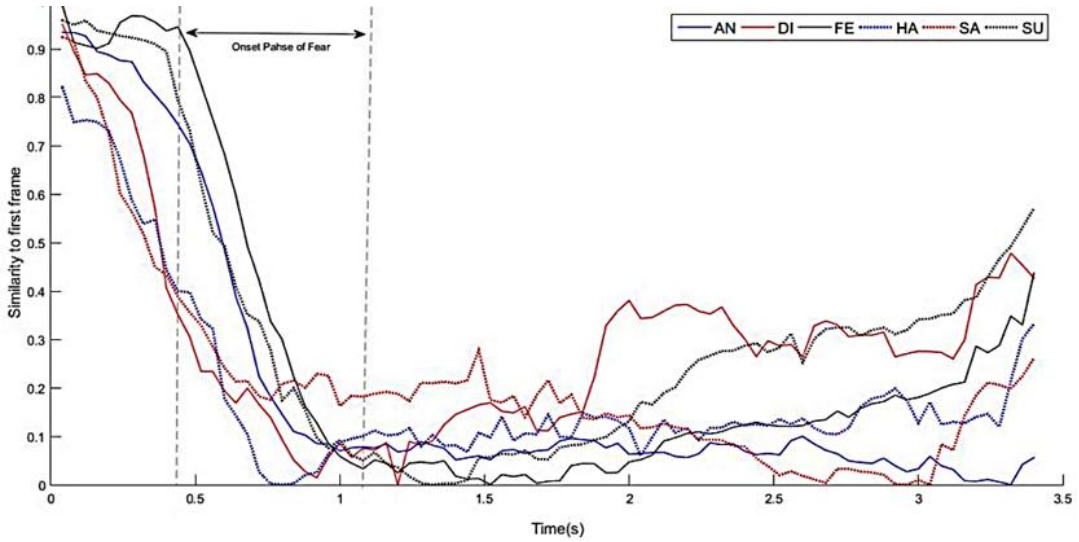


Figure 5.10: Similarity curves and onset phase of expressions.

Expression recognition confusion matrices for 60 and 100 subjects are represented in Table 5.1 and Table 5.2 respectively. Each column of the tables corresponds to the true expression and each row reports the classification result. It is observed that maximum recognition results are achieved for surprise expression in both cases while the results on fear expression are poor. This finding is consistent with previous studies in D-FER which are based on texture and depth sequences. As stated before, we also conducted experiments on 95 subjects and the average recognition accuracy is very close to what obtained on 100 subjects (85.09%).

Table 5.1: Confusion matrix on 60 subjects

Expression	Recognition Accuracy (%)					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	**88.33**	8.33	0.00	0.00	3.33	0.00
Disgust	3.33	**86.67**	5.00	1.67	1.67	1.67
Fear	3.33	6.67	**65.00**	13.33	1.67	10.00
Happy	0.00	0.00	0.00	**96.67**	3.33	0.00
Sadness	10.00	0.00	3.33	1.67	**85.00**	0.00
Surprise	0.00	0.00	1.67	0.00	0.00	**98.33**
Overall	**86.67**					

Table 5.2: Confusion matrix on 100 subjects

Expression	Recognition Accuracy (%)					
	Anger	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Anger	83.0	11.0	0.0	2.0	3.0	1.0
Disgust	4.0	82.0	8.0	4.0	0.0	2.0
Fear	5.0	7.0	67.0	9.0	2.0	10.0
Happy	0.0	0.0	1.0	95.0	4.0	0.0
Sadness	10.0	0.0	3.0	3.0	84.0	0.0
Surprise	0.0	1.0	1.0	0.0	0.0	98.0
Overall	**84.84**					

5.4 Discussion

The experiments are conducted on texture and depth sequences of BU-3DFE data sets. In fact, as the main motivation for this work is the study by Shao et al. [59], our first experiments are performed on 95 subjects to compare the recognition rate to [59]. The first contribution of this phase of the study is that LRSC is applied for the first time in dynamic 3D facial expression recognition. Secondly, the novel spatiotemporal regions of interest are constructed by landmarks which are automatically detected in the first frame and tracked using a multi-point tracker (DE-

MC) in the subsequent frames. Lastly, the region of interest pooling is introduced as an alternative to conventional SPP.

In order to compare recognition results with other studies, average recognition accuracies of recent papers conducted on BU-4DFE are reported in Table 5.3. Average accuracies on 100, 95 and 60 subjects are reported and compared to state-of-the-art. The results indicate that the proposed algorithm provides comparable and relatively superior results in comparison with previous studies considering that low-resolution videos are used in this work. More specifically, proposed system results in a significant improvement when conducted on the whole data set including 100 subjects. In addition, although Shao et al. have [59] selected their 95 subjects manually by discarding corrupted sequences and biased the results by this selection; proposed method outperforms their work on 95 randomly selected subjects with the same resolution. For 60 subjects, recognition results of the proposed system are higher than the accuracies reported in [61] , [63] and [86] nevertheless our experiments are performed on low-resolution videos. In fact, for practical applications, low-complexity systems based on low-quality and noisy videos are preferred. It is also worth mentioning that in [60], the original quality videos are processed. Furthermore, accurate nose tip detection is a crucial step that the proposed system in [60] relies on. In case of [76], the system is not automated and the high accuracy is achieved by using 83 landmark points annotated in the data set. Li et al. [110] has obtained 92.25% average accuracy using their deep learning neural network named as Dynamic Geometrical Image Network (DGIN). However, as mentioned it previous before, deep learning approaches are computationally costly, require large data sets and elaborate preprocessing stage.

Table 5.3: Comparison of proposed method with recent literature

Research Work	Method	Auto.	Classifier	ES*	ACC**
Reale et al. [111]	Spatiotemporal Volume + Nebula Feature	yes	SVM	100S, 6E, LOO	76.10
Fang et al.[71]	MeshHOG + LBP-TOP	yes	SVM	100S,6E,10-CV	75.82
Fang et al.[71]	Spin Image + LBP-TOP	yes	SVM	100S,6E,10-CV	74.63
Proposed Method	**LBP-TOP + Spatiotemporal ROI Pooling**	**yes**	**HCRF**	**100S,6E,10-CV**	**84.84**
Shao et al. [59]	LBP-TOP + SPP	yes	CRF	95S,6E,10-CV	83.07
Proposed Method	**LBP-TOP + Spatiotemporal ROI Pooling**	**yes**	**HCRF**	**95S,6E,10-CV**	**85.09**
Berretti et al. [61]	Pairwise Distance of 3D Landmarks and SIFT	yes	HMM	60S,6E,10-CV	79.40
Amor et al. [60]	Geometric 3D Motion Extraction	yes	LDA + HMM	60S,6E,10-CV	93.25
Kumar et al. [112]	Euclidian distance (geometric)	no	SVM	60S,6E,10-CV	93.06
Xue et al. [63]	3D-DCT + mRMR	yes	LDA + KNN	60S,6E,10-CV	78.80
Sandbach et al. [86]	3D motion-based Features	yes	GentleBoost + HMM	60S,6E,6-CV	64.60
Li et al. [110]	Geometric Images (DPI, NCI, SII)	yes	Deep Leraning	60S,6E,10-CV	92.22
Proposed Method	**LBP-TOP + Spatiotemporal Region of Interest**	**yes**	**HCRF**	**60S,6E,10-CV**	**86.67**

*Experimental Setting
**Accuracy (%)

5.5 Conclusion

In the third stage of this study, a 3D facial expression recognition method using LRSC pooled from automatically detected ROIs, is proposed. ROIs are defined using detected landmarks based on facial AUs, which are assumed to be more representative for the respective facial expressions than subject appearance. LBP-TOP feature descriptors are computed from cuboids inside spatiotemporal ROIs in both texture and depth sequences. Features extracted from texture and depth images are fused, where LRSC is utilized to obtain sparse codes from feature descriptors. In

the classification stage, HCRFs are used for the classification of six basic expressions. Experimental results have shown that the average accuracy of the proposed system in recognition of six basic expressions on 60 subjects and 100 subjects in BU-4DFE data set is equal to 86.67% and 84.84% respectively. These results verify that proposed algorithm improves the accuracy of D-FER in comparison to recent studies.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1 Comparison and Discussion of Proposed Methods

In this thesis, dynamic 3D facial expression recognition (D-FER) also known as 4D facial expression recognition is studied from two main perspectives. Generally speaking, there are two main streams in D-FER studies: facial landmark-based methods and appearance-based methods. In the preliminary phase, static FER is addressed using conventional distance features extracted from facial landmark locations. The aim is to find out if using an effective feature selection method and a well-developed classifier, geometric landmark-based feature result in an acceptable performance. The experiments conducted on BU-3DFE have confirmed that both sequential forward feature selection and SVM-FSVM classification improve the recognition rate. These findings provide the motivation for the second phase of the study, namely geometric landmark-based D-FER.

The proposed system is designed to address the dynamics of facial expression by extracting multimodal time series features from facial landmarks movements in the sequences. A comprehensive set of geometric deformations including point, distance and angle features are extracted from 3D coordinates of facial key points. The novelty of the method is applying time series analysis tools for feature extraction and classification of test sequences. As mentioned before, feature selection also plays a critical role in the performance of D-FER systems. Hence, a recently developed

feature selection method named as NCFS is utilized to address the dimensionality problem and to prune the redundant features. The selected classifier, AC-DTW is a new classification approach for time series analysis. A multivariate version of this approach has been adapted for our geometric landmark-based D-FER system. Experimental results proved that performance of the system is comparable to the state-of-the-art regardless of the low complexity of the system.

In order to complete the study, the effectiveness of facial landmark-based information in combination to appearance-based features is addressed in the third phase of the study. In other words, after verifying that a successful D-FER system can be designed relying only on facial landmarks, a fully automated system for low-quality videos is proposed in which the landmark information is dovetailed in an appearance-based system. This time, facial landmarks are utilized to define spatiotemporal regions of interest where coded LBP-TOP features are pooled from. The system is fully automated as the landmark locations are detected and tracked by the multi-point tracker. LBP-TOP features are extracted from the cuboids of low-quality texture and depth sequences. This phase of the work is compared to similar recent studies to evaluate the effectiveness. Although the performance of this approach is lower that our geometric landmark-based system, considering the practical applications of D-FER systems, the improvement is worthwhile. In fact, the proposed appearance-based D-FER system aims to have a closer look at the real world D-FER systems where the locations of landmarks are unknown and the quality of the registered images and videos is mainly low.

To have a general perspective of the study and its different phases, a summary table is represented in Table 6.1. In the rest of the chapter, we may refer to each proposed

method by its name for convenience. The reference systems are the baseline approaches that have been improved by suggested methods. It should be noted that even though static FER system is not directly comparable to D-FER systems, the results advocate the importance of facial landmarks in facial expression recognition problems providing motivation for the other proposed methods.

Table 6.1: A summary of the phases of study

System	Name	Method
FER	A	3D landmark distances + SFFS + SVM-FSVM
	Reference	3D landmark distances + SFFS + SVM
D-FER	B	Geometric landmark-based mean deformation time series + NCFS + AC-DTW
	Reference	Geometric landmark-based mean deformation time series + NCFS + DTW
	C	LBP-TOP of texture and depth + LRSC + ROI pooling + HCRF
	Reference*	LBP-TOP of texture and depth + LLC + SPP + CRF

*Conducted on 95 subjects

The proposed FER and D-FER systems are tested on BU-3DFE and BU-4DFE data sets respectively. For D-FER systems, the experiments are conducted on different number of subjects in order to be able to compare the recognition rates with state-of-the-art. The results of the proposed approaches on 100 subjects and their main reference systems are summarized in Table 6.2. Although, the maximum recognition accuracy and maximum improvement belongs to A FER system, there are many practical limitations in the system. In fact, BU-3DFE data set is a set of still images recorded at the peak of the expression. Each subject has expressed 4 different levels of each of the 6 basic expressions. The coordinates of 83 landmarks are available without any head pose rotation artifact. Furthermore, as stated before, the highest intensity level (level 4) images are used for this study. These conditions are far different from real world applications when there is no control over the head pose,

intensity and the peak of expression. However, this controlled registration procedure of BU-3DFE images potentials the high recognition accuracy of system A (87.67%) compared to B and C.

On the other hand, in D-FER systems which simulate a more realistic situation, the head rotation occurs within the sequences making the landmark coordinates less precise. Furthermore, unlike BU-3DFE data set, in BU-4DFE data set there is no control over the intensity of the expression across the subjects and the sequences contain both low and high intensity expressions. It is also worth mentioning the extra computational burden infused when scholars suggest identifying the peak expression from the video converting a D-FER system into an FER system. Hence, the recognition rate of D-FER systems designed based on practical limits is in general lower than that of FER systems.

Table 6.2: A summary of the results and comparisons

Scheme	Data Set	System	Automated	Accuracy (%)
FER	BU-3DFE	A	No	**87.67**
		Reference	No	78.33
D-FER	BU-4DFE	B	No	**83.50**
		Reference	No	69.83
		C	Yes	**84.84**
		Reference	Yes	83.07

Table 6.٢ confirms that proposed system based on AC-DTW classification of time series features extracted from mean deformations of facial landmarks performs significantly better than its conventional version using DTW classifier. More precisely, the recognition rate rises from 69.83% to 83.50%. The B system is

characterized by its simplicity and low-complexity. However, in real world problems when the resolution of the videos and images is limited, identifying the precise coordinates of facial landmarks is demanding. Therefore, the D-FER system can benefit from texture and depth information of the video sequences. In phase C, this problem is addressed by integration of facial landmark information into the pooling step and designing an automated recognition system. In other words, the coded LBP-TOP feature descriptors are pooled from the ROIs defined based on facial landmark points. The system performs successfully on 100 subjects and outperforms system B regardless of its low-resolution video input and automatic detection and tracking of landmarks. There is also improvement over the main reference study for system C.

6.2 Conclusions

Dynamic 3D facial expression recognition is an emerging topic having attracted the interest of researchers in computer vision and image processing. In this thesis, dynamic 3D facial expression recognition is studied from different aspects. The aim of the study is to uncover and address the main challenges in recognizing human emotions from 3-dimensional facial video sequences. This objective is fulfilled by analysis of the fundamental stages in general D-FER systems and exploring the issues and possible solutions. These basic stages include feature extraction, feature selection and classification. The work is performed in four main phases including literature review to comprehend the background, 3D facial expression recognition for a preliminary practical evaluation, geometric dynamic 3D facial expression recognition and non-geometric dynamic 3D facial expression recognition.

In the first phase of the thesis, previous studies on both static and dynamic FER systems are examined. The literature review section is outlined according to the

fundamental stages in FER systems. Since the difference between the FER and D-FER systems is basically in the feature extraction procedure, previous works in feature selection are assessed for static and dynamic systems individually. Two mainstreams in feature extraction i.e. geometric and non-geometric approaches are described. The feature descriptors adapted from still images to videos are explained and consequently the advantages and failings are disclosed.

In the second phase, a typical geometric 3D facial expression recognition system is implemented aiming at exploring the potential of landmark-based features by applying decent feature selection and classification methods. Pairwise distances between facial key points are extracted to construct the feature set. For selecting the useful subset of the features, SFFS based on Naive Bayesian classifier error is applied. Since SVM is a very popular classifier in FER studies, an improved multiclass version of SVM called fuzzy SVM is applied for classification. The experiments conducted on BU-3DFE data set have shown that in a system using conventional majority voting SVM, replacing the typical t-test feature selection with SFFS results in 6% improvement in average recognition accuracy of six basic expressions. Then by applying proposed FSVM classifier, average accuracy reaches from 72.33% to 87.67%. The results are motivating in the sense that by applying appropriate feature selection and classification methods, landmark-based features represent the discriminative facial deformations for facial expression recognition.

Accordingly, in the third phase of the thesis, a dynamic 3D facial expression recognition system based on time series analysis of geometric landmark based deformations is proposed. More precisely, time series features are computed from local temporal deformation of landmark locations to capture the dynamics of the

whole sequence. After head pose detection, correction and normalization of landmark positions, a comprehensive set of geometric deformations are extracted from 3D coordinates of facial landmarks. The point, distance and angle deformations are computed in each frame of the expression sequences with respect to the first frame. Then, time series features are constructed by applying a sliding window averaging on deformation values. In fact, the notion of time series analysis is introduced in D-FER systems for the first time to capture local mean dynamics. Since the dimensionality of feature space is high and there is a high level of correlation among features, a feature selection method named NCFS is applied to discard redundant features. NCFS is an effective and simple supervised embedded feature selection method suitable for high dimensional data with a large number of irrelevant features. As a result, it selects a small subset of useful geometric features for classification. Based on selected features, multimodal time series features are formed in reduced space. Finally, a recently introduced variant of DTW known as AC-DTW is utilized to classify these multivariate time series. Experiments conducted on BU-4DFE data set have resulted in 92.50% and 83.50% average recognition rate on 60 and 100 subjects respectively which confirm the effectiveness of the proposed system when compared to state-of-the-art.

Finally, in the last phase, an automatic dynamic 3D facial expression recognition system using LRSC and ROI pooling is proposed. Facial landmarks are automatically detected and tracked in texture sequences using candidate point's identification by scale space extrema, DoG features descriptors and Adaboost. ROIs are defined using detected landmarks based on facial AUs, which are assumed to be more representative for the respective facial expressions than subject appearance.

Appearance-based dynamic feature descriptors, namely LBP-TOP are computed from cuboids inside spatiotemporal ROIs in both texture and depth sequences. Texture and depth features descriptors are then fused and LRSC is utilized to obtain sparse codes. In the classification stage, HCRF is applied to recognize six basic expressions. Experimental results conducted on BU-4DFE data set have shown that the average accuracy of the proposed system on 60, 95 and 100 subjects is equal to 86.67%, 85.09% and 84.84 respectively. This verifies the proposed algorithm is efficient in comparison with recent studies.

In summary, this study has followed a comprehensive manner for dynamic 3D facial expression recognition in view of theoretical aspects and practical challenges. The problem is studied from different sides in three different phases and the existing methods are modified and improved to obtain higher recognition rate. The major practical limits are taken into account and at the same time the most recent approaches in machine learning, pattern recognition and image processing are adapted to tackle the problems.

6.2 Future Work

As a potential future work, similar to LBP-TOP and HOG-TOP novel spatiotemporal feature descriptors motivated by conventional feature descriptors may be adapted. Moreover, appearance-based and landmark-based features can be fused to construct the feature matrix for improved performance. For real-time applications, time and complexity limitations should be considered when implementing different phases of the system. Exploring the descriptors which perform even on low-quality videos, in presence of noise, occlusion, illumination variation and other real-world encounters are also challenging topics not having been addressed well in dynamic 3D facial

expression recognition. In addition, deep learning approaches which have attracted the interest of many researchers in recent years are another alternative. The promising results of CNN and other deep-learning neural networks can be applied in FER and D-FER systems to obtain higher recognition accuracy.

REFRENCES

[1] S. F. Yeganli, H. Demirel, and R. Yu, “Noise removal from MR images via iterative regularization based on higher-order singular value decomposition,” *Signal, Image Video Process.*, vol. 11, no. 8, pp. 1477–1484, 2017.

[2] A. Bolotnikova, H. Demirel, and G. Anbarjafari, “Real-time ensemble based face recognition system for NAO humanoids using local binary pattern,” *Analog Integr. Circuits Signal Process.*, vol. 92, no. 3, pp. 467–475, 2017.

[3] P. Zarbakhsh and H. Demirel, “Low-rank sparse coding and region of interest pooling for dynamic 3D facial expression recognition,” *Signal, Image Video Process.*, vol. 12, no. 8, pp. 1611–1618, 2018.

[4] I. Beheshti, H. Demirel, and H. Matsuda, “Classification of Alzheimer’s disease and prediction of mild cognitive impairment-to-Alzheimer’s conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm,” *Comput. Biol. Med.*, vol. 83, no. February, pp. 109–119, 2017.

[5] P. Bolourchi, H. Demirel, and S. Uysal, “Target recognition in SAR images using radial Chebyshev moments,” *Signal, Image Video Process.*, vol. 11, no. 6, pp. 1033–1040, 2017.

[6] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, “Emotion Recognition and Its Applications,” in *Advances in Intelligent*

Systems and Computing, vol. 300, 2014, pp. 51–62.

[7] N. T. CAO, A. H. TON-THAT, and H. IL CHOI, "Facial Expression Recognition Based on Local Binary Pattern Features and Support Vector Machine," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 06, p. 1456012, Sep. 2014.

[8] Z. Sun, Z.-P. Hu, M. Wang, F. Bai, and B. Sun, "Robust Facial Expression Recognition with Low-Rank Sparse Error Dictionary Based Probabilistic Collaborative Representation Classification," *Int. J. Artif. Intell. Tools*, vol. 26, no. 04, p. 1750017, Aug. 2017.

[9] Z. Wu, R. Xiamixiding, A. Sajjanhar, J. Chen, and Q. Wen, "Image Appearance-Based Facial Expression Recognition," *Int. J. Image Graph.*, vol. 18, no. 02, p. 1850012, Apr. 2018.

[10] S. Krig, "Interest Point Detector and Feature Descriptor Survey," in *Computer Vision Metrics*, Cham: Springer International Publishing, 2016, pp. 187–246.

[11] N. Alugupally, A. Samal, D. Marx, and S. Bhatia, "Analysis of landmarks in recognition of face expressions," *Pattern Recognit. Image Anal.*, vol. 21, no. 4, pp. 681–693, 2011.

[12] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing Action Units for Facial Expression Analysis," vol. 23, no. 2, pp. 1–19, 2001.

[13] H. Soyel and H. Demirel, “Optimal feature selection for 3D facial expression recognition using coarse-to-fine classification,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 18, no. 6, pp. 1031–1040, 2010.

[14] K. Yurtkan and H. Demirel, “Feature selection for improved 3D facial expression recognition,” *Pattern Recognit. Lett.*, vol. 38, no. 2, pp. 26–33, Mar. 2014.

[15] K. Chitta and N. N. Sajjan, “A reduced region of interest based approach for facial expression recognition from static images,” *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 2806–2809, 2017.

[16] D. Ghimire and J. Lee, “Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines,” *Sensors (Switzerland)*, vol. 13, no. 6, pp. 7714–7734, 2013.

[17] S. Berretti, B. Ben Amor, M. Daoudi, and A. Del Bimbo, “3D facial expression recognition using SIFT descriptors of automatically detected keypoints,” *Vis. Comput.*, vol. 27, no. 11, pp. 1021–1036, 2011.

[18] Y. Sun and L. Yin, “Facial Expression Recognition Based on 3D Dynamic Range Model Sequences,” 2008, pp. 58–71.

[19] R. Picard, *Affective Computing*. MIT Press, Cambridge, 1997.

[20] P. Zarbakhsh and H. Demirel, “Fuzzy SVM for 3D facial expression

classification using sequential forward feature selection," in *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2017, pp. 131–134.

[21] K. Yurtkan and H. Demirel, "Entropy-based feature selection for improved 3D facial expression recognition," *Signal, Image Video Process.*, vol. 8, no. 2, pp. 267–277, 2014.

[22] R. Shbib and S. Zhou, "Facial Expression Analysis using Active Shape Model," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 8, no. 1, pp. 9–22, Jan. 2015.

[23] Sujono and A. A. S. Gunawan, "Face Expression Detection on Kinect Using Active Appearance Model and Fuzzy Logic," *Procedia Comput. Sci.*, vol. 59, no. Iccsci, pp. 268–274, 2015.

[24] P. Ekman, W. Friesen, J. Hager, "Facial Action Coding System (FACS)," *A Hum. Face*, 2002.

[25] E. N. Arcoverde Neto *et al.*, "Enhanced real-time head pose estimation system for mobile device," *Integr. Comput. Aided. Eng.*, vol. 21, no. 3, pp. 281–293, Apr. 2014.

[26] J. J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Detection, tracking, and classification of subtle changes in facial expression," *J. Robot. Auton. Syst.*, vol. 31, pp. 131–146, 2000.

[27] D. Derkach, A. Ruiz, and F. M. Sukno, “Head Pose Estimation Based on 3-D Facial Landmarks Localization and Regression,” *Proc. - 12th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2017 - 1st Int. Work. Adapt. Shot Learn. Gesture Underst. Prod. ASL4GUP 2017, Biometrics Wild, Bwild 2017, Heteroge*, pp. 820–827, 2017.

[28] S. L. Happy and A. Routray, “Automatic facial expression recognition using features of salient facial patches,” *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, 2015.

[29] H. Patil, A. Kothari, and K. Bhurchandi, “3-D face recognition: features, databases, algorithms and challenges,” *Artif. Intell. Rev.*, vol. 44, no. 3, pp. 393–441, 2015.

[30] F. R. Al-Osaimi, M. Bennamoun, and A. Mian, “Spatially optimized data-level fusion of texture and shape for face recognition,” *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 859–872, 2012.

[31] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, “3D face discriminant analysis using gauss-markov posterior marginals,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 728–739, 2013.

[32] H. Mahersia and K. Hamrouni, “Using multiple steerable filters and Bayesian regularization for facial expression recognition,” *Eng. Appl. Artif. Intell.*, vol. 38, pp. 190–202, 2015.

[33] M. Ilbeygi and H. Shah-Hosseini, “A novel fuzzy facial expression recognition system based on facial feature extraction from color face images,” *Eng. Appl. Artif. Intell.*, vol. 25, no. 1, pp. 130–146, 2012.

[34] A. Chakrabarty, H. Jain, and A. Chatterjee, “Volterra kernel based face recognition using artificial bee colonyoptimization,” *Eng. Appl. Artif. Intell.*, vol. 26, no. 3, pp. 1107–1114, 2013.

[35] J. Wang, L. Yin, X. Wei, and Y. Sun, “3D facial expression recognition based on primitive surface feature distribution,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 1399–1406, 2006.

[36] K. W. Bowyer, K. Chang, and P. Flynn, “A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition,” *Comput. Vis. Image Underst.*, vol. 101, no. 1, pp. 1–15, 2006.

[37] X. Zhao, E. Dellandréa, L. Chen, and I. A. Kakadiaris, “Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 41, no. 5, pp. 1417–1428, 2011.

[38] H. Tabia, M. Daoudi, J. P. Vandeborre, and O. Colot, “A new 3D-matching method of nonrigid and partially similar models using curve analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 852–858, 2011.

[39] H. Mohammadzade and D. Hatzinakos, “Iterative closest normal point for 3D face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 381–397, 2013.

[40] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,” *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[41] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, “Samkah Mn Zahb.Pdf.”

[42] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, “Facial expression recognition and histograms of oriented gradients: a comprehensive study,” *Springerplus*, vol. 4, no. 1, 2015.

[43] F. Ahmed, H. Bari, and E. Hossain, “Person-independent facial expression recognition based on Compound Local Binary Pattern (CLBP),” *Int. Arab J. Inf. Technol.*, vol. 11, no. 2, pp. 195–203, 2014.

[44] F. Bashar, A. Khan, F. Ahmed, and M. H. Kabir, “Robust facial expression recognition based on median ternary pattern (MTP),” *2013 Int. Conf. Electr. Inf. Commun. Technol. EICT 2013*, pp. 1–5, 2013.

[45] S. S. Meher and P. Maben, “Face recognition and facial expression identification using PCA,” *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, pp. 1093–1098, 2014.

[46] P. Marasamy and S. Sumathi, “Automatic recognition and analysis of human faces and facial expression by LDA using wavelet transform,” *2012 Int. Conf. Comput. Commun. Informatics, ICCCI 2012*, pp. 1–4, 2012.

[47] S. A. Al-agha, H. H. Saleh, and R. F. Ghani, “Geometric-based Feature Extraction and Classification for Emotion Expressions of 3D Video Film,” *J. Adv. Inf. Technol.*, vol. 8, no. 2, pp. 74–79, 2017.

[48] J. de A. Fernandes, L. N. Matos, and M. G. dos S. Aragao, “Geometrical Approaches for Facial Expression Recognition Using Support Vector Machines,” in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 347–354.

[49] L. Gang, L. Xiao-hua, Z. Ji-liu, and G. Xiao-gang, “Geometric feature based facial expression recognition using multiclass support vector machines,” *2009 IEEE Int. Conf. Granul. Comput. GRC 2009*, pp. 318–321, 2009.

[50] H. Soyel and H. Demirel, “3D facial expression recognition with geometrically localized facial features,” *2008 23rd Int. Symp. Comput. Inf. Sci. Isc. 2008*, pp. 1–4, 2008.

[51] J. Chen, Z. Chen, Z. Chi, and H. Fu, “Facial Expression Recognition Based on Facial Components Detection and HOG Features,” no. August, pp. 64–69, 2014.

[52] M. M. F. Donia, A. A. A. Youssif, and A. Hashad, “Spontaneous Facial

Expression Recognition Based on Histogram of Oriented Gradients Descriptor," *Comput. Inf. Sci.*, vol. 7, no. 3, pp. 31–37, 2014.

[53] L. Wang, R. Li, and K. Wang, "A Novel Automatic Facial Expression Recognition Method Based on AAM," *J. Comput.*, vol. 9, no. 3, pp. 608–617, 2014.

[54] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 40, no. 3, pp. 461–474, 2010.

[55] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A High-Resolution 3D Dynamic Facial Expression Database State University of New York at Binghamton High-Resolution Data Acquisition," *Autom. Face Gesture Recognition, 2008. FG '08. 8th IEEE Int. Conf.*, no. 1, pp. 1–6, 2008.

[56] H. Fang *et al.*, "Facial expression recognition in dynamic sequences: An integrated approach," *Pattern Recognit.*, vol. 47, no. 3, pp. 1271–1281, 2014.

[57] Y. Guo, G. Zhao, and M. Pietikainen, "Dynamic Facial Expression Recognition with Atlas Construction and Sparse Representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1977–1992, 2016.

[58] S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences," *Pattern Recognit. Lett.*, vol. 92, pp.

25–32, 2017.

[59] J. Shao, I. Gori, S. Wan, and J. K. Aggarwal, “3D dynamic facial expression recognition using low-resolution videos,” *Pattern Recognit. Lett.*, vol. 65, pp. 157–162, Nov. 2015.

[60] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, “4-D facial expression recognition by learning geometric deformations,” *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2443–2457, 2014.

[61] S. Berretti, A. Del Bimbo, and P. Pala, “Automatic facial expression recognition in real-time from dynamic sequences of 3D face scans,” *Vis. Comput.*, vol. 29, no. 12, pp. 1333–1350, 2013.

[62] V. P. Kalyan Kumar, P. Suja, and S. Tripathi, “Emotion Recognition from Facial Expressions for 4D Videos Using Geometric Approach,” 2016, pp. 3–14.

[63] M. Xue, A. Mian, W. Liu, and L. Li, “Automatic 4D facial expression recognition using DCT features,” *Proc. - 2015 IEEE Winter Conf. Appl. Comput. Vision, WACV 2015*, pp. 199–206, 2015.

[64] J. Liang, C. Xu, Z. Feng, and X. Ma, “Hidden Markov Model Decision Forest for Dynamic Facial Expression Recognition,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 29, no. 07, p. 1556010, Nov. 2015.

[65] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Recognizing facial expression: Machine learning and application to spontaneous behavior,” *Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005*, vol. II, pp. 568–573, 2005.

[66] Y. Yao, D. Huang, X. Yang, Y. Wang, and L. Chen, “Texture and Geometry Scattering Representation-Based Facial Expression Recognition in 2D+3D Videos,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 1s, pp. 1–23, 2018.

[67] J. Chen, Z. Chen, Z. Chi, and H. Fu, “Facial Expression Recognition in Video with Multiple Feature Fusion,” *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 38–50, 2018.

[68] M. Liu, S. Shan, R. Wang, and X. Chen, “Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1749–1756, 2014.

[69] G. Zhao and M. Pietikainen, “Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[70] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, “4D facial expression recognition,” *2011 IEEE Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pp. 1594–1601, 2011.

[71] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, “3D/4D facial expression analysis: An advanced annotated face model approach,” *Image Vis. Comput.*, vol. 30, no. 10, pp. 738–749, 2012.

[72] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, “A dynamic approach to the recognition of 3D facial expressions and their temporal models,” *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 406–413, 2011.

[73] Q. Zhen, D. Huang, H. Drira, B. Ben Amor, Y. Wang, and M. Daoudi, “Magnifying Subtle Facial Motions for Effective 4D Expression Recognition,” *IEEE Trans. Affect. Comput.*, pp. 1–1, 2017.

[74] N. Kaur and E. V. Kaur, “Improved Emotion Detection by Regression Algorithm with SURF Feature and SVM,” *Int. J. Eng. Comput. Sci.*, vol. 3, no. 10, pp. 8639–8642, 2014.

[75] A. Majumder, L. Behera, and V. K. Subramanian, “Local binary pattern based facial expression recognition using Self-organizing Map,” *Proc. Int. Jt. Conf. Neural Networks*, pp. 2375–2382, 2014.

[76] H. Soyel, U. Tekguc, and H. Demirel, “Application of NSGA-II to feature selection for facial expression recognition,” *Comput. Electr. Eng.*, vol. 37, no. 6, pp. 1232–1240, 2011.

[77] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained

Linear Coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.

[78] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely – Laplacian sparse coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3555–3561.

[79] L. Zhang and C. Ma, "Low-rank, sparse matrix decomposition and group sparse coding for image classification," *Proc. - Int. Conf. Image Process. ICIP*, pp. 669–672, 2012.

[80] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "PCA-based dictionary building for accurate facial expression recognition via sparse representation," *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1082–1092, 2014.

[81] L. Li, Z. Ying, and T. Yang, "Facial expression recognition by fusion of gabor texture features and local phase quantization," *Int. Conf. Signal Process. Proceedings, ICSP*, vol. 2015-Janua, no. October, pp. 1781–1784, 2014.

[82] Q. W. Wang and Z. L. Ying, "Facial Expression Recognition Algorithm Based on Gabor Texture Features and Adaboost Feature Selection via Sparse Representation," *Appl. Mech. Mater.*, vol. 511–512, pp. 433–436, Feb. 2014.

[83] H. Tang and T. S. Huang, "3D Facial expression recognition based on automatically selected features," *2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work.*, pp. 1–8, 2008.

[84] S. Canavan, Y. Sun, X. Zhang, and L. Yin, "A dynamic curvature based approach for facial activity analysis in 3D space," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 14–19, 2012.

[85] X. Zhang *et al.*, "A high-resolution spontaneous 3D dynamic facial expression database," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[86] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image Vis. Comput.*, vol. 30, no. 10, pp. 762–773, 2012.

[87] M. Rosato, X. Chen, and L. Yin, "Automatic registration of vertex correspondences for 3D facial expression analysis," *BTAS 2008 - IEEE 2nd Int. Conf. Biometrics Theory, Appl. Syst.*, 2008.

[88] F. Dornaika, A. Moujahid, and B. Raducanu, "Facial expression recognition using tracked facial actions: Classifier performance analysis," *Eng. Appl. Artif. Intell.*, vol. 26, no. 1, pp. 467–477, 2013.

[89] L. Yin, X. Wei, P. Longo, and A. Bhuvanesh, "Analyzing facial expressions using intensity-variant 3D data for human computer interaction," *Proc. - Int.*

Conf. Pattern Recognit., vol. 1, pp. 1248–1251, 2006.

[90] M.-C. Su, C.-K. Yang, S.-C. Lin, D.-Y. Huang, Y.-Z. Hsieh, and P.-C. Wang, "An SOM-based Automatic Facial Expression Recognition System," *Int. J. Soft Comput. Artif. Intell. Appl.*, vol. 2, no. 4, pp. 45–57, Aug. 2013.

[91] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, 2019.

[92] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[93] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Trans. Multimed.*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[94] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition."

[95] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *Vis. Comput.*, 2019.

[96] F. An and Z. Liu, “Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM,” *Vis. Comput.*, 2019.

[97] N. Sebe, I. Cohen, A. Garg, and T. Huang, *Facial Expression Recognition*. Machine Learning in Computer Vision, 2005.

[98] C. Zhan, W. Li, P. Ogunbona, and F. Safaei, “A Real-Time Facial Expression Recognition System for Online Games,” *Int. J. Comput. Games Technol.*, vol. 2008, pp. 1–7, 2008.

[99] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato, “A 3D Facial Expression Database For Facial Behavior Research,” in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 211–216.

[100] Chun-Fu Lin and Sheng-De Wang, “Fuzzy support vector machines,” *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.

[101] S. Abe, “Fuzzy support vector machines for multilabel classification,” *Pattern Recognit.*, vol. 48, no. 6, pp. 2110–2117, 2015.

[102] H. Tang and T. S. Huang, “3D facial expression recognition based on properties of line segments connecting facial feature points,” in *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, 2008.

[103] O. K. Oyedotun, G. Demisse, A. El Rahman Shabayek, D. Aouada, and B. Ottersten, “Facial expression recognition via joint deep learning of RGB-Depth map latent representations,” in *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018, vol. 2018-January, pp. 3161–3168.

[104] M. F. Valstar and M. Pantic, “Fully automatic recognition of the temporal phases of facial actions,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 42, no. 1, pp. 28–43, 2012.

[105] K. M. Goh, C. H. Ng, L. L. Lim, and U. U. Sheikh, “Micro-expression recognition: an updated review of current trends, challenges and solutions,” *Vis. Comput.*, 2018.

[106] W. Yang, K. Wang, and W. Zuo, “Neighborhood component feature selection for high-dimensional data,” *J. Comput.*, vol. 7, no. 1, pp. 162–168, 2012.

[107] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, “Mapping the emotional face. How individual face parts contribute to successful emotion recognition,” *PLoS One*, vol. 12, no. 5, pp. 1–15, 2017.

[108] Y. Wan, X. L. Chen, and Y. Shi, “Adaptive cost dynamic time warping distance in time series analysis for classification,” *J. Comput. Appl. Math.*, vol. 319, pp. 514–520, 2017.

[109] Q. Zhen, D. Huang, Y. Wang, and L. Chen, “Muscular Movement Model-

Based Automatic 3D/4D Facial Expression Recognition," *IEEE Trans. Multimed.*, vol. 18, no. 7, pp. 1438–1450, 2016.

[110] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4D facial expression recognition using dynamic geometrical image network," *Proc. - 13th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2018*, pp. 24–30, 2018.

[111] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis," *2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, pp. 1–8, 2013.

[112] V. P. Kalyan Kumar, P. Suja, and S. Tripathi, "Emotion Recognition from Facial Expressions for 4D Videos Using Geometric Approach," 2016, pp. 3–14.

[113] H. Soyel and H. Demirel, "Improved SIFT matching for pose robust facial expression recognition," *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 585–590, 2011.

[114] U. Mlakar and B. Potočnik, "Automated facial expression recognition based on histograms of oriented gradient feature vector differences," *Signal, Image Video Process.*, vol. 9, pp. 245–253, 2015.

[115] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained Linear Coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3360–

3367.

[116] N. Rathee and D. Ganotra, “An efficient approach for facial action unit intensity detection using distance metric learning based on cosine similarity,” *Signal, Image Video Process.*, vol. 12, no. 6, pp. 1141–1148, 2018.

[117] P. Bian, Z. Xie, and Y. Jin, “Multi-task feature learning-based improved supervised descent method for facial landmark detection,” *Signal, Image Video Process.*, vol. 12, no. 1, pp. 17–24, 2018.

[118] D. G. Lowe, “Distinctive Image Features from,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[119] Y. Tie and L. Guan, “Automatic landmark point detection and tracking for human facial expressions,” *Eurasip J. Image Video Process.*, vol. 2013, pp. 1–15, 2013.

[120] Y. Qi, C. Wu, D. Chen, and X. Yu, “Robust object tracking with multiple basic mean shift tracker,” *2012 IEEE Int. Conf. Robot. Biomimetics, ROBIO 2012 - Conf. Dig.*, no. 1, pp. 2300–2305, 2012.

[121] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. Wiley-Interscience, 2006.

[122] M. Du and L. Guan, “MONOCULAR HUMAN MOTION TRACKING WITH THE DE-MC PARTICLE FILTER Ryerson Multimedia Research Lab

, Dept . of Electrical and Computer Engineering Ryerson University , Toronto , Canada," no. Mcmc, pp. 205–208, 2006.

[123] S. Jaiswal, B. Martinez, and M. F. Valstar, "Learning to combine local models for facial Action Unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, vol. 06, pp. 1–6.

[124] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," *Proc. - Int. Conf. Pattern Recognit.*, pp. 1776–1781, 2014.

[125] S. Zhalehpour, Z. Akhtar, and C. Eroglu Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image Video Process.*, vol. 10, no. 5, pp. 827–834, 2016.

[126] S. K. A. Kamarol, M. H. Jaward, R. Parthiban, and S. K. A. Kamarol, "Spatiotemporal feature extraction for facial expression recognition," *IET Image Process.*, vol. 10, no. 7, pp. 534–541, Jul. 2016.

[127] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely Laplacian sparse coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3555–3561.

[128] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment

by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, 2012.

[129] Sy Bor Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden Conditional Random Fields for Gesture Recognition," *2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Vol. 2*, vol. 2, pp. 1521–1527, 2006.

[130] C. Chen, J. Zhang, and Z. Gan, "Human action recognition based on latent-dynamic Conditional Random Field," *2013 Int. Conf. Wirel. Commun. Signal Process. WCSP 2013*, pp. 1–5, 2013.