FACIAL IMAGE ANALYSIS AND ITS APPLICATIONS TO FACIAL EXPRESSION RECOGNITION

CIGDEM TURAN

PhD

The Hong Kong Polytechnic University

2019

THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF ELECTRONIC AND INFORMATION ENGINEERING

Facial Image Analysis and Its Applications to Facial Expression Recognition

Cigdem Turan

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

August 2018

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

_____Cigdem Turan_____(Name of student)

Abstract

Facial expression recognition (FER), defined as the task to identify someone's emotional or affective state based on face images, has been studied widely in the last few decades. With the development of computer-vision techniques and the availability of better computational power, FER methods can now achieve great performance on the recognition of posed expressions. For instance, the latest recognition rates on the widely used database, the Extended Cohn-Kanade (CK+) database, consisting of posed expressions, have reached over 98% accuracy. However, there are many possible applications of facial behavior analysis, which vary from advertising to teaching, from pain detection to lie detection, and those applications require more complicated recognition methods that can deal with spontaneous expressions and real-life conditions, such as pose, intensity and illumination variations. Therefore, the objectives of this thesis are to research the basic steps of FER, such as classification and face representation, including feature extraction, dimensionality reduction and feature fusion, and to review and develop efficient and robust methods to improve the FER performance.

In this thesis, we first present the histogram-based local descriptors applied to FER from static images, and provide a systematic review and analysis of them. We start with introducing a taxonomy for histogram-based local descriptors and highlight the representative examples of the specific steps, while analyzing their strengths and weaknesses. Then, we compare the performance of 27 local descriptors on four popular databases with the same experiment set-up, including the use of two classifiers, different image resolutions, and different numbers of sub-regions. In addition to their accuracy, other important aspects, such as face resolutions for the best performances, are also studied. Moreover, we compare the results achieved by

handcrafted features, e.g. histogram-based local features, with the results obtained by feature learning and the state-of-the-art deep features. We also evaluate the robustness of the respective local descriptors in the scenario of a cross-dataset facial expression recognition problem. This part of the thesis aims to bring together different studies of the visual features for FER by evaluating their performances under the same experiment set-up, and critically reviewing various classifiers making use of the local descriptors.

Having conducted a review of existing local descriptors with different settings, we propose different methods for FER. In the literature, features extracted from two or more modalities, such as audio, video or image, have been combined for FER, as well as features extracted from different regions of the same face images to enhance FER accuracy. In this thesis, we propose a two-level classification framework with region-based feature fusion. In the first level, the features from the eye and the mouth windows, which are the most salient regions in faces for representing facial expressions, are concatenated to form an augmented feature for facial expression. If the expression of a query input cannot be determined confidently by a Support Vector Machine classifier, the features in the second level of classification are obtained by fusion using Canonical Correlation Analysis (CCA), which can explore and enhance the correlation between the eye and the mouth features, since these two regions should have a high correlation in describing a specific facial expression.

There have been many image features or descriptors proposed for FER, which can achieve different recognition rates. Also, different descriptors can achieve different recognition rates for a specific expression class. In this thesis, we propose an emotion-based feature-fusion method, using the Discriminant-Analysis of Canonical Correlations (DCC) with an adaptive descriptor selection algorithm. The adaptive

descriptor selection algorithm determines the best two features for each expression class on a given training set followed by the fusion of these two features so as to achieve a higher recognition rate for each expression. Our aim is to find the best discriminant features by combining the different descriptors for recognizing each facial expression. To the best of our knowledge, we are the first to use different coherent descriptors for the recognition of different expressions.

Dimensionality reduction is a fundamental problem in any classification problem, since many real-world computer-vision and pattern-recognition applications are involved with large volumes of high-dimensional data. In this thesis, we propose a new and more efficient manifold-learning method, named Soft Locality Preserving Map (SLPM). SLPM is a graph-based subspace-learning method, with the use of *k*-neighborhood information and the class information. The key feature of SLPM is that it aims to control the level of spread of the different classes, because the spread of the classes in the underlying manifold is closely connected to the generalizability of the learned subspace. We also propose an efficient way to further enhance the generalizability of the manifolds of the different expression classes by feature generation, so as to represent each expression manifold completely.

The automatic recognition and interpretation of facial expressions has much to offer to various disciplines, such as psychology, cognitive science and neuroscience. These possible applications have motivated us to extend our research scope to interdisciplinary studies. To do so, a behavioural experiment is designed to study the multidimensionality of comprehension shown by facial expressions, which brings together studies, results and questions from different disciplines by focusing on the computational analysis of human behavior. A new multimodal facial expression database, named Facial Expressions of Comprehension (FEC), consisting of the videos

recorded during the behavioral experiments, is created and released for further academic research. The multidimensionality of comprehension is analyzed in two aspects: 1) the level of comprehension shown by expressions, and 2) the level of engagement with the corresponding feedback. We also propose a new methodology that aims to explore the changes in facial configuration caused by an event, namely Event-Related Intensities (ERIs).

All the methods proposed in this thesis have been evaluated and compared to the state-of-the-art methods. Experimental results and comprehensive analyses show that our algorithms and frameworks can achieve convincing and consistent performances.

List of Publications

[1]. Cigdem Turan and Kin-Man Lam, "Histogram-based Local Descriptors for Facial Expression Recognition (FER): A comprehensive Study," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 331-341, 2018.

[2]. Cigdem Turan and Kin-Man Lam, "Region-based feature fusion for facial-expression recognition," Proceedings, *2014 IEEE International Conference on Image Processing (ICIP'2014)*, pp. 5966-5970, 2014.

[3]. Cigdem Turan, Kin-Man Lam, and Xiangjian He, "Facial expression recognition with emotion-based feature fusion," Proceedings, *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC'2015)*, pp. 1-6, 2015.

[4]. Cigdem Turan, Kin-Man Lam, and Xiangjian He, "Soft Locality Preserving Map (SLPM) for Facial Expression Recognition," *submitted to Journal of Image and Vision Computing*, 2018.

[5]. Cigdem Turan, Yixin Wang, Shun-Cheung Lai, Karl David Neergaard, and Kin-Man Lam, "Facial Expressions of Sentence Comprehension," *accepted for presentation in IEEE International Conference on Digital Signal Processing (DSP2018).*

[6]. Cigdem Turan, Karl David Neergaard, and Kin-Man Lam, "Facial Expressions of Comprehension (FEC)," *submitted to IEEE Transactions on Affective Computing*, 2018.

Acknowledgement

Many people have supported me during my Ph.D study, and I would like to thank them all. First and foremost, I would like to thank my supervisor, Prof. Kin-Man Lam, and express my extreme gratitude to him for his enduring support, motivation and patience throughout my Ph.D study and research. He is a great mentor with his insightful professional advice, solid foundation of knowledge and wisdom of life. When I was thousand miles away from home, his constant encouragements and guidance have kept me motivated. Without his belief in me, this thesis would not have been possible.

My sincere thanks also go to Prof. Xiangjian He for his support and motivation during my stay in University of Technology, Sydney, and Prof. Stephen Politzer-Ahles for his insightful suggestions. I have learned much from their knowledge and experience that have widened my research.

I am fortunate to have a chance to work with Huiling Zhou, Muwei Jian, Hailiang Li, Dong Li, Khaled W. Aldebei, Mohammed Ambusaidi, Engr Saad Shakeel Chughtai, Shun-Cheung Lai, and Chenhang He in the past four years. I am thankful to their share of knowledge and thoughts at work, and the life we have enjoyed together.

Morever, I am thankful to many people in my life who always care for me. I couldn't finish my PhD without Anushree Mahapatra’s and Cristina Michelini's endless emotional support, Karl Neergaard's constant intellectual stimulation, and Christopher Schwiewager's never-ending compassion. Last but not the least, all that I have done would be impossible without the support, encouragement, and patience of my parents, Aydin Turan and Nebahat Turan, and my brother, Kerim Turan. They have provided me with the opportunity for growth and education through my entire life and I am forever indebted to them.

Table of Contents

List of Abbreviations

AN	Anger
AnU	Animation Unit
AU	Action Unit
AUC	Area under the Curve
BPPC	Binary Pattern of Phase Congruency
CBFD	Compact Binary Face Descriptor
CCA	Canonical Correlation Analysis
CMVM	Constrained Maximum Variance Mapping
CO	Contempt
DCC	Discriminant-Analysis of Canonical Correlations
DFT	Discrete Fourier Transform
DI	Disgust
ERI	Event-Related Intensity
ERP	Event-Related Potential
FACS	Facial Action Coding System
FE	Fear
FEC	Facial Expressions of Comprehension
FER	Facial Expression Recognition
GDP	Gradient Directional Pattern
GDP2	Gradient Direction Pattern
GLTeP	Gradient Local Ternary Pattern
HA	Happiness
HCI	Human-Computer Interaction
HOG	Histogram of Oriented Gradients

HOG-TOP Histogram of Oriented Gradients from Three Orthogonal Planes

ILSDA Improved Locality Sensitive Discriminant Analysis

IWBC Improved Weber Binary Coding

***k*-NN** *K* Nearest Neighbour

LBP Local Binary Pattern

LAP Local Arc Pattern

LDA Linear Discriminant Analysis

LDiP Local Directional Pattern

LDiPv Local Directional Pattern Variance

LDN Local Directional Number Pattern

LDP Local Derivative Pattern

LDTP Local Directional Texture Pattern

LFD Local Frequency Descriptor

LGBPHS Local Gabor Binary Pattern Histogram Sequence

LGBP-TOP Local Gabor Binary Patterns from Three Orthogonal Planes

LGDiP Local Gabor Directional Pattern

LGIP Local Gradient Increasing Pattern

LGP Local Gradient Pattern

LGTrP Local Gabor Transitional Pattern

LME Linear Mixed Effects

LMP Local Monotonic Pattern

LOSO Leave-One-Subject-Out

LOTO Leave-One-Trial-Out

LPMIP Locality-Preserved Maximum Information Projection

LPP Locality Preserving Projection

LPQ	Local Phase Quantization
LPQ-TOP	Local Phase Quantization from Three Orthogonal Planes
LSDA	Locality Sensitive Discriminant Analysis
LS-SVM	Least Square Support Vector Machine
LTeP	Local Ternary Pattern
LTrP	Local Transitional Pattern
LXNORP	Local XNOR Pattern
MBC	Monogenic Binary Coding
MBP	Median Binary Pattern
MFA	Marginal Fisher Analysis
MMC	Maximum Margin Criterion
MMDA	Multi-Manifolds Discriminant Analysis
MRELBP	Median Robust Extended Local Binary Pattern
MTP	Median Ternary Pattern
NN	Nearest Neighbour
OLPP	Orthogonal Locality Preserving Projection
PAD	Pleasure, Arousal and Dominance
PC	Phase Congruency
PCA	Principal Component Analysis
PDM	Point Distribution Model
PHOG	Pyramid of Histogram of Oriented Gradients
PSF	Point Spread Function
RCM	Region Covariance Matrices
ROC	Receiver Operating Characteristic
SA	Sadness

SDM Soft Discriminant Map

SLPM Soft Locality Preserving Map

SLPP Supervised Locality Preserving Projection

SOLPP Supervised Orthogonal Locality Preserving Projections

SSP Social Signal Processing

SSS Small Sample Size

STM Spatiotemporal Manifold

SU Surprise

SVM Support Vector Machine

Volume-LBP Volume Local Binary Pattern

Volume-LPQ Volume Local Phase Quantization

WLD Weber Local Descriptor

List of Figures

List of Tables

Chapter 1. Introduction

This chapter aims to introduce the general research background of social signal processing and highlights the importance of facial expressions as a social signal. Our research motivation, as well as the original works presented in this thesis, are discussed with the outline of the content.

1.1. Research background

People tend to focus only on verbal messages while communicating with other people. However, nonverbal aspects of social communication are as much important as, if they are not more important than, verbal aspects of social communication to construct our perception of social context. Moreover, such behaviours can occur without the presence of another person. For instance, we still use gestures to express ourselves when we are talking on the phone, even though the corresponding listener cannot see us [121]. We display our emotions on our face even when we are alone or we do not have any mean to express ourselves to a person-of-interest. In a sense, we are hard wired in our brains to display nonverbal cues, i.e. social signals, and extract social information from them.

Social signals can carry varying functionalities, such as deceiving, managing interaction, expression emotion and forming impressions, in which facial expressions and gestures are the primary means for expressing and interpreting emotions. Facial expressions, as a mean for expressing and interpreting emotions, have been widely studied in psychology and neuroscience in the last century, where the first study dates back to Darwin [44].

If facial expressions and other social signals are an important part of our daily communication as described, why not detecting and understanding those social signals

automatically using machine analysis? This is exactly what the domain, known as social signal processing (SSP), aims. In all of the social cues, facial and vocal behaviours, gesture and postures have been studied extensively for their affective content [65, 255].

A general SSP system should include the basic steps as data capture, person detection, behavioural cue extraction and social behaviour understanding. In this thesis, we focus on studying facial behaviour analysis as a social signal, mainly due to its extensive range of real-life applications related to security, teaching and diagnosis. Most of the studies presented in this thesis take granted of the first two steps with the help of the ready-to-use databases, and focus on the recognition of the certain facial configurations.

1.2. Motivation

Facial expression recognition (FER), defined as the task to identify someone's emotional or affective state based on face images, has been studied widely in the last few decades. With the development of computer-vision techniques and the availability of better computational power, FER methods can now achieve great performance on the recognition of posed expressions. For instance, the latest recognition rates on the widely used database, the Extended Cohn-Kanade (CK+) database, consisting of posed expressions, have reached over 98.6% accuracy [52]. However, there are many possible applications of facial behaviour analysis, which vary from advertising to teaching, from pain detection to lie detection, and those applications require more complicated recognition methods that can deal with spontaneous expressions and real-life conditions, such as pose, illumination variations, etc. These facts have encouraged us to dig deeper into the basic steps of FER, such as face representation,

dimensionality reduction, and feature fusion, and also to review and work on the improvement of the state-of-the-art methods.

Researchers in different disciplines, such as computational linguistics and computational neuroscience, are often not aware of the advances in recognizing facial information. This disconnection between disciplines limits the interdisciplinary multimodal studies to understand human facial behavior. This disconnection has motivated us to extend our research scope to interdisciplinary study by attempting to bring studies, results and questions together from different disciplines. Thus, the sentence comprehension shown by the facial expressions, which is a novel problem in the area of facial behaviour understanding, is investigated, which can practically provide methodological support to investigate people's facial behavior and mental states.

1.3. Statement of originality

The following contributions presented in this thesis are claimed to be original.

a. Histogram-based local descriptors applied to FER from static images are presented and a systematic review and analysis of them are provided. First, the main steps in encoding binary patterns in a local patch, which are required in every histogram-based local descriptor, are described. Then, the existing local descriptors are listed while analyzing their strengths and weaknesses. Finally, the experimental results of all these descriptors, in total 27 descriptors, on commonly used facial expression databases with varying resolution, noise, occlusion, and number of sub-regions are presented and compared with the results obtained by the state-of-the-art deep learning methods. The robustness of the respective local descriptors in the scenario of a cross-dataset FER problem is also evaluated.

b. A two-level classification framework for facial expression recognition is proposed. In the first level, the features from the eye and the mouth windows are concatenated to form an augmented feature for facial expression. If the expression of a query input cannot be determined confidently by a Support Vector Machine classifier, the features in the second level of classification are obtained by fusion using Canonical Correlation Analysis (CCA), which can explore and enhance the correlation between the eye and the mouth features.
c. An emotion-based feature fusion method using the Discriminant-Analysis of Canonical Correlations (DCC) with an adaptive descriptor selection algorithm is proposed. Great amount of image features and descriptors proposed for FER, in which different features may be more accurate for the recognition of different expressions. The adaptive descriptor selection algorithm determines the best two features for each expression class on a given training set followed by the fusion of these two features so as to achieve a higher recognition rate for each expression.
d. A new and more efficient manifold-learning method, named Soft Locality Preserving Map (SLPM), is proposed. SLPM is a graph-based subspace-learning method, with the use of k-neighborhood information and the class information. The key feature of SLPM is that it aims to control the level of spread of the different classes, because the spread of the classes in the underlying manifold is closely connected to the generalizability of the learned subspace. Also, a feature generation method is proposed to learn the manifold for each expression class more accurately. In the proposed feature generation method, the features for low-intensity expressions are generated directly in the feature domain.
e. A behavioural experiment is designed to study the multidimensionality of comprehension shown by facial expressions, which brings together questions,

studies and results from different disciplines. A new multimodal facial expression database, named Facial Expressions of Comprehension (FEC), consisting of the videos recorded during the behavioral experiments is released for further academic research. The multidimensionality of comprehension is analyzed in two aspects: 1) the level of comprehension shown by expressions, and 2) the level of engagement with the corresponding feedback. For the first aspect, an SVM classifier with facial appearance information extracted using a spatiotemporal local descriptor named LPQ-TOP is employed. For the second aspect, the statistical analysis is employed on Event-Related Intensities (ERIs), which is proposed to explore the changes in facial configuration caused by an event.

1.4. Outline of the thesis

This thesis is organized as seven chapters and the chapters are outlined as follows.

Chapter 2 gives an overview of FER techniques by introducing the general concepts followed by the history and the development of FER. The development of the FER methods is explained with respect to several aspects of a FER system such as face representation, dimensionality reduction and classification. Then, some current trends of FER are presented, which have shifted the recent research substantially.

Chapter 3 presents histogram-based local descriptors applied to FER from static images, and provide a systematic review and analysis of them. First, the main steps in encoding binary patterns in a local patch, which are required in every histogram-based local descriptor, are described. Then, the existing local descriptors are listed while analyzing their strengths and weaknesses. Finally, the experimental results of all these descriptors on commonly used facial expression databases with varying resolution, noise, occlusion, and number of sub-regions are presented and compared with the results obtained by the state-of-the-art deep learning methods. This chapter aims to

bring together different studies of the visual features for FER by evaluating their performances under the same experimental setup, and critically reviewing various classifiers making use of the local descriptors.

Chapter 4 presents a two-level feature-fusion framework based on Canonical Correlation Analysis (CCA) for FER. In the framework, features from the eye and the mouth windows are extracted separately, which are correlated with each other in representing a facial expression. For each of the windows, two effective features, namely the Local Phase Quantization (LPQ) and the Pyramid of Histogram of Oriented Gradients (PHOG) descriptors, are employed to form low-level representations of the corresponding windows. The features are then represented in a coherent subspace by using CCA in order to maximize the correlation. In the experiments, the Extended Cohn-Kanade dataset is used; its face images span seven different emotions, namely anger, contempt, disgust, fear, happiness, sadness, and surprise.

Chapter 5 introduced an emotion-based feature fusion method using the Discriminant-Analysis of Canonical Correlations (DCC) for FER. There have been many image features or descriptors proposed for FER and the different features may be more accurate for the recognition of different expression. Also, experiments show that descriptors are sensitive to the conditions of images, such as race, lighting, pose, etc. Thus, an adaptive descriptor selection algorithm is proposed, which determines the best two features for each expression class on a given training set. These two features are, then, fused, so as to achieve a higher recognition rate for each expression. In the emotion-based feature fusion method, four effective descriptors for face representation, namely Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Weber Local Descriptor (WLD), and Pyramid of Histogram of Oriented Gradients

(PHOG), are considered. Supervised Locality Preserving Projection (SLPP) is applied to the respective features for dimensionality reduction and manifold learning.

Chapter 6, firstly, reviews the most popular and the state-of-the-art methods for dimensionality reduction, which vary from unsupervised to supervised, and from statistics to graph-theory based. Then a new and more efficient manifold-learning method, named Soft Locality Preserving Map (SLPM), is presented. Furthermore, feature generation and sample selection are proposed to achieve better manifold learning. The proposed manifold-learning method can be applied to various pattern recognition applications, and its performance on FER is evaluated using databases, such as the Bahcesehir University Multilingual Affective Face Database (BAUM-2), the Extended Cohn-Kanade (CK+) Database, the Japanese Female Facial Expression (JAFFE) Database, and the Taiwanese Facial Expression Image Database (TFEID).

Chapter 7 presents a study on the multidimensionality of comprehension shown by facial expressions. To do so, a behavioral experiment is designed where participants took part in a roughly 30-minute computer mediated task. In the experiment, they were asked to answer either “true” or “false” to knowledge-based questions, then immediately given feedback of “correct” or “incorrect”. Their faces were recorded during the task using the Kinect v2 device. A new multimodal facial expression database, named Facial Expressions of Comprehension (FEC), consisting of the videos recorded during the behavioral experiments is presented. In order to identify the level of engagement with the corresponding feedback, a new methodology is proposed that aims to explore the changes in facial configuration caused by an event: Event-Related Intensities (ERIs). ERIs, calculated using the Animation Units, obtained by the Kinect v2 device, are then used in statistical analysis. To identify the level of comprehension

shown by expressions, the SVM classifier with facial appearance information extracted using a spatiotemporal local descriptor named LPQ-TOP is employed.

Chapter 8 concludes the works described in this thesis with suggestions for further development.

Chapter 2. Literature review

In this chapter, general concepts and the development of facial expression recognition (FER) algorithms are introduced. A detailed review of FER algorithms is presented with respect to several aspects of FER, such as computational emotion models, face representation, and subspace learning. Lastly, current trends on FER algorithms are discussed.

2.1. Problem statement

FER, defined as the task to identify someone's emotional or affective state based on face images, has been widely studied in the past decade. Compared to other forms of nonverbal communication, such as gesture and touch, facial expressions can disclose a variety of emotions and are among the most universal forms of nonverbal communication that can be recognized and displayed across cultures. It has great potential to be used in human-computer interaction (HCI), such as assistive driving [231], embodied agents [74], and in applications, such as diagnosis [37, 105]; and computer games [120].

In general, a FER method has three steps: (a) face detection, (b) feature extraction, and (c) facial expression recognition (classification). In the face-detection step, the locations of the faces are found in a given image, in addition to several facial features, such as the eyes and the mouth, which are crucial to FER. Since face detection is an important step for any problem that uses faces as a form of input, it should perform well under several challenging variations, such as pose, occlusion and illumination. Face detection has been widely studied in the past few decades [40, 137, 212, 230, 273]. The next step is extracting geometrical or appearance features from faces to obtain information that can be used for recognizing people's different facial

expressions. The features extracted in this step should preferably discard the subject's identity information and highlight the expression-specific information of a face. The extracted features are then fed to a classification model that can match the facial expression as one of the emotion classes of interest. The feature-extraction step could be followed by a feature-processing step, such as dimensionality reduction and feature fusion, before they are used in the classification model. In this thesis, local descriptors that are applied to the FER problem as the feature-extraction method are reviewed in Chapter 3, while the rest of the chapters, i.e. Chapter 4 to Chapter 6, mainly focus on the different feature-processing methods, where the features are extracted and applied to FER.

To evaluate a FER system, the recognition accuracy, which is the percentage of the testing images being correctly recognized, is often used as a measurement of its robustness. To obtain the recognition accuracy, two main approaches adapted: *n*-fold cross validation and leave-one-subject-out (LOSO). In *n*-fold cross validation, the dataset is divided into *n* folds, where *n*–1 folds are used to form a training set to test its performance on the fold left in each run. The program runs *n* times and the recognition accuracy is calculated by averaging the accuracy of each run. LOSO, on the other hand, defines the training and the testing sets in a way that the testing set would consist of all the samples from one subject at a time, while the training set would consist of the rest.

Although a great number of studies have been done in the area of FER, several issues have remained unsolved, because a FER system should be able to handle various input formats, such as image and video, produced under different environments. Some of the current challenges of FER are listed as follows:

Illumination: Face images obtained in real-life conditions often have changing lighting conditions, which could affect the appearance of facial expressions.

Pose: Most of the current works can handle well the recognition of facial expressions from frontal faces. However, in real-life settings, faces are often captured under various poses and it can be challenging to recognize the expressions in such settings.

Occlusion: Face occlusion can affect the classification process by decreasing the discriminative information available, because some parts of a face may be missing, due to various reasons, such as wearing sunglasses or a scarf.

Variations in expression: Although there have been studies on the universality of several emotions, facial expressions can vary depending on someone's age, race and cultural background. This can affect the generalizability of the FER system.

Expression intensity: Expression can be displayed as explicit, i.e. macro-expressions, or subtle, i.e. micro-expressions, or somewhere in between. Most of the current studies focus either on macro-expression recognition or micro-expression recognition.

Processing time: The recognition rate of a facial expression algorithm is as important as the processing time of a single face image, especially when the recognition is required to be performed in real time.

2.2. Review on development of facial expression recognition algorithms

The earliest work on FER dates back to 1872 with Darwin's study [44]. However, automatic analysis of FER was observed first in 1978, in the engineering literature [214], and then gained momentum after the work of Mase and Pentland [160] in the 1990s, mainly due to the advancements in related research areas, such as face detection

and tracking. Since then, automatic FER has been intensively studied in the areas of psychology, neuroscience, engineering, etc. The development of FER can be observed over time and through the methods used, which vary from acted to spontaneous expressions, from macro-expressions to micro-expressions, from frame-based to sequence-based FER, from handcrafted features to feature learning, from the nearest neighbour classifier to deep learning, from simple facial expression recognition to facial behaviour understanding, etc. In this section, we give a review of the development of FER methods based on these aspects.

2.2.1. Computational emotion models

Darwin, in his book, "The Expression of the Emotions in Man and Animals", investigated scientifically the meaning of varying facial expressions with the muscle groups that cause them [182]. He further emphasized the functional role of facial expressions in communication. Since then, emotions and affects have been explored in various scientific disciplines, such as cognitive science, neuroscience and psychology. Since automatic affect analysis is strongly correlated with the advances in the emotion theory, in this section, we present the three most common emotion theories, with their uses in automatic affect analysis.

Emotion modelling, based on the research in psychology, is mainly divided into three approaches: categorical, dimensional, and appraisal-based approach. The categorical approach claims that emotions can be categorized as basic and non-basic emotions where the basis for inclusion of any emotion to the category of basic emotions has been varied with bodily involvement, relation to instinct and unlearned emotion step [182]. Plutchik [189] categorized the emotions as eight basic emotions such as fear, anger, sorrow, joy, disgust, acceptance, anticipation, and surprise, and argued that any emotion can be represented by a mixture of these principal emotions.

Ekman et al. [61] have proposed six universal facial expressions as basic emotions which have analogous muscle movements for each expression among race, age and gender: anger, disgust, fear, joy, sadness, surprise, In [63], they also have showed that although contempt is mixed emotion of disgust and anger, it is a candidate to be primary expression in addition to previous six expressions due to the fact that expression of contempt also can be understood universally across Western and Non-Western people. Six basic emotions proposed by Ekman et al. have been broadly accepted as a computational emotion model and adopted extensively to the automatic analysis of human affect. Most of the early facial expression databases based the annotation of the human affect on the six basic emotions [34, 116, 155, 225].

Due to the fact that basic emotions are often not able to reflect the complexity of the wide range of expressions in daily life, a number of researchers proposed the use of dimensional model of human affect, where affective states are not represented by a single or a group of discrete emotion labels but collocated in a dimensional space, e.g. two dimensional space as arousal and valence, i.e. circumplex model of affect [198], and three dimensional space as pleasure, arousal and dominance (PAD) [167]. The arousal dimension refers to the level of excitement of the emotion where the valence dimension refers to the positivity of the emotion. The dominance dimension, on the other hand, refers to the degree of power of the emotion. There exist several studies that focus on the modelling affect based on circumplex model of affect [166, 175, 209] or PAD [26, 107, 236].

The dimensional emotion model is a better approach than the categorical emotion model for the automatic analysis of human affect in the wild because of the fact that the dimensional emotion model can reflect a large variety of expressions. However, there are several disadvantages of the dimensional emotion model. For example,

although the theorists were able to map the basic emotions to the dimensional space, not all the emotions can fit perfectly to the dimensional space, e.g. confusion. Also, the annotation of the facial expressions suffers from the subjectivity of the annotators since the dimensional emotion theory has an infinite number of values.

A model, known as OCC, is formed as the standard cognitive appraisal model for emotions by Ortony, Clore and Collins [181] where emotions according to OCC arise from the positive or negative reactions to situations constructed as the desirability of events, actions of actors and attitudes of objects [205]. Although the OCC, which specifies 22 emotions categories, was used to model the learning content by mapping students' emotions when they played an educational game [38], they obtained cognitive-related measurements such as the participants' goal for the game, through indirect evaluation. It highlights the complexity of the use of the cognitive appraisal model in automatic analysis of effect without direct input from the participants themselves.

Even after a century of research in the areas, such as psychology, neuroscience and cognitive science, with all of the aforementioned issues, the question of the appropriate model of emotion for automatic analysis of human affect remains under discussion.

2.2.2. Face representation

FER systems aim to determine one's emotional state, based on face images, regardless of one's age, gender or race. In recent years, FER has shown its importance in HCI, such as assistive driving [231], embodied agents [74], and in applications such as diagnostic [37, 105], and computer games [120]. This is mainly due to the fact that facial expression is an important aspect of non-verbal communication. Thus, the demand for an effective FER technology has been increasing. Although much progress has been made on recognizing facial expressions, it is still a difficult task, due to the

complexity and variability of facial expressions, and an effective facial-representation method is a vital step to improve the recognition rate in FER.

Facial feature representations proposed in the literature can now be divided into three categories: geometrical, appearance-based, and deep features. Geometrical features [159] take advantage of shape and location information of facial components and salient points, i.e. the eyes, lips, nose tip, etc. FER with Action Unit (AU) recognition is a geometrical feature-based approach, which has achieved more attention recently with the advancement in deep neural-network structures [103, 218]. However, geometrical features still require an accurate and reliable reconstruction and tracking of the facial landmarks. Therefore, it is difficult to achieve in real-life situations. Furthermore, AU-based facial expression recognition may require training data whose Action Units are already labelled by experts, which is a labor-intensive and time-consuming process. Recent studies have shown that appearance-based methods can achieve similar or better performance than AU recognition-based methods [224].

Appearance-based features are based on texture information related to the expressions on a face, e.g. wrinkles, skin changes, etc., which can be applied to the whole face or specific face regions. One of the first attempts of FER based on texture classification is to use Local Binary Pattern (LBP), which was proposed by Ojala et al. [179]. LBP is one of the most widely used descriptors, due to its computational simplicity, discriminative power, and insensitivity to monotonic grayscale changes.

The successful application of LBP on the FER problems has inspired further studies for local descriptors. These studies focus on enhancing the coding techniques, e.g. different neighbourhood sizes, processing of input images, e.g. linear filtering, transformations, etc., to emphasize the expression-specific information. Numerous

variants of LBP have been proposed for the problems, such as face recognition [132, 246], facial expression recognition [117], texture classification [210], spatiotemporal feature representation [266], and medical image analysis [173]. Some comprehensive studies of LBP variants can be found in [54, 124, 174].

Recently, local binary feature learning methods have been proposed for efficient and data-adaptive face representation, because LBP and other hand-crafted features require strong prior knowledge of the problem in order to engineer them by hand [11, 58, 150]. The objective behind the feature learning methods is to learn a feature mapping using raw pixels to project each local pixel difference vector into a low-dimensional binary vector that can efficiently represent the face data. Therefore, a codebook constructed using the learned binary codes can be used to obtain a histogram feature for each image [58]. To the best of our knowledge, local binary feature learning methods have not been applied to the FER problem, but only to age estimation [148] and face recognition [57, 147, 149]. As LBP has been successfully applied to the tasks for facial image analysis, it is worthwhile evaluating the recently proposed local binary feature learning methods on FER.

Deep neural networks have been studied widely for many pattern-recognition tasks, such as human pose estimation [219], face recognition [201], gender recognition [156], image recognition [88, 208], which require learning from a large amount of data. The increasing popularity and the success of deep features are also rooted in the FER problems [52, 111, 140-142, 169, 267]. Although the increase in recognition rate for FER is undeniable, the debate between hand-crafted features and deep features is still active. Benitez-Garcia et al. [17] proposed a local descriptor, i.e. a handcrafted feature, which can achieve a higher recognition rate than any deep neural-network structure until now. This suggests that the domain-specific knowledge and the

handcrafted features are still effective and favourable for visual classification. A comprehensive analysis of the current local features is presented in Chapter 3.

The facial features representation methods mentioned above often aim an image-based FER which aims recognizing facial expressions of static images. Sequence-based FER, on the other hand, focuses on the recognition of facial expressions in a sequence of static images. One of the first attempts on recognizing emotions in a video was applying a feature extraction method designed for image-based FER to each sequence followed by employing advanced classification methods such as Hidden Markov Model [136]. Another approach to sequence-based FER targeted finding a single image which can represent the whole video sequences by holding the expression with the highest intensity, i.e. peak frame selection [257]. The weakness of the latter approach is the fact that a video would have more than one frame with the highest intensity. Also, a peak frame would not be able to represent the temporal changes of facial expressions on the face which are, nowadays, essential to FER in the applications just as drowsiness detection [231], pain recognition [10, 114, 154, 195] etc.

To encode the temporal changes of facial expressions, several dynamic texture descriptors have been proposed. The Volume Local Binary Pattern (Volume-LBP) and the Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) [265] are, to the best of our knowledge, the very first attempts of extending spatial local descriptors to spatiotemporal local descriptors. The successful application of them has inspired researchers to implement more spatiotemporal descriptors such as, Volume Local Phase Quantization (Volume-LPQ) [183], Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) [108], Spatiotemporal Local Monogenic Binary Patterns [92], Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-

TOP) [7], and Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) [32].

Chapter 3 presents a comprehensive analysis and comparison of histogram-based local descriptors applied to FER.

2.2.3. Subspace learning

Dimensionality reduction, which aims to find the distinctive features to represent high-dimensional data in a low-dimensional subspace, is a fundamental problem in classification. Many real-world computer-vision and pattern-recognition applications, e.g. facial expression recognition, are involved with large volumes of high-dimensional data. Principal Component Analysis (PCA) [90, 158] and Linear Discriminant Analysis (LDA) [73, 158] are two notable linear methods for dimensionality reduction. PCA aims to find principal projection vectors, which are those eigenvectors associated with the largest eigenvalues of the covariance matrix of training samples, to project the high-dimensional data to a low-dimensional subspace. Unlike PCA, which is an unsupervised method that considers common features of training samples, LDA employs the Fisher criteria to maximize the between-class scattering and to minimize the within-class scattering. Although LDA is superior to PCA for pattern recognition, it suffers from the small-sample-size (SSS) problem [192] because the number of training samples available is much smaller than the dimension of the feature vectors in most of the real-world applications. To overcome the SSS problem, Li et al. [131] proposed the Maximum Margin Criterion (MMC) method, which utilizes the difference between the within-class and the between-class scatter matrices as the objective function. In [143], it is shown that intra-class scattering has an important effect when dealing with overfitting in training a model. Unlike the conventional wisdom, too much compactness within each class decreases

the generalizability of the manifolds. Since LDA and MMC are too "harsh", they need to be softened. Liu et al. [143] proposed the Soft Discriminant Map (SDM), which tries to control the spread of the different classes. MMC can be considered as a special case of SDM, where the softening parameter $\beta = 1$.

Linear methods, like PCA, LDA and SDM, may fail to find the underlying nonlinear structure of the data under consideration, and they may lose some discriminant information of the manifolds during the linear projection. To overcome this problem, some nonlinear dimensionality reduction techniques have been proposed. In general, the nonlinear dimensionality reduction techniques are divided into two categories: kernel-based and manifold-learning-based approaches. Kernel-based methods, as well as the linear methods mentioned above, only employ the global structure while ignoring the local geometry of the data. However, manifold-learning-based methods can explore the intrinsic geometry of the data. Popular nonlinear manifold-learning methods include ISOMAP [217], Locally Linear Embedding [197], and Laplacian Eigenmaps [16], which can be considered as special cases of the general framework for dimensionality reduction named "graph embedding", proposed by Yan et al. [242]. Although these methods can represent the local structure of the data, they suffer from the out-of-sample problem. Locality Preserving Projection (LPP) [176] was proposed as a linear approximation of the nonlinear Laplacian Eigenmaps [16] to overcome the out-of-sample problem. LPP considers the manifold structure via the adjacency graph. The manifold-learning methods presented so far are unsupervised methods, i.e. they do not consider the class information. Several supervised-based methods [31, 202, 237] have been proposed, which utilize the discriminant structure of the manifolds. With the Marginal Fisher Analysis (MFA) [242], which uses the Fisher criterion and constructs two adjacency graphs to represent the within-class and

the between-class geometry of the data, several other methods have been proposed with similar ideas, such as Locality-Preserved Maximum Information Projection (LPMIP) [234], Constrained Maximum Variance Mapping (CMVM) [130], and Locality Sensitive Discriminant Analysis (LSDA) [25]. In real-life applications, unlabeled data can exist because of various reasons. To deal with this problem, various semi-supervised learning algorithms have also been proposed [191, 235, 260]. In Chapter 6, a new subspace learning method that overcomes the shortcoming of the existing subspace learning methods is presented.

2.2.4. Classification

In the literature, there are plenty of studies regarding classifiers, since it is the last and a critical step for a successful FER system. Early studies broadly used linear regression methods for binary classification, such as the least square classifier, perceptron [196], logistic regression [6] and linear Support Vector Machines (SVMs) [21]. However, linear methods may fail to achieve satisfactory results, unless the features are linearly separable. To overcome this problem, nonlinear classifiers based on kernel methods, like nonlinear SVM, have become more popular. These days, widely used classifiers include the *k*-nearest neighbour (*k*-NN) [42], SVM and the Convolutional Neural Network [118, 125, 146, 161] for the frame-based FER; and SVM, the dynamic Bayesian network [66, 172, 263] and Recurrent Neural Network [19, 53, 60, 71] for the sequence-based FER. In the experiments presented in this thesis, the *k*-NN classifier and the SVM are used for the frame-based FER.

The *k*-NN classifier, which is the most fundamental classification method, is commonly based on Euclidean distance, where the prediction is decided in the favour of the closest sample, and it has been widely used in FER because of its simplicity [15, 204, 221, 222]. Yet, SVM is often preferred over the *k*-NN classifier since it can attain

satisfactory success while dealing with the high dimensional data. SVM was first proposed for binary classification, i.e. two classes as -1 and 1. Since classification problems in the real-life often consist of more than two classes, several approaches have been proposed to apply the SVM to multiclass problems. The one-versus-all classification approach is the oldest one, which suggests training one classifier for each class, i.e. one class as 1 and the rest as -1. In Chapter 4 and Chapter 5, two different algorithms are presented to increase the recognition rate of the one-versus-all classification approach. Second approach to extend the SVM classifier to a multiclass classifier is the one-versus-one classification approach which trains a binary classifier for each pair of classes, i.e. $n(n-1)/2$ classifiers where n is the number of classes. In the rest of the chapters in the thesis, the one-versus-one classification approach is adopted.

2.2.5. Facial expression databases

A great number of facial expression databases have been created to encourage researchers to study facial affect analysis. Most of the early databases consisted of posed expressions recorded in a controlled environment. Some of the widely used databases consisting of posed expressions can be listed as below:

Extended Cohn-Kanade (CK+) [116] database is one of the most popular acted databases which contains the total of 593 sequences posed by 123 subjects in a laboratory environment -322 of the sequences are labelled as one of the seven discrete emotions (anger, contempt, disgust, fear, happiness, sadness and surprise). Each sequence starts with a neutral expression and ends with a peak frame of the particular expression where 68 facial points are also provided for each frame.

Japanese Female Facial Expression (JAFFE) [155] database**,** which is also a widely used database, was also collected manually in a laboratory environment

contains 219 images labelled as one of the six basic expressions (anger, disgust, fear, happiness, sadness and surprise) of 10 Japanese females.

MMI [225] consists of video recording from 75 subjects displaying six basic emotions in addition to neutral expression.

Multi-PIE [27], which was initially released for the face recognition problem, contains both images and videos with varying conditions such as illumination variations and different viewpoints from 337 subjects expressing five facial expressions such as smile, surprise, squint, disgust, scream and neutral.

Taiwanese Facial Expression Image Database (TFEID) [34] database contains 268 images, with the six basic expressions and the neutral expression, from 40 Taiwanese subjects.

There are several databases that include 3D data, such as **BU-3DFE [251], BU-4DFE [250]** and **Bosphorus [200].**

The databases mentioned above, which consist of images or videos obtained in a laboratory environment, failed to demonstrate real-life conditions. Also, much evidence has shown the differences between spontaneous and modelled expressions in terms of the underlying cognitive processes taking place during expressions and the expressions' features [64, 159]. These differences highlight the need for a shift from posed expressions to spontaneous expressions. Recently, numerous databases have been introduced to support the development of recognizing facial expressions in the wild.

Acted Facial Expressions in the Wild (AFEW) [48] database was created by extracting short video clips from movies based on subtitles. **Static Facial Expressions in the Wild (SFEW) [31]** is a database that consists of images covering seven emotions obtained from the video clips in AFEW. Since the actors in the movies are

trained to express almost spontaneous expressions and the scenes are often outside of laboratory settings, the AFEW and the SFEW databases exhibits close-to-real-life conditions with the presence of several expression under varying illumination conditions and head pose. Since the extraction of the video clips in AFEW depends on the related subtitles in the movie, the facial expressions presented in the considering frame might not overlap the related subtitles or they might not exist any face at all.

To overcome the limitations that are present in AFEW, a new semiautomatic method for extracting video clips from movies has been released. **Bahcesehir University Multilingual Affective Face Database (BAUM-2) [67],** which is a multilingual database consists of short videos extracted from movies, was created using the considering method. In BAUM-2, there are in total 1047 video clips from 286 subjects. An image dataset, namely BAUM-2i, consisting of images with peak expressions from the videos in BAUM-2, is also provided by the authors for static facial expression recognition.

SMIC [134], SAMM [45], CASME [244] and **CASME-2 [243]** consist of video sequences with micro-expressions. **MMSE [264]** is also a new database containing features of multimodal data, with spontaneous facial expressions. Recently, two studies have focused on affective modelling, with special attention given to identifying learning-related emotions. **DAISEE [83]** and **BNU-LSVED 2.0 [236]** were designed with the goal of embodying a computer-based learning tutor, which is equipped with a semblance of emotional intelligence.

All of these databases suffer from the cost of the human labelling. For example, all the videos in the BAUM-2 database were annotated by one of the seven prototypical expressions and neutral by 5 to 7 people with the intensity from 1 to 5, then, the emotion of a video was determined by majority voting and the intensity

values provided by each annotator were averaged. This leads to several problems. Firstly, with the increase attention to the big data, human labelling of videos has been the major cost of creating a database. Secondly, human labelling of expressions can cause noise to a database simply because of lack of objectivity in the labelling process. Although recent trend focuses reducing the effort of human labelling by labelling only a portion of a database and transferring the knowledge to the rest of the database, less studies investigated the objectivity of the labelling process.

In Chapter 7, a novel facial expression database, namely Facial Expressions of Comprehension (FEC) which avoids the labelling process made by humans, is presented which aims to offer a new novel research problem to the literature of FER.

2.2.6. Current trends of facial expression recognition

With the great advancements in computer vision and machine learning techniques, a promising new literature is developing that uses dynamic facial expression data to interpret the facial expressions in the wild. The use of dynamic, multimodal data, while at its naissance, has a great deal yet to offer. Vural et al. [231], whose findings may lead to an early alarm system to avoid car accidents due to fatigue, studied the drowsiness expressions of drivers. Cohn et al. [37] analyzed the facial expressions of patients during psychoanalytic sessions with the intent of diagnosing depression. Recently, Dibeklioglu and Gevers [51] investigated the automatic estimation of the level of taste liking through facial dynamics and showed that the proposed method is more reliable for estimation the taste than the human subjects. Concurrent to studies in facial recognition, bodily movement is also being used to enrich dynamic and multimodal data. Jaques et al. [106] focused on understanding and predicting bonding between interlocutors during conversations using facial expressions and body language. A similar multimodal approach to recognition was used in videos of adults

experiencing pain [10, 114, 154, 195]. In 2017, Jaiswal et al. [105] detected the presence of ADHD/ASD using facial movements with the Kinect device, a tool that gives multiple streams of data for both facial and bodily movement.

An application, currently being explored with recognition technology, is facial expression in teaching environments [122, 164, 199, 205]. The study in [205] proposed an affective e-learning model to investigate emotional states of learners and predict their future interaction with a learning system. Related affective states were based on a cognitive appraisal approach [181] that includes twenty emotions. Kort et al. [122] built a computer-based model that identifies users' affective states and responds accordingly, i.e. a learning companion. The paper also proposed a spiral affective model that combines the different phases of learning with the emotional axes. Sathik et al. [199] investigated student learning in a classroom setting through the correlation of successful comprehension with positive expressions, and failed comprehension with negative expressions. The authors did not build a model to predict students' comprehension but instead explored the statistical correlation of expressions towards comprehension types. Aligned with these new novel questions in the area of computer vision and social signal processing, the study presented in Chapter 7 aims to investigate multidimensionality of comprehension through facial expressions.

2.3. Conclusion

This chapter presents the principles and development of FER systems, with the current trends in FER. The development of FER has been observed through several aspects, such as face representation, subspace learning, etc. In the following chapters, the methods for FER proposed in this thesis will be presented, and they will be compared with various state-of-the-art methods presented in this chapter.

Chapter 3. Histogram-based local descriptors for facial expression recognition

3.1. Introduction

In Section 2.2.2, the importance of face representation in a FER system is described with the development of the face representation methods. In this chapter, we present a comprehensive study of appearance-based facial features, i.e. handcrafted features, and then we compare their best results with those methods based on recently proposed local binary feature learning methods and deep features for FER.

The steps of a basic FER framework with the use of appearance-based features can be listed as follows: 1) detecting and aligning the face images, 2) dividing each face image into several overlapping or non-overlapping regions, 3) extracting local features from these regions based on the local descriptors, 4) concatenating the respective local features to form a single feature vector, followed by unsupervised or supervised dimensionality reduction, 5) training a classifier based on the feature vectors from training samples, and 6) predicting the class label of a new query based on the trained classifier. The classification results depend on almost every step listed above. However, most of the recent studies have focused only on developing more robust local features [17, 28, 70, 133]. A robust feature should be highly discriminative, easily computed, of low dimensionality, insensitive to noise, such as illumination changes, and have low intra-class variations.

It is difficult to balance these properties for a local descriptor. For example, Local Binary Pattern (LBP) is computationally simple and discriminative, but sensitive to random noise. Similarly, although Gabor-based local descriptors have shown their achievements, especially in face recognition [35, 190, 259, 262], the features suffer

from the expensive computational requirement and high dimensionality. Thus, developing a robust local descriptor is still an open issue for many fields of image representation and classification, such as texture representation [28, 56, 68, 229] and face representation [28, 132].

In the field of computer vision, popular local descriptors are often employed to different problems or applications. For instance, although LBP was originally devised for texture classification, it has been applied to face recognition [4], image retrieval [216], facial expression recognition [204], etc. However, it might not always be true for a new descriptor. Facial expression recognition is a problem different from face recognition or other types of recognition. "A good face-recognition local descriptor" should represent discriminative identity information about face images, while "a good facial-expression local descriptor" should discard the subject's identity information and highlight the expression-specific information of a face. Therefore, it is important to be attentive to the nature of a problem in choosing an appropriate local descriptor.

The local descriptors, proposed in the literature, often benchmark their results against previously reported ones. However, the reliability of the benchmarking may not be high, due to the following reasons:

- A few benchmark databases were used, and the descriptors were evaluated with different databases.
- Each of the databases may have a different set of expression categories.
- Different image preprocessing techniques, e.g. face alignment, illumination, different normalization, etc., are used in experiments.
- The evaluation procedures/testing protocols, e.g. the choice of the classifier, the cross-validation scheme used, etc., are different.

- The overall experiments cannot be reproduced because not all the experimental setup is known.

In the literature, there have been several attempts to compare the performances of LBP-like descriptors using the same experimental settings. One of the most recent experimental studies on the LBP-like descriptors was conducted by Liu et al. [138], which evaluated thirty-two LBP variants for texture classification. However, there are still many other texture descriptors for facial expression recognition, which should be compared.

Kristensen et al. [124] presented an overview of "binary flavored features" for FER. Although a set of commonly used terms was defined so as to encourage consistency in terminology and to explain the current challenges, the depth of the survey in terms of performance comparison is limited. Another aim of this paper is to fill this gap by providing a comprehensive performance analysis on those recent local descriptors used for FER.

In this chapter, we compare the performances of 27 local descriptors on four popular databases with the same experimental setup, including the use of two classifiers, different image resolutions, and different numbers of sub-regions. In addition to their accuracy, other important aspects, such as face resolutions for best performances, are also studied. Moreover, we compare the results achieved by handcrafted features, e.g. histogram-based local features, with the results obtained by the "Compact Binary Face Descriptor (CBFD) [150]" and the state-of-the-art deep features. We also evaluate the robustness of the respective local descriptors in the scenario of a cross-dataset facial expression recognition problem. In our evaluation, we found that the best overall performances are obtained by Local Phase Quantization

(LPQ) and Local Gabor Binary Pattern Histogram Sequence (LGBPHS), with consistency across most of the databases used in our experiments.

The rest of the chapter is organized as follows: Section 3.2 introduces a taxonomy for histogram-based local descriptors and highlights the representative examples of the specific steps. In Section 3.3., some of the commonly-used local descriptors are explained in more details. In Section 3.4, the experimental setup is first described, then comprehensive experimental results are presented. Section 3.5 concludes the paper.

3.2. Construction of the Histogram-based Local Descriptors

Histogram-based local descriptors compute local statistical information at key points, and describe the features in a region using a histogram representation. Almost all the local statistical feature methods, as described in [246], have two main parts: statistical histogram feature extraction and statistical feature combination. Unlike [246] which divides the statistical histogram feature extraction further into three steps, we divide it into five steps in this paper, in order to describe different local descriptors in more detail. In the rest of this section, each step is explained while the corresponding representative descriptors are highlighted with their strengths and weaknesses.

3.2.1. Local variation coding

Histogram-based local-feature descriptors represent the centre pixel of a local region as a decimal number, according to its values compared to its neighbouring pixels. Regardless of the input image, local variation coding is a general method used to encode the pattern features in a local patch. For each local patch, with a given neighbourhood, a typical local variation coding has five steps, including linear

filtering, quantization, binarization, encoding and binary to decimal conversion. In the following sub-sections, these five steps will be explained in detail.

3.2.1.1. Linear filtering

The first step of local variation coding is to convolve a patch with a predefined set of linear filters. The most commonly used linear filters in histogram-based local descriptors are Kirsch [193, 194], Prewitt [3, 36, 194], Sobel [3, 36, 117, 153, 194], and Derivative-Gaussian [194].

From the computational point of view, Sobel operators are more efficient than the Kirsch operators, as less pixels and multiplications are involved. These linear filters operate on a local patch with a 3×3 mask, and custom linear filters, which consider higher-order derivatives, have also been proposed. For example, Local Arc Pattern (LAP) [97] and Local Monotonic Pattern (LMP) [168] encode the first and the second-order derivatives of a local patch in different orientations, using a set of custom filters. Although LAP and LMP can represent a bigger micro pattern with multiple radii, they use intensity values, as LBP, and are therefore sensitive to non-monotonic changes. Local Transitional Pattern (LTrP) [98] and LMP encode the transition of intensity change in different directions over a local patch. Local Derivative Pattern (LDP) [258] encodes the second and higher-order derivatives of a local patch. Although the higher-order derivatives can represent local variations with more details, the dimensionality of the resulting feature vector will become higher, as well as the computational cost.

3.2.1.2. Quantization

The second step of the local variation coding is the quantization of the linear-filter responses. The most common way of quantization used in the different descriptors is the unit step function. The local descriptors, such as LBP, Median Binary Pattern (MBP) [13], etc., quantize their filter responses using the unit-step function. However,

this will generate inconsistent binary codes in uniform and near-uniform face regions, because the filter responses may vary slightly around the threshold value, usually zero. Local Ternary Pattern (LTeP) [13], Median Ternary Pattern (MTP) [13], Gradient Directional Pattern (GDP) [3, 36], and Gradient Local Ternary Pattern (GLTeP) [117, 224] add an extra level of thresholding, which facilitates the generation of more consistent codes for local patterns in smooth facial regions, as well as highly textured regions.

Quantization of the filter responses does not necessarily result in binary values. A common way of non-binary quantization is the *k*-bin method. Histogram of Oriented Gradients (HOG) [43] and Pyramids of Histogram of Oriented Gradients (PHOG) [18] are two examples, which quantize the gradient angles to *k* intervals, and then count the gradient magnitudes of those pixels whose gradient orientations are within a specific interval. Another method of non-binary quantization of the filter responses, such as the angle or phase information, is to use the quadrant information [129, 246, 259], i.e. the 2-D Cartesian coordinate system, where four quadrants are defined by the *x*- and *y*-axes.

3.2.1.3. Binarization

After quantization, the filter responses of some descriptors, such as LBP [179], GDP [3, 36], have already been in binary form, i.e. 0 and 1. However, the other descriptors need a binarization process. The filter responses can be binarized in two ways:

Binarization by splitting into different levels of binary codes: One example of this method is LTeP [13], which has three levels after thresholding. A common way of encoding these three-level responses is to split the responses into two binary codes: "1" and "0" form one binary code, while "0" and "-1" form the other one. Therefore, two histograms are formed, and this results in a higher dimensional feature vector.

Binarization by logical operators: This method can be utilized in two different circumstances: when the quantized values are in binary form [98, 168], or not in binary form [129]. The common logical operators are "AND" [168] and "XOR" [98, 245]. These two logical operators have their unique advantages in information encoding. "AND" encodes the alikeness/sameness of the values, while "XOR" encodes the opposition between the values.

3.2.1.4. Encoding

The bits in a binary codeword correspond to the binarized responses of the different abovementioned filters. A basic way of creating a codeword is to use all the resultant binary codes to form a string. In the case of 3×3 neighborhood, i.e. 8 neighbours, each code string will be 8-bit long, which forms a decimal value between 0 and 255. LBP, LTeP, MBP, MTP and GDP utilize this basic code. Local Directional Pattern (LDiP) [100] computes the eight directional edge responses, by using the Kirsch masks. However, as the response values are not equally important in all the directions, LDiP encodes the k most prominent directions, i.e. a customized codeword. LDiP can provide more stable codes, in the presence of gray-level distortion, such as noise and non-monotonic illumination changes. High-frequency regions in a face carry more information about texture information, such as the human eye regions. Therefore, to achieve a more competent face representation, textural regions with high contrast/frequency should influence the LDiP code more. However, LDiP considers both low and high-frequency regions equally. To incorporate this importance into the LDiP codes, an extension of LDiP, named Local Directional Pattern Variance (LDiPv) [113], was proposed, which introduces the variance of the codes as weights in constructing the histograms. However, both LDiP and LDiPv consider the filter responses in absolute value, which lose the important direction information, e.g.

different transitions in a region. Furthermore, they are sensitive to rotation variations, because a fixed start position has to be defined for encoding a binary string, and they are profoundly dependent on the number of the most prominent directions considered. Local Directional Number Pattern (LDN) [194] also encodes the principal directions, i.e. the most positive and negative directions, so a more discriminative representation of directions can be achieved. Local Directional Texture Pattern (LDTP) [193] also encodes the principal directions, which discards the insignificant details that may vary on the samples belonging to the same class. However, different from the other descriptors, LDTP encodes both the principal directions and the intensity information (the intensity difference of the two principal directions). Therefore, LDTP is robust against both rotation and illumination changes.

Recently in the fields of texture classification, image retrieval, and facial feature representation, an extensive amount of customized coding schemes has been proposed [28, 68, 132, 229]. All these coding schemes aim at producing robust features, which are important for the image-classification problem.

3.2.1.5. Binary to decimal conversion

The last step of local variation coding is to convert a binary codeword into a decimal value, which represents the local pattern of the pixel under consideration. After computing the feature values for all the pixels in a patch, the statistics of these numbers, in the form of a histogram, can be used to represent the patch.

3.2.1.6. Local binary patterns and other local variation coding schemes

LBP, as a local variation coding method, has four steps as discussed previously: linear filtering, quantization with the unit step function, encoding the binary codeword, and binary to decimal conversion. LBP has also been extended to use different neighbourhood sizes, as well as uniform LBP codewords, i.e. those codewords have

no more than two transitions from 1 to 0 or 0 to 1. A codeword is non-uniform if it has more than two transitions. This idea was inspired by the fact that the uniform codewords occur more frequently than those non-uniform codewords in images.

LBP encodes the relationship between the central pixel and its neighbours. Some local descriptors extract high-order local information. A high-order descriptor can capture more detailed discriminative information. Other local descriptors also encode different distinctive spatial relationships in a local region. More information about LBP variants can be found in [138].

3.2.2. Local feature representation

LBP and other histogram-based local descriptors encode the distribution of local variation codes within a region. A frequency-based or weighted-vote-based histogram constructed for a whole face image will lose the spatial information about the patterns encoded by a local descriptor. To represent the facial features more effectively, face images are divided into a number of overlapping or non-overlapping small sub-regions. Local features extracted from the sub-regions can achieve better recognition rates than those using holistic features, such as Eigenfaces and Fisherfaces [15].

Different regions in a face carry different amount of information about an expression. To eliminate the excessive and non-informative features for face or expression recognition, weighted histogram representation has been adopted. In this representation, weights are often set according to the discriminability of the regions [204], e.g. a small weight near the image's borders, and a higher weight around the eye and mouth regions.

Another local-feature representation uses only those regions that carry salient information about facial expressions. Benitez-Garcia et al. [17] developed an algorithm to detect salient regions based on fiducial points for feature extraction. In

[220], we observed that the features extracted from the eye and mouth regions can achieve higher recognition rates than the features extracted from the sub-regions divided from a whole face.

Region covariance matrices (RCM) is a local feature that has been used in several computer-vision applications, such as face recognition, texture classification, and facial expression recognition [82, 151, 184, 223]. Unlike the previously mentioned descriptors, RCM, which computes the covariance matrix of image features, e.g. colour, coordinates, the first- and second-order gradients, etc. in a region, is not a histogram-based local descriptor. The covariance matrices computed over all specified regions are directly used to represent the image under consideration. To improve the discriminating ability of this feature, Gabor-based region covariance matrices [184] and local binary covariance matrices [82] were proposed for face and facial expression recognition, which use the Gabor and LBP features, respectively, to construct the covariance matrices. Covariance-based local descriptors are computationally expensive and require more memory [82], but covariance matrix is an effective method to combine multiple features for describing an image region.

3.2.3. Inputs to local variation coding

Most of the early descriptors extract local features from intensity information, using a local variation coding method. However, the intensity information is sensitive to noise and illumination variations. Therefore, other types of input have been considered for local variation coding. Since gradients are more stable than intensity under the presence of illumination variations, several descriptors utilize gradient information to encode local variations. For example, GDP encodes gradient angles, while GLTeP encodes gradient magnitudes.

After the successful applications of LBP, several descriptors, which are based on Gabor filtering with a predefined number of scales and orientations, have been proposed. Examples of these descriptors include Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [262], Local Gabor Directional Pattern (LGDiP) [94], and Local Gabor Transitional Pattern (LGTrP) [5]. These descriptors often encode the magnitude information of the transform, i.e. the Gabor Magnitude Image, because the magnitude information is robust to misalignment. Gabor features are robust to image variations in terms of illumination and noise, but extracting the features is computationally expensive and the resulting feature vector has a high dimensionality.

Binary Pattern of Phase Congruency (BPPC) [206] applies wavelet transform to the logarithmic Gabor features, followed by computing the phase congruency (PC). PC is a dimensionless quantity, and can be considered as the gradient where high energy values of PC occur on edges, corners, etc. Monogenic signal analysis [72], which is a 2-D generalization of the 1-D analytic signal, is an alternative method to Gabor filtering. Monogenic signal analysis can estimate the multi-resolution amplitude, orientation, and phase components of a signal, which represent the signal energetic, structural, and geometric information, respectively. One advantage of monogenic signal analysis over Gabor transformations is that it has a lower time and space complexity.

In 2010, two local descriptors, which use monogenic signal analysis, were proposed for texture classification [261] and face recognition [247], where only the monogenic phase information and both the amplitude and orientation information, respectively, are encoded. Several other this kind of local descriptors exist in the literature [132, 177, 256]. Monogenic signal analysis has also been used for

spatiotemporal facial expression recognition, with the local descriptor named "Spatiotemporal Local Monogenic Binary Patterns" [92]. However, to the best of our knowledge, Monogenic Binary Coding (MBC) [246] is the only descriptor that applied monogenic signal analysis to static facial expression images [238]. MBC encodes the amplitude (MBC_A), phase (MBC_P), and orientation (MBC_O) information separately.

Local Phase Quantization (LPQ) [180] is a local descriptor, which extracts features from the discrete Fourier transform (DFT) over an image. LPQ is robust against blur and low resolution because it quantizes the phase information in local neighbourhoods. However, LPQ requires the point spread function (PSF) to be positive and valued in the low-frequency domain. Local Frequency Descriptor (LFD) [129], which also extracts information from DFT, encodes both the magnitude and phase information using LBP and Local XNOR Pattern (LXNORP). LFD does not require PSF to be positive, and can carry more information than LPQ, but the dimension of the feature vector is doubled.

Weber Local Descriptor (WLD) [69, 70] was inspired by the Weber's Law, which states that the significance of a change in the stimuli depends on the initial value of the stimuli. WLD, which computes the differential excitation and the orientation of an image, forms a joint histogram for the differential excitation and the orientation. WLD has been applied to several problems successfully, including facial expression recognition [145]. However, WLD discards the orientation information of the differential excitation and neighbouring pixel pairs.

Recently, Jang et al. [245] proposed an extension of WLD, named Improved Weber Binary Coding (IWBC), to solve the drawbacks of WLD. IWBC generates two images, which are called the Novel Weber Magnitude Image and the Novel Weber

Orientation Image, which are then encoded using Local XOR Pattern and LBP, respectively. Although IWBC can represent a face more accurately than WLD by including the orientation information about the neighbouring pixels, it suffers from the problem of high dimensionality. To the best of our knowledge, IWBC has never been applied to the FER problem. Since IWBC has been shown to outperform WLD on the face recognition problem, so we include IWBC in our experiments to evaluate its performance as a local descriptor for FER.

3.3. Construction of the Selected Descriptors

In this section, local descriptors used in the methods presented in this thesis is explained in more details.

3.3.1. Local Binary Pattern (LBP)

LBP [178] is a texture descriptor, which creates a label for each pixel by representing it as an 8-bit binary number. The binary number is obtained by thresholding the pixel's value with those of its 3×3 neighbouring pixels. The labels in a region are then converted to decimal values to represent the region using a 256-bin histogram. LBP has been mostly used, because it is insensitive to monotonic variations caused by illumination changes, and it is computationally simple to extract the feature. In our experiments presented in this thesis, all the non-uniform LBP codes are grouped into a single bin, so a 59-bin histogram is used, rather than 256.

3.3.2. Local Phase Quantization (LPQ)

LPQ [180] was proposed as a blur-invariant texture descriptor, based on the blur-invariance property of the Fourier phase information, with the assumption that the blur is centrally symmetric. To extract the LPQ feature, the short-term Fourier transform at each pixel over a rectangular $M \times M$ neighborhood is computed. The signs of the real and imaginary parts of the Fourier coefficients, at four different frequencies, are

used to record the phase information of the pixel being considered, i.e. using a scalar quantizer. The distribution of the obtained 8-bit numbers is then represented by using a histogram.

3.3.3. Pyramid of Histogram of Oriented Gradients (PHOG)

PHOG [18] is an extension of the commonly used local descriptor, HOG [43]. PHOG represents an image, using its local shape at different scales, i.e. with different pyramid levels. After the application of the Canny edge detector, followed by calculating the oriented gradients of the edge contours using the 3×3 Sobel masks, the image under consideration is divided into spatial cells, according to the number of levels. The orientation gradients are represented by using a K-bin histogram, and the histograms of each level and each spatial cell are then concatenated to form a single feature vector. This final feature vector is of dimension $K \times \Sigma 4l$, where l is the number of pyramid levels and K is the number of bins in the histograms.

3.3.4. Weber Local Descriptor (WLD)

WLD [33] is based on Weber's Law, which states that the change of a stimulus can be recognized, if the ratio of the change to the original stimulus is larger than a certain

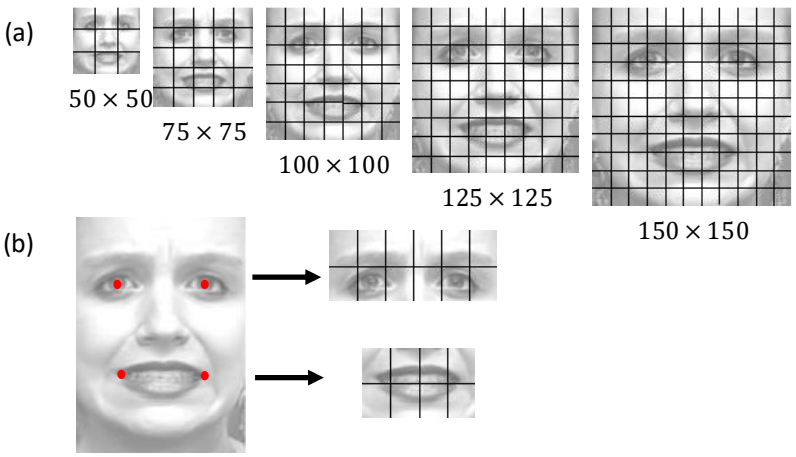

Fig. 3-1. Examples of the sub-regions used in our experiments: (a) regular sub-regions in an image, and (b) the sub-regions for the eye window and mouth window.

value. The WLD feature value depends on each current pixel value, but also puts emphasis on the difference between the current pixel and its neighboring pixels. WLD consists of two components: differential excitation that considers the ratio between the current pixel and the relative intensity difference against the neighboring pixels, and orientation that is the ratio between the vertical and horizontal gradients. Weber magnitude δ_m and orientation δ_o are defined as follows:

$$\delta_m(x_c) = \cos^{-1}\left(\alpha \sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c}\right), \text{ and} \tag{3.1}$$

$$\delta_o(x_c) = \cos^{-1} \frac{x_1 - x_5}{x_3 - x_7}, \tag{3.2}$$

$$\delta_o(x_c) = \cos^{-1} \frac{x_1 - x_5}{x_3 - x_7}, \tag{3.3}$$

where x_c denotes the center pixel and x_i represents a neighboring pixel, where $i = 0, 1, \ldots, p-1$. Then, WLD is represented as a histogram, by concatenating the two components.

3.4. Experiments

In this section, a number of histogram-based local descriptors are evaluated for facial-expression recognition, with the same experiment settings. We will first describe the experimental setup, including the benchmark databases, pre-processing, feature extraction, and classification schemes, and then analyse the experimental results.

3.4.1. Experimental setup

3.4.1.1. Databases and the corresponding numbers of expression classes

The performances of the local descriptors are compared on commonly-used, acted databases, as well as spontaneous databases. The facial-expression databases used in our experiments are BAUM-2 [67], CK+ [116], JAFFE [155], and TFEID [34].

The CK+ database, which is one of the acted facial-expression databases mostly used, contains a total of 593 posed sequences across 123 subjects. 327 of the sequences were labelled with one of the seven discrete expressions — anger, contempt, disgust, fear, happiness, sadness, and surprise. The last three frames of each sequence and their landmarks provided are used for experiments. JAFFE and TFEID are two acted face databases with six prototypical expressions and the neutral expression, which contain 213 images from 10 Japanese females and 268 images from 40 Taiwanese subjects, respectively. The BAUM-2 database consists of expression videos extracted from movies. The expressions in the videos are in the close-to-real-life conditions, i.e. with

Table 3-1. A list of the descriptors, and the corresponding feature dimensions, used in our experiments.

	Abbreviation	Descriptor Name	Dimension
1	BPPC [206]	Binary Pattern of Phase Congruency	1062
2	GDP [3, 36]	Gradient Directional Pattern	256
3	GDP2 [95]	Gradient Direction Pattern	8
4	GLTeP [117, 224]	Gradient Local Ternary Pattern	512
5	IWBC [245]	Improved Weber Binary Coding	2048
6	LAP [97]	Local Arc Pattern	272
7	LBP [179]	Local Binary Pattern	59
8	LDiP [100]	Local Directional Pattern	56
9	LDiPv [113]	Local Directional Pattern Variance	56
10	LDN [194]	Local Directional Number Pattern	56
11	LDTP [193]	Local Directional Texture Pattern	72
12	LFD [129]	Local Frequency Descriptor	512
13	LGBPHS [262]	Local Gabor Binary Pattern Histogram Sequence	256
14	LGDiP [94]	Local Gabor Directional Pattern	280 *
15	LGIP [153]	Local Gradient Increasing Pattern	37
16	LGP [96]	Local Gradient Pattern	7
17	LGTrP [5]	Local Gabor Transitional Pattern	256
18	LMP [168]	Local Monotonic Pattern	256
19	LPQ [49, 180]	Local Phase Quantization	256
20	LTeP [13]	Local Ternary Pattern	512
21	LTrP [98, 99]	Local Transitional Pattern	256
22	MBC [238, 246]	Monogenic Binary Coding	3072 *
23	MBP [13]	Median Binary Pattern	256
24	MRELBP [139]	Median Robust Extended Local Binary Pattern	800
25	MTP [13]	Median Ternary Pattern	512
26	PHOG [18]	Pyramid of Histogram of Oriented Gradients	168 *
27	WLD [133, 145]	Weber Local Descriptor	32 *

pose, age, and illumination variations. In our experiments, the image dataset, namely BAUM-2i, consisting of images with peak expressions extracted from the videos in BAUM-2 are considered. There are 1,057 face images from 250 subjects, which have seven discrete expressions and the neutral expression in BAUM-2i.

The abovementioned databases have their own characteristics in terms of where the expression images were taken, the expression classes, the race, age and gender of the participants, etc.

3.4.1.2. Descriptors

From the list of descriptors in [124], we select those descriptors based on spatial features, because this paper considers FER on static images only. The descriptors described in this paper and used in FER are also included in the experiments. Because of numerous LBP variants, only the basic LBP variants and MRELBP, which achieved the best performances in a recent comparative study for texture classifications [138],

Table 3-2. A list of selected descriptors for our experiments and a comparison of the types of input data used and the local variation coding methods.

Descriptors	Input for local variation coding	Local variation coding
IWBC	Weber magnitude, Weber orientation	Local Xor Pattern (LXP) and Local Binary Pattern (LBP)
LAP	Intensity	first- and second-order derivatives using a set of custom filters
LBP	Intensity	-
LGBPHS	Gabor image	Local Binary Pattern (LBP)
LGIP	Intensity	Horizantal and vertical responses of sobel masks
LMP	Intensity	Local And Pattern with sign information of two level intensity differences
LPQ	Phase from Fourier Transform	Quantization
LTeP	Intensity	Two-level LBP
MBC_P	Phase, orientation or amplitude from Riesz transform	Local Nand (not and) Pattern
WLD	Differential excitations and orientations	Quantization

are chosen for our comparative analysis. The other local descriptors, which are not based on LBP, but inspired by LBP, for facial expression recognition are also included. Most of the descriptors presented and evaluated in this paper belong to the sixth category defined in [138], which is called "other methods inspired by LBP". A feature learning method, named "Compact Binary Face Descriptor (CBFD) [150]", is also used in our experiments to evaluate its performance on FER, in comparison to other state-of-the-art methods.

The descriptors (represented by their abbreviations) evaluated in our experiments are listed in Table 3-1. To conduct a more detailed performance analysis, the best ten descriptors, along with the corresponding input information used and coding methods, are listed in Table 3-2. It is worth noting that MRELBP, IWBC, and CBFD are the first time applied to the FER problem.

3.4.1.3. Preprocessing and feature extraction

For the first set of experiments, face images from the different databases are scaled to different resolutions, including 50×50, 75×75, 100×100, 125×125, and 150×150. Then, features are extracted from the images with different numbers of sub-regions.

In the second set of experiments, face images from the different databases are all scaled to the size of 126×189 pixels, with a distance of 64 pixels between the two eyes. To locate the eye and mouth windows, the facial landmarks, i.e. the eye and mouth corners, are used. If facial landmarks are not provided for a database, the required facial-feature points are marked manually. The eye window and the mouth window are further divided into 12 and 8 sub-regions, respectively. Fig. 3-1 shows examples of selected sub-regions in both the first and the second set of experiments.

3.4.1.4. Dimensionality reduction and classification

In the first two sets of experiments, the local descriptors listed in Table 3-1 were first extracted. Then, the subspace-learning method, Soft Locality Preserving Projection (SLPM) [222], is applied for manifold learning and dimensionality reduction. SLPM is a graph-based subspace-learning method, which uses the *k*-neighborhood information and the class information. The key feature of SLPM is that it aims to

Table 3-3. The recognition rates for different resolutions, different numbers of sub-regions, on the CK+ database. "-" means that the corresponding results are unavailable because the dimensionality of the feature vectors are too high for experiments.

Database		CK+ – LOSO – 6-class											
Resolution		50x50	75x75		100x100		125x125			150x150			
# of sub-regions		3x3	3x3	5x5	5x5	7x7	5x5	7x7	9x9	5x5	7x7	9x9	11x11
1	BPPC [206]	85.33	85.76	90.40	89.75	**90.83**	90.40	90.40	89.86	87.06	90.40	89.21	89.86
2	GDP [3, 36]	74.54	75.73	83.82	86.30	86.62	86.30	85.98	**86.84**	85.65	85.76	86.08	86.08
3	GDP2 [95, 224, 232]	57.71	57.39	83.06	81.98	90.51	82.20	89.97	92.45	83.06	89.43	94.28	**94.82**
4	GLTP [117]	82.85	85.98	91.15	92.66	92.66	92.34	91.69	91.05	92.13	**93.64**	91.59	92.99
5	IWBC [245]	88.67	90.72	91.69	90.51	93.42	91.15	93.10	**94.82**	90.83	92.13	93.31	92.99
6	LAP [97]	83.17	80.26	89.75	89.21	91.69	90.29	91.26	93.42	91.05	91.05	93.42	**94.17**
7	LBP [179]	84.68	84.03	92.23	91.48	91.69	91.91	93.53	93.85	91.80	93.20	93.74	**95.25**
8	LDiP [100]	68.72	71.52	86.73	85.44	89.00	86.08	89.54	89.54	84.68	89.64	89.32	**89.75**
9	LDiPv [70, 113]	68.93	71.20	83.17	82.85	86.95	83.50	87.70	89.00	85.11	85.98	88.67	**89.21**
10	LDN [194]	80.91	82.96	88.46	88.24	90.29	90.40	91.15	92.66	90.83	90.40	**92.66**	91.91
11	LDTP [193]	82.74	80.69	85.87	85.65	90.08	86.19	89.75	93.10	83.60	87.06	**93.53**	89.75
12	LFD [129]	86.62	82.09	**90.61**	88.78	90.51	87.38	89.43	89.21	86.62	88.57	88.78	87.49
13	LGBPHS [262]	86.19	87.27	92.02	92.88	92.99	91.26	90.72	91.48	90.29	89.75	91.48	**95.25**
14	LGDiP [94]	71.09	69.15	75.19	80.15	79.72	77.35	78.86	79.07	80.04	**83.39**	80.80	78.64
15	LGIP [153]	83.28	84.14	93.20	91.59	92.88	91.69	92.66	93.96	91.69	92.34	93.31	**95.15**
16	LGP [96]	50.70	51.13	79.50	77.13	87.38	76.27	85.33	92.45	76.27	86.41	93.10	**93.31**
17	LGTrP [5]	48.76	50.16	62.46	64.51	68.72	65.26	64.40	66.67	62.03	**69.26**	64.40	68.82
18	LMP [168]	86.30	87.38	90.83	92.23	92.34	92.99	95.04	**95.25**	91.59	94.50	93.85	93.96
19	LPQ [49, 180]	90.08	92.45	93.96	**94.39**	93.31	93.31	94.28	94.17	92.77	93.74	93.74	93.64
20	LTeP [13]	88.35	89.10	91.80	92.45	93.31	92.99	93.96	**95.69**	92.56	94.50	95.04	94.93
21	LTrP [98, 99]	74.76	75.73	85.65	85.44	88.13	84.36	87.70	88.24	87.38	**89.54**	89.43	87.27
22	MBC_A [238, 246]	**92.56**	89.54	89.97	90.51	88.35	89.43	89.43	-	90.08	89.32	-	-
23	MBC_P [238, 246]	88.89	89.32	92.88	**94.28**	90.51	91.80	93.42	-	92.56	92.45	-	-
24	MBC_O [238, 246]	88.89	87.81	92.02	91.80	91.37	90.94	**92.56**	-	**92.56**	92.02	-	-
25	MBP [13]	83.71	82.85	90.08	90.61	90.94	91.05	91.48	93.53	90.40	91.69	93.42	**94.07**
26	MRELBP [139]	87.70	88.13	92.13	90.29	92.45	90.72	92.02	**93.53**	91.05	92.88	92.88	93.31
27	MTP [13]	90.72	87.92	90.51	90.08	89.97	89.64	89.97	**92.77**	89.43	89.00	91.59	90.94
28	PHOG [18]	87.59	89.54	89.54	90.29	89.32	89.00	**91.80**	90.72	89.21	90.83	90.51	90.40
29	WLD [133, 145]	81.45	79.61	91.37	90.94	92.23	90.83	93.10	**95.90**	92.23	93.31	95.47	95.47

control the level of spread of the different classes, because the spread of the classes in the underlying manifold is closely connected to the generalizability of the learned subspace. In our experiments, we employ SLPM for dimensionality reduction and for increasing the discriminative ability of the extracted features. Finally, the nearest neighbour (NN) classifier is used for classification. The third set of experiments were conducted, with the best setting for each of the databases, using the Support Vector

Table 3-4. The recognition rates for different resolutions and different numbers of sub-regions, on the BAUM-2i database. "-" means that the corresponding results are unavailable because the dimensionality of the feature vectors are too high for experiments.

Database		BAUM-2i – 10-fold – 6-class											
Resolution		50x50	75x75		100x100		125x125			150x150			
# of sub-regions		3x3	3x3	5x5	5x5	7x7	5x5	7x7	9x9	5x5	7x7	9x9	11x11
1	BPPC [206]	51.18	52.48	53.90	56.26	58.98	54.37	**59.10**	57.45	56.15	55.67	57.21	54.85
2	GDP [3, 36]	40.78	45.39	50.71	46.10	**52.84**	45.98	50.83	51.89	46.22	50.24	51.89	49.76
3	GDP2 [95, 224, 232]	23.88	26.12	29.91	27.90	37.35	28.01	40.54	48.94	28.84	40.66	48.11	**53.31**
4	GLTP [117]	50.00	54.14	57.57	55.44	59.57	53.90	59.10	58.27	54.73	59.22	57.57	**60.05**
5	IWBC [245]	55.67	57.33	**59.22**	58.39	58.16	56.97	57.92	57.57	57.09	57.21	56.86	56.26
6	LAP [97]	46.22	44.80	53.66	48.70	51.77	48.35	50.24	56.03	49.05	50.24	56.03	**56.38**
7	LBP [179]	48.35	48.23	54.26	54.02	56.62	53.07	56.86	58.63	53.31	54.61	56.62	**59.46**
8	LDiP [100]	27.30	30.61	43.62	45.39	53.07	48.70	52.25	53.66	45.15	52.60	**56.03**	**56.03**
9	LDiPv [70, 113]	24.11	28.25	40.19	36.05	49.88	40.54	48.94	52.25	40.78	48.82	52.84	**53.90**
10	LDN [194]	37.23	34.16	48.11	47.52	51.06	47.28	54.14	55.67	47.16	54.61	55.91	**60.99**
11	LDTP [193]	30.26	34.04	48.11	43.38	48.82	42.79	46.81	49.41	44.33	47.64	46.81	**51.06**
12	LFD [129]	46.69	44.56	53.90	50.59	57.21	51.77	56.74	57.57	51.06	54.73	56.50	**57.92**
13	LGBPHS [262]	49.41	50.00	56.62	57.57	59.46	59.22	60.28	60.76	59.81	61.11	**62.41**	57.92
14	LGDiP [94]	30.97	32.62	39.60	42.67	44.09	41.25	46.10	**47.52**	39.83	42.55	44.44	43.85
15	LGIP [153]	30.50	30.97	49.88	47.04	54.61	49.17	54.02	56.74	48.11	52.96	55.44	**58.39**
16	LGP [96]	23.40	21.75	23.52	26.36	31.32	25.41	32.98	42.43	25.30	30.02	41.84	**46.22**
17	LGTrP [5]	24.47	23.05	31.44	30.38	31.09	32.03	34.04	35.11	31.68	36.29	**40.07**	36.29
18	LMP [168]	49.76	50.35	55.91	56.86	58.75	56.50	58.16	**60.64**	52.36	55.32	58.27	60.17
19	LPQ [49, 180]	56.38	56.03	61.35	61.35	60.28	59.46	**61.47**	61.23	57.68	59.57	60.28	**61.47**
20	LTeP [13]	52.36	50.59	55.32	52.96	58.87	52.96	59.46	59.22	54.02	59.57	**60.28**	**60.28**
21	LTrP [98, 99]	35.34	38.89	42.79	46.22	51.06	45.15	49.17	52.01	46.57	50.47	51.77	**53.78**
22	MBC_A [238, 246]	56.62	56.62	**59.81**	57.57	59.34	56.38	58.63	55.08	56.03	55.67	55.20	-
23	MBC_P [238, 246]	56.03	54.96	59.93	59.81	61.58	59.57	61.47	61.94	59.46	60.87	**62.06**	-
24	MBC_O [238, 246]	57.68	55.79	61.35	**61.82**	60.99	60.05	60.17	61.23	58.27	60.99	60.40	-
25	MBP [13, 141, 142]	43.62	47.04	54.37	54.37	54.49	53.55	55.32	**59.46**	52.60	53.43	55.08	57.33
26	MRELBP [139]	46.34	48.70	55.56	56.86	57.68	57.80	57.92	**59.34**	57.45	57.57	58.98	**59.34**
27	MTP [13]	43.97	41.96	51.54	50.24	**54.02**	47.87	53.78	51.65	42.91	51.65	52.13	52.60
28	PHOG [18]	47.52	50.35	51.42	53.43	53.90	51.77	54.26	52.96	54.14	54.02	**54.61**	53.43
29	WLD [133, 145]	30.26	24.82	51.77	46.10	57.80	47.52	55.44	56.15	46.93	54.73	55.79	**58.75**

Machine (SVM) classifier, with the linear kernel. The results are then compared to those based on the nearest neighbour classifier. Two different cross-validation

Table 3-5. The comparison of recognition rates obtained by the selected local descriptors on the BAUM-2i database (the best of sub-regions) using 10-fold cross validation. 6-class: AN, DI, FE, HA, SA, and SU. 7-class: AN, CO, DI, FE, HA, SA, and SU. 8-class: AN, CO, DI, FE, HA, NE, SA, and SU.

	BAUM-2i		
	6-class	7-class	8-class
IWBC	59.22	55.53	52.53
LAP	56.38	54.97	49.00
LBP	59.46	58.32	52.44
LGBPHS	**62.41**	57.99	54.15
LGIP	58.39	54.75	49.86
LMP	60.64	57.54	52.53
LPQ	61.47	**58.99**	**54.73**
LTeP	60.28	57.21	52.63
MBC_P	62.06	58.10	54.25
WLD	58.75	54.41	50.53

Table 3-6. The recognition rates of selected local descriptors on the CK+ database, with 6 classes (AN, DI, FE, HA, SA, and SU) and 7 classes (AN, CO, DI, FE, HA, SA, and SU), using LOSO.

	CK+			
	Eye and mouth windows		Best of sub-regions	
	6-class	7-class	6-class	7-class
IWBC	94.61	93.68	94.82	94.50
LAP	91.37	91.44	94.17	92.86
LBP	93.31	92.56	95.25	93.99
LGBPHS	92.23	90.72	95.25	93.99
LGIP	91.26	92.35	95.15	94.50
LMP	**94.71**	94.19	95.25	**94.90**
LPQ	94.61	**94.90**	94.39	94.19
LTeP	93.53	93.17	95.69	94.80
MBC_P	91.69	89.40	94.28	92.46
WLD	93.31	91.44	**95.90**	94.80

schemes are adopted in our experiments: Leave-One-Subject-Out (LOSO) to encourage the reproducibility of the experiments, and 10-fold cross-validation, which is used when there are sufficient number of images for each subject in the database,

i.e. BAUM-2i. Furthermore, both the 10-fold and LOSO cross-validation schemes are used for comparison on the JAFFE and TFEID databases.

3.4.2. Experimental results

In this section, the experiment results on the four facial-expression databases (BAUM-2i, CK+, JAFFE, TFEID) under different experimental settings are presented and

Table 3-7. The recognition rates of the selected best local descriptors on the JAFFE database.

	JAFFE			
	Eye and mouth windows		Best of sub-regions	
	LOSO	10-fold	LOSO	10-fold
IWBC	58.69	88.73	65.73	90.61
LAP	**68.08**	90.61	68.54	94.84
LBP	61.50	86.38	65.73	93.43
LGBPHS	63.38	**93.90**	**71.83**	93.90
LGIP	62.91	87.32	66.20	93.90
LMP	60.09	85.92	67.14	93.43
LPQ	67.61	92.02	69.95	93.43
LTeP	61.03	89.20	62.44	94.37
MBC_P	63.38	92.96	66.67	93.90
WLD	63.38	86.85	69.01	**96.24**
CBFD	66.20	89.67	-	-

Table 3-8. The comparison of recognition rates obtained by the selected local descriptors on the TFEID database.

	TFEID			
	Eye and mouth windows		Best of sub-regions	
	LOSO	10-fold	LOSO	10-fold
IWBC	89.55	90.67	92.91	91.79
LAP	91.04	91.04	94.40	**95.15**
LBP	91.79	92.54	93.66	94.78
LGBPHS	**94.40**	91.04	**95.15**	93.66
LGIP	89.18	86.19	94.78	93.28
LMP	91.42	92.16	94.03	94.03
LPQ	**94.40**	**93.28**	94.03	94.40
LTeP	90.30	92.16	94.40	**95.15**
MBC_P	94.30	91.79	94.40	93.66
WLD	92.16	91.42	94.78	94.40
CBFD	93.66	92.16	-	-

discussed. The experiments are designed to measure the performances of the respective descriptors, for face images at different resolutions and divided into different sub-regions, and with different classifiers.

3.4.2.1. Performance analysis for varying resolution and number of sub-regions

All the face images are first aligned based on the positions of the two eye pupils, and cropped to the different resolutions. For each resolution, face images are divided into different numbers of sub-regions, say $l \times l$, where l varies from 3 to 11.

Table 3-3 and Table 3-4 present the results on CK+ and BAUM-2i for all the descriptors. As observed from the results shown in **Table 3-3** and Table 3-4, in general, the classification performances improve when the image resolution and the number of sub-regions increase. Therefore, higher resolution and more sub-regions lead to better classification performances. However, with more sub-regions, the feature dimension will become very high. In other words, the better performance is at the expenses of higher computational requirements.

For more detailed performance analysis, the best ten descriptors, which have achieved promising results, were chosen to repeat the first set of experiments on the four databases separately with different numbers of expression classes, as well as the two different classification schemes. **Table 3-5** shows the best classification rates on BAUM-2i with different numbers of expression classes. In **Table 3-6** to **Table 3-8**, the columns named "best of sub-regions" show the best classification rates for the number of sub-regions being used. We only show the best results, otherwise there are too many data to be shown.

3.4.2.2. Performance analysis of the eye and mouth regions

The second set of experiments was conducted with the features extracted from the eye and mouth windows of face images. The CK+, JAFFE and TFEID databases are used

to test the performances of the respective features extracted from the eye and the mouth regions. The BAUM-2i database is not used because it consists of images in the wild. Labelling the facial landmarks is a complicated task. In Table 3-6 to Table 3-8, the two columns under "eye and mouth windows" show the classification accuracies of the selected features, using the LOSO and 10-fold cross-validation

Table 3-9. The comparison of the recognition rates obtained with features extracted from the eye and mouth regions by the nearest neighbor classifier (NN) and SVM classifier using LOSO.

	CK+		JAFFE		TFEID	
	SLPM + NN	SVM	SLPM + NN	SVM	SLPM + NN	SVM
LGBPHS	92.23	91.91	63.38	61.50	94.40	94.40
LPQ	94.61	**94.93**	**67.61**	67.14	94.40	94.40

schemes.

As observed from the tables, using the features extracted from the eye and the mouth windows achieves lower classification accuracies than that using features extracted from the sub-regions of whole face images. However, for the results based on sub-regions, we show the best classification accuracies achieved for the five different resolutions and the five different numbers of sub-regions. Furthermore, each descriptor achieves the best performance on a different resolution and a different number of sub-regions. Experiment results show that there are not a particular resolution and a particular number of sub-regions that can work the best for all the descriptors.

3.4.2.3. Performance analysis of the classifiers

Table 3-9 presents the experiment results obtained with the NN and the SVM classifiers. We can observe that both LGBPHS and LPQ achieve similar performances in the use of NN and SVM. However, the NN classifier can achieve equal or higher

performance than the SVM classifier if a supervised dimensionality reduction method is employed. In our experiments, we utilize SLPM for dimensionality reduction.

Table 3-10. The comparison of the recognition rates of the ten selected descriptors on cross-dataset facial expression recognition, with features extracted from the eye and mouth windows.

Trained on	CK+		JAFFE		TFEID	
Tested on	JAFFE	TFEID	CK+	TFEID	CK+	JAFFE
IWBC	25.00	33.77	34.52	42.98	38.30	30.98
LAP	29.89	32.46	24.16	42.98	**45.85**	26.63
LBP	21.74	34.65	29.02	44.74	35.81	28.26
LGBPHS	18.48	33.33	37.22	60.09	39.48	44.02
LGIP	**30.43**	31.14	31.18	41.23	42.61	31.52
LMP	29.35	32.46	**37.32**	48.25	37.32	23.37
LPQ	19.57	**38.16**	32.58	50.44	42.07	35.33
LTeP	19.57	35.96	25.03	25.88	26.86	35.87
MBC_P	25.54	31.14	37.00	**63.60**	38.83	**47.28**
WLD	15.76	35.09	27.18	32.89	42.61	24.46

3.4.2.4. Performance analysis of cross-dataset facial expression recognition

In real-life applications, query samples are often different from the training samples in terms of uncontrolled variations such as illumination. Therefore, it is important for a local descriptor to have a good generalization power, and the descriptor can still achieve a good performance when the training and test sets are from different databases. In this paper, we also conduct experiments to test the robustness and accuracy of the best selected descriptors in the scenario of cross-dataset FER.

Table 3-10 shows the experiment results when the training and the testing sets are two different datasets, which have different acquisition conditions. As you can observe in Table 3-10, the recognition rates for the 6 basic emotions decrease significantly, because cross-dataset FER is a challenging task. Although no local descriptor can perform consistently better than the others, MBC_P achieves the highest recognition rates when the model is trained using JAFFE while tested on TFEID, and vice versa. MBC_P uses monogenic signal analysis to estimate the phase component of the

Table 3-11. The comparison of recognition rates of deep learning methods and the best recognition rate obtained with handcrafted features.

Method	Feature Type	Accuracy (%)
3DCNN-DAP [140]	Deep features	92.4
BDBN [142]	Deep features	96.7
STM-ExpLet [141]	Deep features	94.2
DTAGN [111]	Deep features	97.3
Inception [169]	Deep features	93.2
PPDN [267]	Deep features	97.3
LFC + FFD [17]	Handcrafted features	97.9
FN2EN [52]	Deep features	**98.6**
LPQ-SLPM-NN	Handcrafted features	**95.9**

images, which represents the images' geometric information. Since the JAFFE database consists of images of Japanese women, while TFEID consists of images of Taiwanese men and women, we can observe that the phase information of the monogenic signal analysis is insensitive to cross-cultural face representation for FER.

3.4.2.5. Comparison with deep features

Recently, convolutional deep neural networks have been applied to FER [52, 111, 140-142, 169, 267]. Table 3-11 presents the performances of deep learning methods applied on the CK+ database. 3DCNN-DAP [140] adapts a deformable parts learning component to detect discriminative facial action parts for spatiotemporal FER, where a Boosted Deep Belief Network [142] (BDBN) is used to learn and select the expression-related facial features to develop a strong classifier in a unified loopy framework iteratively. Iterative learning of the BDBN framework strengthens the discriminative capabilities of the features. STM-ExpLet [141] learns a spatiotemporal manifold (STM) from low-level features from each expression video clip, followed by learning a universal manifold model that statistically unify all the STMs. With this method, expression videos are also aligned. Different from these methods, DTAGN [111] trains two models, with temporal geometry features and temporal appearance

features, respectively, from image sequences, and these two features are complementary to each other. In [169], a network, which consists of two convolutional layers with max pooling and four inception layers, was proposed. The network was evaluated for its generalizability by experiments, with cross-database classification. To boost the generalizability of learning, [267] presented a peak-piloted deep network (PPDN), which uses the samples with high-intensity expressions to supervise the samples with low-intensity expression that are hard to classify. Until now, FN2EN [52], which uses a two-stage training algorithm, achieved the best performance on the CK+ dataset. FN2EN, in the first stage, trains the convolutional layers, whose outputs from the last pooling layer are used to supervise the expression net in the second stage.

As observed in Table 3-11, LPQ with NN outperforms several deep learning methods. LFC + FFD [17] is also a histogram-based feature extraction method, which achieves higher classification accuracy than all the listed methods, except FN2EN [52]. To the best of our knowledge, FN2EN achieves the highest classification accuracy on the CK+ database. However, expensive computational cost is a drawback of most of the methods based on deep convolutional neural networks. Furthermore, the CK+ database consists of images taken under controlled environments, i.e. posed expressions and the number of expression samples are limited in the CK+ database. These factors direct us the need of a large-scale facial-expression database in the wild. There have been several attempts to collect facial-expression images in the wild [50, 55, 165]. [188] and [170] are two recently published databases, which contain large-scale face images with varying expressions. These databases will be very useful for FER based on deep learning.

3.5. Conclusion

This chapter provides a systematic review and analysis of current histogram-based local feature descriptors, which have been applied for facial-expression recognition. The weaknesses and strengths of the existing descriptors, as well as their underlying connections, have also been discussed and analysed. Then, a comprehensive evaluation of the performances of different descriptors for facial-expression recognition is conducted and presented. In total, 27 local descriptors have been applied on four facial-expression databases, under the same experimental settings. The robustness of the respective local descriptors is tested under different conditions, such as varying image resolutions and number of sub-regions, and the classifiers. Moreover, a brief performance comparison with seven recent deep features and two handcrafted features has been conducted.

Several remarks from the experiment results are listed as follows:

- The databases have different characteristics, which affect the choice of the ideal descriptor for a particular database. Even the number of expression classes can also affect the performances of the descriptors.
- The results show a trade-off between the number of sub-regions and the overall classification accuracy. The use of the eye and the mouth windows decreases the number of sub-regions and the dimensionality of the resulting feature vectors, with a slight loss in terms of accuracy.
- The resolution of face images and the number of sub-regions are the two most important factors that affect the overall classification accuracies.
- The highest classification accuracies are obtained mostly by LGBPHS and LPQ. This shows that Gabor wavelets and phase information are important

features for representing expression-specific information. However, we should keep in mind that Gabor features suffer from high computational cost.

- According to the comprehensive analysis shown in this paper, the best local descriptors for FER, by considering the feature length, computational cost, and the classification accuracy simultaneously, is LPQ.
- Deep neural-network-based methods indeed can achieve excellent classification accuracies on FER. However, these methods also suffer from time and space complexities as LGBPHS.

In conclusion, our comprehensive experiment results show that the trade-off between the computational cost and the classification accuracy still exists today.

Chapter 4. Region-based feature fusion for facial-expression recognition

4.1. Introduction

During the last several decades, facial-expression recognition (FER) has become a popular research topic in the field of computer vision. Approaches proposed for increasing the accuracy of facial-expression recognition have been extended from the feature level to the decision level. In the literature, features extracted from two or more modalities – such as audio, video or image – have been combined for FER [240]. On the other hand, features extracted from different regions of the same face images can also be fused to enhance FER accuracy [233, 271]. There are three levels for feature fusion in information processing: pixel level, feature level, and decision level [211]. Fusion at the decision level is achieved by combining multiple classifiers. [39, 93, 109] have applied decision-level fusion to face recognition and handwritten character recognition.

Canonical Correlation Analysis (CCA), proposed by H. Hotelling in 1936 [91], is a powerful tool for exploring the linear correlation between two feature sets. In [152], CCA has been used to fuse two different feature sets and to improve the recognition rate for facial expression. Three-dimensional CCA [78] and CCA with spectral component selection [272] have also been proposed to improve CCA for facial-expression recognition.

In this chapter, we propose to extract features from two non-overlapping facial regions, and to fuse these features in two different levels. In the first level, the features from the eye and the mouth windows are concatenated to form an augmented feature for facial expression. If the expression of a query input cannot be determined

confidently, the features in the second level of classification are obtained by fusion using CCA, which can explore and enhance the correlation between the eye and the mouth features.

The Support Vector Machine (SVM) [41] is used extensively for classification as it can achieve a satisfactory level of accuracy when dealing with high-dimensional data. SVM was first proposed for binary classification, i.e. to classify two classes, say, labelled as -1 or 1. SVM has also been extended for multiclass classification. One-versus-all classification is the simplest approach, which trains one classifier for each class, while the training samples of all the other classes are grouped to form the other class. A decision on a testing input is made if only one of the classifiers produces a positive result, while all the other classifiers reject the input [41]. However, it is not unusual to have more than one classifier produce a positive result for an input. To tackle this issue, the output of each of the SVM-based classifier is considered to be the probability of the testing input belonging to the corresponding class [228]. The testing input is assigned to the class whose corresponding classifier's output has the highest value. In our algorithm, we propose a two-level classification scheme for facial-expression recognition in order to improve the one-versus-all classification.

The remaining parts of this chapter are organized as follows. Details of our proposed method are presented in Section 4.2. In Section 4.3, the experimental setup is described and the experiment results are given. Section 4.4 gives a conclusion to the paper.

4.2. Details of Our Approach

In our approach, the eye window and mouth window, which are the most salient regions in faces for representing facial expressions, are considered for feature

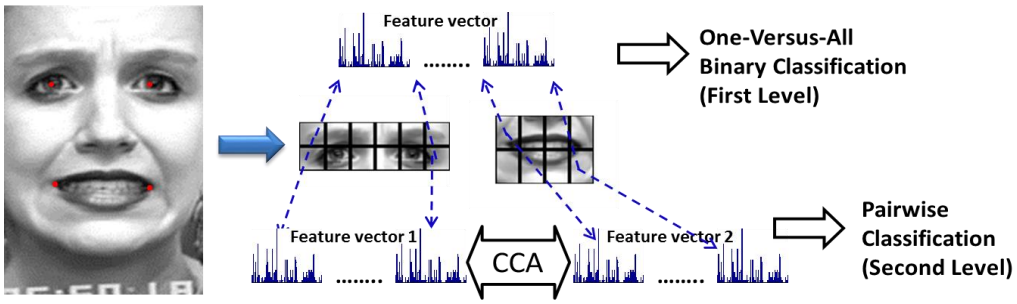

Fig. 4-1. Our region-based feature-fusion scheme for facial-expression recognition.

extraction. In addition, these two regions should have a high correlation in describing a specific facial expression.

Before setting the facial regions, all of the face images under consideration are scaled and aligned based on the positions of the two eyes. Then, the eye window and the mouth window are set according to the eyes' positions and the positions of the mouth corners, respectively. Fig. 4-1 shows the eye and mouth windows, and the red dots represent the positions of the eyes and the mouth corners. In our method, all images are normalized to a size of 126×191 pixels, with 64 pixels between the two eyes. We can see that the eye window and the mouth window can jointly show the expression. In order to represent the facial features more effectively, the eye window is further divided into 12 sub-regions, while the mouth window into 8 sub-regions, as illustrated in Fig. 4-1. In our method, the Local Phase Quantization (LPQ) and the Pyramid of Histogram of Oriented Gradients (PHOG) descriptors are extracted separately from each of the 20 sub-regions (for a detailed explanation of PHOG and LPQ, please refer to Section 3.3.2-3). LPQ is used with a window size of 3×3 and α = 0.7 where the parameters of PHOG are set L = 2 and K = 8, and the angular range used is [0, 360].

Principal Component Analysis (PCA) [90] is first applied to the features extracted from the eye window and the mouth window separately so as to decrease the feature

dimensionality, with 99% of the energy retained. Then, these local LPQ and PHOG features from the two windows are concatenated to form a feature vector, which is used in the first-level classification. After labelling the samples exhibiting one emotion as positive and all others as negative, a binary classifier for each emotion is trained using SVM – a total of 7 classifiers. The decision on a query input is made based on the output values of the SVM classifiers. If a decision cannot be made using the first-level classifiers, the three most probable facial expressions, i.e. the corresponding classifiers having the highest output values, will be selected to perform the second-level classification. In our algorithm, the second-level classification is conducted if the difference between the highest and the second-highest output is smaller than 0.1. As we consider 7 different facial expressions, and 3 of the expressions are considered in the second level, there are 35 groups of classifiers to be trained for the second-level classification. Three classifiers are trained for each group, and the features used are generated by fusing the eye and the mouth features using CCA. The purpose of performing this fusion is to obtain a richer and more distinctive description for each expression.

In the rest of this section, the second-level classification is described in detail, and the CCA is presented for fusing the features from the eye window and the mouth window to produce a coherent feature for expression representation and recognition; this can produce coherent LPQ and PHOG features for the eye and the mouth windows in the second-level classification.

4.2.1. Second-level classification and feature fusion

As described above, the testing input is assigned to the class whose corresponding classifier's output has the highest value. To be confident of a correct classification, it is found by experiments that the highest output value should be higher than, say, 0.6

out of 1.0, and the other classifiers' outputs should be much smaller. However, it is difficult to make the decision in some cases, such as 1) more than one classifier with high probabilities can have similar values, say, around 0.3 – 0.6; and 2) all the classifiers' output probabilities have similarly small values, say, smaller than 0.2. This means that there is no strong preference for a particular class. Under either of these two scenarios, it is probable that the input will be classified wrongly.

To solve this problem, we propose adding an additional classification layer – i.e. the second-level classification – where the 3 classes with the highest output values and whose values differ by no more than a certain threshold td are used to construct the classifier. The classification structure is illustrated in Fig. 4-1. In the first level, the one-versus-all scheme with SVM is employed to learn a classifier for each of the facial expressions to be recognized. If the classifier with the highest output has its value larger than 0.6, and 0.1 higher than the second-highest output, the query input is then declared to be of the expression of the classifier with the highest output. However, if this is not the case, the expression of the input is not obvious, and the second-level classifiers with more sophisticated features are used. CCA is employed to fuse the features from the eye window and the mouth window. The one-versus-one scheme, which is a pairwise classification scheme, is used to further analyse the distinction between classes in the second-level classification.

4.2.1.1. Canonical Correlation Analysis (CCA)

CCA is a subspace method that maximizes the correlation between two feature vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ by the linear mappings $\boldsymbol{x}_a = \mathbf{W}_a^{\mathrm{T}}\boldsymbol{x}$ and $\boldsymbol{y}_b = \mathbf{W}_b^{\mathrm{T}}\boldsymbol{y}$, respectively, where $\mathbf{W}_a$ and $\mathbf{W}_b$ are the projection matrices, such that the projections achieve maximum correlation. The correlation expression can be written as follows:

$$\max \rho = \frac{\widehat{\mathrm{E}}[\boldsymbol{x}_a \boldsymbol{y}_b]}{\sqrt{\widehat{\mathrm{E}}[\boldsymbol{x}_a^2]\widehat{\mathrm{E}}[\boldsymbol{y}_b^2]}} \tag{4.1}$$

$$\rho = \frac{\hat{\mathrm{E}}[\mathbf{w}_a^{\mathrm{T}} \boldsymbol{x}_a \boldsymbol{y}_b^{\mathrm{T}} \mathbf{w}_b]}{\sqrt{\hat{\mathrm{E}}[\mathbf{w}_a^{\mathrm{T}} \boldsymbol{x}_a \boldsymbol{x}_a^{\mathrm{T}} \mathbf{w}_a] \hat{\mathrm{E}}[\mathbf{w}_b^{\mathrm{T}} \boldsymbol{y}_b \boldsymbol{y}_b^{\mathrm{T}} \mathbf{w}_b]}} \tag{4.2}$$

$$\rho = \frac{\mathbf{w}_a^{\mathrm{T}} \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a^{\mathrm{T}} \mathbf{C}_{aa} \mathbf{w}_a \mathbf{w}_b^{\mathrm{T}} \mathbf{C}_{bb} \mathbf{w}_b}} \tag{4.3}$$

where $\mathbf{C}_{aa}$ and $\mathbf{C}_{bb}$ are the within-set covariance matrices of x and y, respectively, while $\mathbf{C}_{ab}$ is their between-sets covariance matrix.

4.3. Experimental Protocol and Results

4.3.1. Experimental Protocol

In our experiments, the extended Cohn-Kanade (CK+) [116] database is used because intense emotions are included in its images. This database has also been used for evaluation by many other FER methods. The CK+ dataset contains a total of 593 posed sequences across 123 subjects. 322 of the sequences have been labelled with one of the seven discrete emotions that are considered in our experiments; these are anger, disgust, contempt, fear, happiness, sadness and surprise. Each sequence starts with a neutral face and ends with a frame of peak expression. The last frame of each sequence, and its landmarks, are used for recognition. In our experiments, LIBSVM [29], which is a MatLab tool for SVM, was used.

4.3.2. Experiment Results

To evaluate the robustness of the proposed method, we compare our proposed methods with a number of existing FER methods [81, 123, 135, 204, 207, 241, 270]. For the existing methods considered in our experiments, we simply align the faces based on the eye positions, divide the faces into sub-regions, and concatenate the features from all the sub-regions to form an augmented feature, The results are shown in the first column of Table 4-1. For our proposed algorithm, only the features from the eye

Table 4-1. Recognition rates (in %) of different methods on the CK+ dataset.

Descriptors	Face Alignment	2 regions	2 regions with 2nd level classification
LPQ	92.86%	93.79%	**95.03%**
PHOG	86.65%	90.37%	**91.30%**

window and the mouth window are considered, rather than the whole face. For both methods, linear SVM is employed to learn the classifiers.

To improve the reliability of the experiment results, samples in the dataset are shuffled, and then divided into 7 folds. In each fold, there is at least one image with each of the seven expressions. Thus, in our experiments, 7-fold cross-validation has been conducted.

As shown in Table 4-1, an increase of 3% in terms of average accuracy can be achieved for both the LPQ and the PHOG descriptors if the conventional method is replaced by our proposed fusion method in the second level of algorithm which also uses specific regions for the eyes and mouth. Moreover, we have compared our method

Table 4-2. Comparison of the performances of some current facial-expression recognition methods.

Methods	Classes	Evaluation	Recognition Rate (%)
2008 [123]	7	5-fold	92.3
2009 [204]	7	10-fold	91.4
2010 [135]	6	LOSO	96.33
2011 [207]	6	LOSO	92.97
2012 [270]	6	10-fold	89.89
2012 [81]	7	10-fold	91.51
2013 [241]	6	5-fold	89.2
the proposed method (PHOG)	7	7-fold	**91.30**
the proposed method (LPQ)	7	7-fold	**95.03**

with other, state-of-the-art methods on the CK+ dataset, as shown in Table 4-2. Our proposed method outperforms the other methods in terms of the average accuracies.

Table 4-3. Confusion matrix for LPQ with CCA.

%	Predicted Labels						
	AN	**CO**	**DI**	**FE**	**HA**	**SA**	**SU**
AN	**86.7**	0.0	6.7	0.0	0.0	6.7	0.0
CO	5.6	**88.9**	0.0	0.0	5.6	0.0	0.0
DI	1.7	0.0	**98.3**	0.0	0.0	0.0	0.0
FE	0.0	0.0	0.0	**92.0**	8.0	0.0	0.0
HA	0.0	0.0	0.0	1.4	**98.6**	0.0	0.0
SA	10.7	0.0	0.0	0.0	0.0	**89.3**	0.0
SU	0.0	1.3	0.0	0.0	0.0	0.0	**98.7**

The confusion matrix of our method based on LPQ with CCA in the second level is shown in Table 4-3. Except for the expressions contempt (CO), fear (FE) and sadness (SA), our method achieves 98% accuracy. Contempt is confused with anger and happiness, while fear is also confused with happiness. Sadness is also confused with anger. The limited number of samples available in the dataset is the reason for this confusion; the numbers of samples labeled as contempt, fear, and sadness are 18, 25, and 28, respectively. It should be noted that, due to using 7-fold cross-validation, the average number of samples belonging to an emotion in one fold cannot be more than five each for contempt, fear, and sadness.

4.4. Conclusion

In this chapter, we have proposed a new approach for facial-expression recognition, by fusing the information about the eye window and the mouth window of a face. The features used are LPQ and PHOG, which have been demonstrated to achieve good performances in facial-expression recognition. The respective features from the two windows are fused by projecting them into a coherent subspace, which is used in the

second level of classification, where the features from the eye and mouth windows have their correlation maximized. In our experiments, seven emotions are considered for testing the performance of our proposed method. Based on the training feature vectors, the SVM is employed to learn a binary classifier for each of the emotions. Experiment results have shown that our method, with the LPQ feature, outperforms the PHOG feature; and that our method can achieve greater accuracy than other, state-of-the-art facial-expression recognition methods.

Chapter 5. Facial expression recognition with emotion-based feature fusion

5.1. Introduction

Facial expression recognition (FER) is one of the most interesting topics in the field of human-computer interaction, and has become a popular research topic during the last few decades. As explained comprehensively in Chapter 2 and Chapter 3, before training classifiers for recognizing facial expressions, feature extraction is performed from face images in order to extract the distinctive features which can distinguish the different expressions. Also, we observed that different descriptors can achieve different over-all recognition rates. Furthermore, it can be seen that, from confusion matrices, different descriptors can achieve different recognition rates for a specific emotion. However, in the past, a single local descriptor was usually studied to achieve the best overall performance for all emotions. In this chapter, we propose to identify the best two features for each expression, which are then fused to form a coherent feature for representing a particular expression.

Manifold learning aims to embed high-dimensional data in a lower dimensional space while preserving the intrinsic characteristics. In [203], Shan et al. compared the performances of different manifold learning techniques on facial expression recognition, and showed that Supervised Locality Preserving Projections (SLPP) [202] achieves the best performance. More importantly, SLPP also considers the class information in the construction of the manifolds.

According to [101], emotions can be classified into four basic classes: 1) Anger-Disgust (AN-DI), 2) Fear-Surprise (Fe-SU), 3) Sadness (SA), and 4) Happiness (HA). In a video sequence, the set of specific facial movements of a particular emotion does

not occur at once but sequentially over time. In the early stages of anger or disgust, accurate discrimination between these two expressions is not obvious, similar to that between fear and surprise. Based on this, the number of expression classes is set at four, and the performances of the respective feature descriptors are measured for each of the expression classes. Then, the best two descriptors for each expression are identified and fused using Discriminant-Analysis of Canonical Correlations (DCC) [119] to form a coherent feature set. Our aim is to find the best discriminant features by combining the different descriptors for recognizing each facial expression. To the best of our knowledge, we are the first to use different coherent descriptors for the recognition of different expressions. Based on the coherent features, a classifier is learned for each expression. In other words, four classifiers are learned for the four expressions, i.e. anger-disgust, fear-surprise, happiness, and sadness.

The rest of this paper is organized as follows: The details of our proposed approach are presented in Section 5.2. In Section 5.3, experimental setup is described, and the experimental results are shown. Section 5.4 concludes the paper.

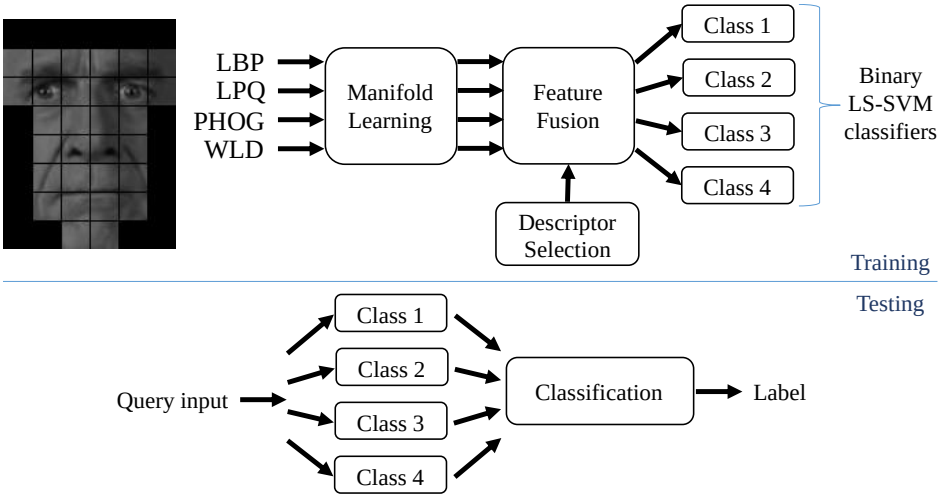

Fig. 5-1. The emotion-based feature fusion scheme for facial expression recognition.

5.2. Details of Our Approach

Before extracting features, the faces are scaled and aligned based on the position of the eyes such that the distance between the two eyes is 64 pixels and the image size is 126×100 pixels. In order to obtain more effective facial features, each image is divided into 8×6 regions, and 30 of the regions are used for feature extraction, as illustrated in Fig. 5-1. We can see that the selected regions contain the salient facial features, so they can represent facial expressions more effectively. After extracting the features, i.e. LBP, LPQ, PHOG, and WLD, supervised LPP is applied for manifold learning.

In the rest of this section, first, the four descriptors, SLPP, and DCC, are explained in detail. Then, the process of evaluating the performance of each descriptor for each expression class is described. Finally, the proposed adaptive descriptor selection algorithm is presented.

5.2.1. Local descriptors

In this paper, four different local descriptors are considered, because: 1) they have been used widely for facial expression recognition [49, 145, 204, 248], and 2) they represent facial expressions in terms of different aspects such as intensity, phase, and shape. These four descriptors are Local Binary Pattern (LBP) [179], Local Phase Quantization (LPQ) [180], Weber Local Descriptor (WLD) [33], and Pyramid of Histogram or Oriented Gradients (PHOG) [18]. For a detailed explanation of the descriptors used in our experiments, please refer to Section 3.3.

5.2.2. Supervised Locality Preserving Projection (SLPP)

Locality Preserving Projection (LPP) [176], which is a linear approximation of the nonlinear Laplacian Eigenmap [16], employs the following minimization problem:

$$\min_{w} \sum_{i,j} \left(\boldsymbol{w}^{\mathrm{T}} x_i - \boldsymbol{w}^{\mathrm{T}} x_j \right)^2 s_{ij}, \tag{5.1}$$

where $\mathbf{S} = [s_{ij}]$ is the similarity matrix that preserves the local neighbourhood information. An edge is added between nodes *i* and *j* if and are among the k nearest neighbours of each other. Heat kernel sets the edge weight s_{ij} as above if there is an edge between nodes *i* and *j*:

$$s_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}, \tag{5.2}$$

where *t* is the parameter for the method. An extension of LPP, namely supervised LPP (SLPP) [202], uses the class information when constructing the similarity matrix. In other words, an edge is added between nodes *i* and *j* if and only if x_i and x_j belong the same class and are among the *k*-nearest neighbors of each other. More about the graph-based manifold learning methods will be explained in Chapter 6.

5.2.3. Discriminant-Analysis of Canonical Correlations (DCC)

Discriminant-Analysis of Canonical Correlations (DCC) [119] was proposed as a discriminative learning method by Kim et al., inspired by Linear Discriminant Analysis (LDA) [15] which has been used commonly for dimension reduction aiming to preserve the class discriminatory information. Similar to LDA, DCC seeks to find a transformation matrix $\mathbf{W}$ for two feature sets $\mathbf{X}$ and $\mathbf{Y}$ such that $\mathbf{X}_w = \mathbf{W}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{Y}_w = \mathbf{W}^{\mathrm{T}}\mathbf{Y}$, where the matrix $\mathbf{W}$ maximizes the canonical correlations of the within-class sets, while minimizing the canonical correlations of the between-class sets.

In this paper, DCC is applied to two different feature sets extracted using two different descriptors in order to fuse them in a manner that the transformed feature set will have the most discriminant, coherent features to represent each emotion class.

5.2.4. Evaluating the descriptors

In this paper, we evaluate the performance of each descriptor using the one-versus-all classification scheme. The features of those face images of a particular emotion are

labeled as positive, while those of other emotions as negative. Then, a binary classifier is trained using Support Vector Machine (SVM) for each class of emotion, so there are a total of 4 classifiers. The recognition rate for each of the descriptors is measured. In addition, the two best descriptors for each emotion class are paired and then fused using DCC to form a single coherent descriptor. The performances of these coherent features are also evaluated using the one-versus-all scheme.

5.2.5. The proposed automatic descriptor selection algorithm

In the evaluation of the respective descriptors and the coherent descriptors, we found that fusing the two descriptors which achieve the highest recognition rates for a particular emotion can achieve higher accuracy than the individual descriptors. However, the best descriptors for each emotion may be different, as well as for different databases. Thus, fusing fixed descriptors to form a coherent descriptor is not the optimum way to achieve the best results. To achieve robust facial expression recognition, an adaptive descriptor selection step is included in our algorithm. The descriptor selection algorithm analyzes the performances of each pair of descriptors for each expression class on the given training set and determines the best two descriptors for each expression class regarding the training set; a total of 4 pairs of descriptors are selected. As observed before, the best two descriptors may be different for different expression classes. Therefore, a pair of best descriptors is determined for each expression class. In the descriptor selection step, *N*-fold cross validation, where $N = 3$ in our experiments, has been conducted on the training set. After identifying the best descriptors, a binary classifier is trained for each class using the most salient features, which are created by fusing the two best features by using DCC. For a query input, four different feature vectors are created and tested on the four different classifiers. The output of each of the classifiers is viewed as the probability of the

Fig. 5-2. Sample images for (a) the JAFFE, and (b) the BAUM-2 databases.

query belonging to the corresponding class. The query is assigned to the class whose corresponding output has the highest value.

5.3. Experimental Protocol and Results

5.3.1. Experimental protocol

Experiments were conducted on three databases: BAUM-2, JAFFE, and a combination of two databases. JAFFE [155] consists of images from 10 Japanese females that express 6 basic emotions and the neutral. Unlike JAFFE which is a database recorded in a controlled environment, the BAUM-2 [67] database consists of expression videos, extracted from movies. In our experiments, an image dataset, namely BAUM-2i, consisting of images with peak expressions from the videos from BAUM-2 is considered. There are 183 face images from 10 subjects in the JAFFE database that express 6 basic emotions, while there are 829 face images from 250 subjects in the BAUM-2i static expression dataset. Since the BAUM-2 database was created by extracting from movies, the images are in the close-to-real-life conditions (i.e. with pose, age, and illumination variations, etc.) and are more challenging than those in an acted database, as seen in Fig. 5-2.

It has been shown that SVM can achieve satisfactory results even for high-dimensional feature vectors. Furthermore, the more recent Least Square SVM (LS-SVM) [215] has been proposed, which is very efficient on large datasets since it uses

Table 5-2. Experiment results for JAFFE database.

JAFFE	LBP	LPQ	WLD	PHOG
AN-DI	**90.49% ± 1.84%**	85.79% ± 2.97%	89.51% ± 0.90%	**95.85% ± 0.49%**
FE-SU	93.11% ± 1.20%	86.34% ± 1.02%	**96.07% ± 1.05%**	**95.85% ± 1.37%**
HA	**96.28% ± 1.30%**	92.35% ± 1.02%	96.07% ± 1.52%	**97.27% ± 0.39%**
SA	**91.04% ± 1.48%**	88.09% ± 0.81%	**90.82% ± 1.12%**	89.29% ± 0.73%
ALL	**89.18% ± 0.46%**	83.28% ± 0.62%	89.18% ± 0.71%	**90.60% ± 0.71%**

linear programming, rather than convex programming in SVM. LS-SVM has been applied to different recognition problems like face [239] and facial expression [144, 268]. Therefore, our proposed method uses LS-SVM [1] with the Gaussian kernel.

5.3.2. Experiment results for the evaluation of the descriptors

To evaluate the performances of the selected descriptors, 5-fold cross validation was used. In this experiment, it is aimed to present that the performance of each descriptor is different for the different expression classes. Table 5-1 shows the performances of the different descriptors based on the JAFFE dataset. PHOG can achieve the highest accuracy for the expression classes Anger-Disgust and Happiness, while WLD performs better for the class Fear-Surprise. LBP descriptor outperforms other

Table 5-1. Experiment results for BAUM-2 dataset.

BAUM-2	LBP	LPQ	WLD	PHOG
AN-DI	73.92% ± 0.93%	**76.96% ± 0.28%**	**74.98% ± 0.37%**	70.16% ± 0.35%
FE-SU	81.91% ± 0.51%	**83.28% ± 0.56%**	**82.36% ± 0.65%**	81.54% ± 0.12%
HA	**88.37% ± 0.54%**	**89.82% ± 0.46%**	88.25% ± 0.25%	87.48% ± 0.36%
SA	**85.48% ± 0.31%**	**85.62% ± 0.45%**	84.54% ± 0.62%	84.70% ± 0.42%
ALL	62.85% ± 0.62%	**66.71% ± 0.56%**	**63.35% ± 0.48%**	59.86% ± 0.39%

descriptors for the class Sadness. The overall performances of each of the descriptors for all the expression classes are also evaluated. As observed, the overall performances of the classifiers are less than the performances of any other binary classifiers. The reason behind it is that the overall performance considers all the four labels, while the binary classifiers consider the labels as positive and negative. From the results, we can see that PHOG and LBP are the two best descriptors for recognizing all the expressions. Similarly, Table 5-2 shows the corresponding performances based on the BAUM-2i dataset. LPQ outperforms all other descriptors for all the expression classes. LPQ and WLD achieve the best overall performances.

As observed, even for the same expression classes, different descriptors can achieve the best recognition rates with different datasets. The reason for this is due to the fact that the two datasets are different in terms of race, age, resolution, pose, etc. Thus, the two databases are also merged into a single one to explore the form a

Table 5-3. Experiment results for BAUM-2 + JAFFE database.

BAUM-2 + JAFFE	**LBP**	**LPQ**	**WLD**	**PHOG**
AN-DI	74.19% ± 0.28%	**77.35% ± 0.27%**	74.68% ± 0.77%	**74.92% ± 0.11%**
FE-SU	82.21% ± 0.25%	82.59% ± 0.08%	**83.24% ± 0.19%**	**83.26% ± 0.56%**
HA	**89.92% ± 0.28%**	**90.08% ± 0.20%**	88.99% ± 0.27%	88.62% ± 0.26%
SA	84.55% ± 0.28%	**86.50% ± 0.32%**	**84.84% ± 0.85%**	84.66% ± 0.18%
ALL	64.74% ± 0.35%	**68.85% ± 0.37%**	**66.48% ± 0.52%**	64.88% ± 0.46%

database with images having more variations. The two best descriptors are then identified for each expression class. Table 5-3 shows the performances of the descriptors with respect to each of the expression classes. It can be seen that the two best descriptors selected based on BAUM-2 + JAFFE are correlated with the two best descriptors of either dataset. For instance, LPQ and PHOG descriptors achieve the highest accuracies for the AN-DI expression class in BAUM-2 + JAFFE (first row of

the results in Table 5-3). We can also observe that LPQ and PHOG are the descriptors that can achieve the best performances for the AN-DI class on BAUM-2 and JAFFE, respectively.

The results, once again, show that the different expression classes of different datasets can be represented more effectively by a different set of descriptors. Thus, the descriptors to be used for classification should not be fixed for a specific expression class, and should be adaptive to the expressions and the image conditions.

5.3.3. Experiment results for the proposed adaptive descriptor selection algorithm

Based on the results in Table 5-1, Table 5-2, and Table 5-3, the descriptors to be used are adaptive to the expression classes. For the JAFFE database, the fused features for the AN-DI, FE-SU, HA, and SA are PHOG+LBP, WLD+PHOG, PHOG+LBP, and LBP+WLD, respectively. For the BAUM-2i database, the fused features for the AN-DI, FE-SU, HA, and SA are LPQ+WLD, LPQ+WLD, LPQ+LBP, and LPQ+LBP, respectively. For the combined database, i.e. BAUM-2i + JAFFE, the fused features for the AN-DI, FE-SU, HA, and SA are LPQ+PHOG, WLD+PHOG, LPQ+LBP, and LPQ+WLD, respectively. We compare our proposed adaptive algorithm with the non-adaptive algorithm, which uses the same fused features for all the expression classes. For the JAFFE and BAUM-2i databases, PHOG+LBP and LPQ+WLD, respectively,

Table 5-4. Comparison of the performances of best descriptors of each dataset with adaptive descriptor selection method.

	JAFFE	**BAUM-2i**	**BAUM-2i + JAFFE**
LBP-PHOG	91.58% ± 0.30%	67.00% ± 0.50%	69.23% ± 0.15%
LPQ-WLD	87.32% ± 0.46%	68.47% ± 0.41%	69.96% ± 0.60%
Adaptive Descriptor Selection	**92.13% ± 0.91%**	**68.71% ± 0.53%**	**70.99% ± 1.13%**

achieve the best overall performance. These two fused features are used non-adaptively for the recognition of all the expression classes. In the experiments, 5-fold cross-validation has been conducted.

As shown in Table 5-4, using fused features can achieve higher recognition rates than the individual descriptors, and the adaptive algorithm outperforms the non-adaptive one. Also, as observed, the adaptive descriptor selection algorithm increases the accuracy up to 2% for the JAFFE, BAUM-2i and BAUM-2 + JAFFE datasets since the most salient features are used in the recognition of each expression class. The recognition rate for the BAUM-2i dataset is lower than that for JAFFE since BAUM-2i was created with expression images extracted from movies. This makes the dataset more challenging because of the pose, illumination and resolution variations.

5.4. Conclusion

In this chapter, we aim to show the differences in the performances regarding four commonly used descriptors: LBP, LPQ, PHOG and WLD. SLPP is applied as the manifold learning method, which preserves the locality information with the help of class information. Then, DCC is adopted to fuse the best two feature sets by projecting them into a coherent subspace. We have proposed a classification method, which utilizes the adaptive descriptor selection algorithm to further increase the performance of a facial expression recognition system. In our experiments, four expression classes are considered for evaluating the performance of the proposed classification method. The LS-SVM is employed based on the features projected to a coherent subspace to learn a binary classifier for each of the expression classes. Experiment results have shown that the proposed classification method can achieve higher recognition rate than any of the individual descriptors.

Chapter 6. Soft Locality Preserving Maps (SLPM) for facial expression recognition

6.1. Introduction

In this chapter, we propose a new graph-based subspace-learning method to solve the various problems of the existing methods, described in Section 2.2.3, by combining their best components to form a better method. The proposed method, named "Soft Locality Preserving Map (SLPM)" can be outlined as follows:

1. SLPM constructs a within-class graph matrix and a between-class graph matrix using the *k*-nearest neighborhood and the class information to discover the local geometry of the data.
2. To overcome the SSS problem and to decrease the computational cost of computing the inverse of a matrix, SLPM defines its objective function as the difference between the between-class and the within-class Laplacian matrices.
3. Inspired by the idea of SDM on the importance of the intra-class spread, a parameter β is added to control the penalty on the within-class Laplacian matrix so as to avoid the overfitting problem and to increase the generalizability of the underlying manifold.

Although subspace-learning methods have demonstrated promising performances by increasing the discriminative power of training data after transformation, they might fail to exhibit a similar performance on testing data. To improve the generalizability of the manifolds generated by the subspace-analysis methods, more training samples, which are located near the boundaries of the respective classes, are desirable. In this chapter, we apply our proposed SLPM method to facial expression

recognition, and propose an efficient way to enhance the generalizability of the manifolds of the different expression classes by feature generation.

An expression video sequence, which ranges from a neutral-expression face to the highest intensity of an expression, allows us to select appropriate samples for learning a better and more representative manifold for the expression class. For the optimal manifold of an expression class, its center should represent those samples that best represent the facial expression concerned, i.e. those expression face images with the highest intensities. When moving away from the manifold center, the corresponding expression intensity should be reducing. Those samples near the boundary of a manifold are important for describing the expression, which also defines the shape of the manifold. To describe a manifold boundary, images with low-intensity expressions should be considered. Since the feature vectors used to represent facial expressions usually have high dimensionality, many training samples near the manifold boundary are required, so as to represent it completely. However, we usually have a limited number of weak-intensity expression images, so feature generation is necessary to learn more complete manifolds.

In other applications, additional samples have also been generated for manifold learning. In [115], faces are morphed between two people with different percentages so as to generate face images near the manifold boundaries. By generating more face images and extracting their feature vectors, the manifold for each face subject can be learned more accurately. Therefore, the decision region for each subject can be determined for watch-list surveillance. In our algorithm, rather than morphing faces and extracting features from the synthesized face, we generate features for low-intensity expressions directly in the feature domain. Generating features in this way should be more accurate than extracting features from distorted faces generated by

morphing. Several fields of research, such as text categorization [76], handwritten digit recognition [77], facial expression recognition [2, 252], etc., have also employed feature generation to achieve better learning. Unlike these methods, which generate features in the image domain, the proposed method generates features in the feature domain.

In Section 6.2, we further explain the graph-embedding techniques, and give a detailed comparison of those existing subspace-learning approaches similar to our proposed method. In Section 6.3, the proposed method, SLPM, is formulated and its relation to SDM is further explored. In Section 6.4, we listed the local descriptors used in our experiments and the feature-generation algorithm, and describe how to enhance the manifold learning with low-intensity images. In Section 6.5, we present the databases used in our experiments, and the preprocessing of the face images. Then, experiment results are represented, with a discussion. We conclude this chapter in Section 6.6.

6.2. Detailed Review of Subspace Learning

In this section, a review of the graph-embedding techniques is presented in detail, with the different variants. Then, graph-based subspace-learning methods are described in two parts: 1) how the adjacency matrices are constructed, and 2) how their objective functions are defined.

6.2.1. Graph embedding

Given m data points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m\} \in \mathbb{R}^D$, the graph-based subspace-learning methods aim to find a transformation matrix $\mathbf{A}$ that maps the training data points to a new set of points $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_m\} \in \mathbb{R}^d$ $(d \ll D)$, where $\boldsymbol{y}_i = \mathbf{A}^T \boldsymbol{x}_i$ and $\mathbf{A}$ is the projection matrix. After the transformation, the data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which are close

to each other, will have their projections in the manifold space $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ close to each other. This goal can be achieved by minimizing the following objective function:

$$\sum_{ij}(\boldsymbol{y}_i - \boldsymbol{y}_j)^2 w_{ij}, \tag{6.1}$$

where w_{ij} represents the similarity between the training data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. If w_{ij} is non-zero, $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ must be close to each other, in order to minimize (1). Taking the data points in the feature space as nodes of a graph, an edge between nodes i and j has a weight of w_{ij}, which is not zero, if they are close to each other. In the literature, we have found three different ways to determine the local geometry of a data point:

1. **ε-neighbourhood:** This uses the distance to determine the closeness. Given ε ($\varepsilon \in \mathbb{R}$), ε-neighbourhood chooses the data points that fall within the circle around x_i with a radius ε. Those data points fall within the ε-neighbourhood of $\boldsymbol{x}_i$ can be defined as

$$O(\boldsymbol{x}_i, \varepsilon) = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}_i\|^2 < \varepsilon\}. \tag{6.2}$$

2. ***k*-nearest neighbourhood:** Another way of determining the local structure is to use the nearest neighbourhood information. Presuming that the closest k points of $\boldsymbol{x}_i$ would still be the closest data points of $\boldsymbol{y}_i$ in the projected manifold space, we can define a function $N(\boldsymbol{x}_i, k)$, which outputs the set of k-nearest neighbours of $\boldsymbol{x}_i$. Two types of neighbourhood, with label information incorporated, are considered: $N(\boldsymbol{x}_i, k^+)$ and $N(\boldsymbol{x}_i, k^-)$, which represent the sets of k-nearest neighbours of $\boldsymbol{x}_i$ of the same label and of different labels, respectively.
3. **The class information:** The class or label information is often used in supervised subspace methods. In a desired manifold subspace, the data points belonging to the class of $\boldsymbol{x}_i$ are to be projected such that they are close to each

other, so as to increase the intra-class compactness. The data points belonging to other classes are projected, such that they will become farther apart and have larger inter-class separability. The class label information is often combined with either the ε-neighbourhood or the *k*-nearest neighbourhood.

The similarity graph is constructed by setting up edges between the nodes. There are different ways of determining the weights of the edges, considering the fact that the distance between two neighbouring points can also provide useful information about the manifold. Given a sparse symmetric similarity matrix $\mathbf{W}$, two variations have been proposed in the literature.

1. Binary weights: $w_{ij} = 1$ if, and only if, the nodes i and j are connected by an edge, otherwise $w_{ij} = 0$.
2. Heat kernel ($t \in \mathbb{R}$): If the nodes i and j are connected by an edge, the weight of the edge is defined as

$$w_{ij} = \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \Big/ t\right). \tag{6.3}$$

After constructing the similarity matrix with the weights, the minimization problem defined in (1) can be solved by using the spectral graph theory. Defining the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$**,** where $\mathbf{D}$ is the diagonal matrix whose entries are the column sum of $\mathbf{W}$, i.e. $d_{ii} = \sum_j w_{ij}$, the objective function is reduced to

$$\begin{aligned} \min \sum_{ij} \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij} &= \min \sum_{ij} \left(\mathbf{A}^{\mathrm{T}}\boldsymbol{x}_i - \mathbf{A}^{\mathrm{T}}\boldsymbol{x}_j\right)^2 w_{ij} \\ &= \min \mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{A}, \end{aligned} \tag{6.4}$$

where $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$. To avoid the trivial solution of the objective function, the constraint $\mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{A} = 1$ is often added. After specifying the objective function, the optimal projection matrix $\mathbf{A}$ can be obtained by choosing the eigenvectors

corresponding to the d ($d \ll D$) smallest non-zero eigenvalues computed by solving the standard eigenvalue decomposition or generalized eigenvalue problem, depending on the objective function being considered.

6.2.1.1. Constructing the within-class and the between-class graph matrices

As mentioned in the last section, one of the most popular graph-based subspace-learning methods is Locality Preserving Projections (LPP) [176], which uses an intrinsic graph to represent the locality information of the dataset, i.e. the neighbourhood information. The idea behind LPP is that if the data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close to each other in the feature space, then they should also be close to each other in the manifold subspace. The similarity matrix w_{ij} for LPP can be defined as follows:

$$w_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N(\boldsymbol{x}_j, k) \text{ or } \boldsymbol{x}_j \in N(\boldsymbol{x}_i, k), \\ 0, & \text{otherwise,} \end{cases} \tag{6.5}$$

where $N(\boldsymbol{x}_i, k)$ represents the set of k-nearest neighbors of $\boldsymbol{x}_i$. One shortfall of the above formulation for w_{ij} is that it is an unsupervised method, i.e. not using any class-label information. Thinking that the label information can help to find a better separation between different class manifolds, Supervised Locality Preserving Projections (SLPP) was introduced in [202]. Denote $l(\boldsymbol{x}_i)$ as the corresponding class label of the data point $\boldsymbol{x}_i$. SLPP uses either one of the following formulations:

$$w_{ij} = \begin{cases} 1, & \text{if } l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j), \\ 0, & \text{otherwise,} \end{cases} \tag{6.6}$$

$$w_{ij} = \begin{cases} 1, & \text{if } (\boldsymbol{x}_i \in N(\boldsymbol{x}_j, k) \text{ or } \boldsymbol{x}_j \in N(\boldsymbol{x}_i, k)) \text{ and } l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j), \\ 0, & \text{otherwise.} \end{cases} \tag{6.7}$$

Note that (6-7) does not include the neighbourhood information to the adjacency graph, and the similarity matrices defined above can be constructed using the heat kernel. Orthogonal Locality Preserving Projection (OLPP) [24] whose eigenvectors are orthogonal to each other is an extension of LPP. Please note that, in our

experiments we applied Supervised Orthogonal Locality Preserving Projections (SOLPP), which is OLPP with its adjacency matrix including class information.

Yan et al. [242] proposed a general framework for dimensionality reduction, named Marginal Fisher Analysis (MFA). MFA, which is based on graph embedding as LPP, uses two graphs, the intrinsic and penalty graphs, to characterize the intra-class compactness and the interclass separability, respectively. In MFA, the intrinsic graph w_{ij}^{w}, i.e. the within-class graph, is constructed using the neighbourhood and class information as follows:

$$w_{ij}^{w} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N\left(\boldsymbol{x}_j, k_1^{+}\right) \text{ or } \boldsymbol{x}_j \in N(\boldsymbol{x}_i, k_1^{+}), \\ 0, & \text{otherwise.} \end{cases} \tag{6.8}$$

Similarly, the penalty graph w_{ij}^{b}, i.e. the between-class graph, is constructed as follows:

$$w_{ij}^{b} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N\left(\boldsymbol{x}_j, k_2^{-}\right) \text{ or } \boldsymbol{x}_j \in N(\boldsymbol{x}_i, k_2^{-}), \\ 0, & \text{otherwise.} \end{cases} \tag{6.9}$$

Locality Sensitive Discriminant Analysis (LSDA) [25] and Improved Locality Sensitive Discriminant Analysis (ILSDA) [127] are subspace-learning methods proposed in 2007 and 2015, respectively. They construct the similarity matrices in the same way, but LSDA uses binary weights, while ILSDA sets the weight of the edges using the heat kernel. The similarity matrices of LSDA are defined as follows:

$$w_{ij}^{w} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N\left(\boldsymbol{x}_j, k\right) \text{ and } l(\boldsymbol{x}_i) = l\left(\boldsymbol{x}_j\right), \\ 0, & \text{otherwise.} \end{cases} \tag{6.10}$$

$$w_{ij}^{b} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N\left(\boldsymbol{x}_j, k\right) \text{ and } l(\boldsymbol{x}_i) \neq l\left(\boldsymbol{x}_j\right), \\ 0, & \text{otherwise.} \end{cases} \tag{6.11}$$

It can be observed that the intrinsic and the penalty graphs of MFA, LSDA, and ILSDA are similar to each other. In MFA, the numbers of neighboring points for both the similarity matrices are known, i.e. k_1 and k_2. In LSDA and ILSDA, the k neighbors of $\boldsymbol{x}_i$ are selected, which are then divided for constructing the within-class (k^{+} samples

the same class as $\boldsymbol{x}_i$) and the between-class matrices (k^- samples of other classes), i.e. $k = k^+ + k^-$. Let k_1 and k_2 be the numbers of samples belonging to the same class and different classes, respectively, for MFA. The following relation is not always true:

$$N(\boldsymbol{x}_i, k_1) \cup N(\boldsymbol{x}_i, k_2) = N(\boldsymbol{x}_i, k), \tag{6.12}$$

because it is not necessarily true that $k^+ = k_1$ and $k^- = k_2$. Therefore, the neighboring points of x_i in LSDA and ILSDA are not the same as MFA, even if $= k_t = k_1 + k_2$. However, the adjacency matrices constructed in the manifold learning methods are similar to each other. The main difference between the existing methods in the literature is in their definitions of the objective functions. We will elaborate on the differences in the objective functions in the next section.

Locality-Preserved Maximum Information Projection (LPMIP) [234], proposed in 2008, uses the ε-neighbourhood condition, i.e. $O(\boldsymbol{x}_i, \varepsilon)$. Although it was originally applied as an unsupervised learning method, the class labels were used to construct the locality and non-locality information for facial expression recognition. In 2008, Li et al. [130] proposed Constrained Maximum Variance Mapping (CMVM), which aims to keep the local structure of the data, while separating the different manifolds, i.e. different classes, farther apart. The local-structure graphs, i.e. the between-class graph and the dissimilarities graph, are defined as follows:

$$w_{ij}^w = \begin{cases} 1 \text{ or } \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \Big/ t\right), & \text{if } \boldsymbol{x}_i \in O(\boldsymbol{x}_j, \varepsilon), \\ 0, & \text{otherwise,} \end{cases} \tag{6.13}$$

$$w_{ij}^b = \begin{cases} 1, & \text{if } l(\boldsymbol{x}_i) \neq l(\boldsymbol{x}_j), \\ 0, & \text{otherwise.} \end{cases} \tag{6.14}$$

As (6.13) and (6.14) show, the within-class matrix of CMVM only preserves the local structure of the whole data, while the between-class matrix only uses the class label to increase the separability of different class manifolds. In 2015, an extension of CMVM,

namely CMVM+ [269], was proposed to overcome the obstacles of CMVM. CMVM+ adds the class information and neighbourhood information to the similarity matrices. The updated version of the graphs can be written as follows:

$$w_{ij}^{w} = \begin{cases} 1, & \boldsymbol{x}_i \in N(\boldsymbol{x}_j, k) \text{ and } l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j), \\ 0, & \text{otherwise}, \end{cases} \tag{6.15}$$

$$w_{ij}^{b} = \begin{cases} 1, & \text{if } l(\boldsymbol{x}_j) \in C_{inc}(\boldsymbol{x}_i), \\ 0, & \text{otherwise}, \end{cases} \tag{6.16}$$

where $C_{inc}(\boldsymbol{x}_i)$ is a set of neighboring points belonging to different classes, i.e. $l(\boldsymbol{x}_i) \neq l(\boldsymbol{x}_j)$. More details of the function $C_{inc}(\boldsymbol{x}_i)$ can be found in [269].

In 2011, Multi-Manifolds Discriminant Analysis (MMDA) [249] was proposed for image feature extraction, and applied to face recognition. The idea behind MMDA is

Table 6-1. A comparison of the within-class graph and the between-class graph for different subspace-learning methods. (bn: binary weights, hk: heat kernel, *k*-NN: *k*-nearest neighborhood)

Subspace Learning Methods	The within-class graph			The between-class graph		
	Neighbourhood	Class Info	Weight	Neighbourhood	Class Info	Weight
LPP [176]/ OLPP [24]	optional	No	optional	n/a	n/a	n/a
SLPP [202]/ SOLPP	optional	Yes	optional	n/a	n/a	n/a
LSDA [25]	*k*-NN	Yes	bn	*k*-NN	Yes	bn
MFA [242]	*k*-NN	Yes	bn	*k*-NN	Yes	bn
CMVM [130]	ϵ-ball	No	bn/hk	n/a	Yes	bn
LPMIP [234]	ϵ-ball	No	hk	ϵ-ball	No	hk
MMDA [249]	n/a	Yes	hk	class centers	Yes	hk
CMVM+ [269]	*k*-NN	Yes	bn	*k*-NN	Yes	bn
ILSDA [127]	*k*-NN	Yes	hk	*k*-NN	Yes	hk
SLPM (proposed)	*k*-NN	Yes	bn/hk	*k*-NN	Yes	hk

to keep the points from the same class as close as possible in the manifold space, with the within-class matrix defined as follows:

$$w_{ij}^{w} = \begin{cases} \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \Big/ t\right), & \text{if } l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j), \\ 0, & \text{otherwise.} \end{cases} \tag{6.17}$$

MMDA also constructs a between-class matrix in order to separate the different classes from each other. The difference between the between-class matrix of MMDA and the other subspace methods is that its graph matrix is constructed by not taking all the data points as nodes, but rather calculating the weighted centres of different classes by averaging all the data points belonging to the classes under consideration. Let $\mathbf{M} = [\tilde{\boldsymbol{m}}_1, \tilde{\boldsymbol{m}}_2, \ldots, \tilde{\boldsymbol{m}}_c]$ be the class-weighted centres, where c is the number of classes. Then, the between-class matrix of MMDA can be written as:

$$w_{ij}^{b} = \exp\left(-\left\|\tilde{\boldsymbol{m}}_i - \tilde{\boldsymbol{m}}_j\right\|^2 \Big/ t\right). \tag{6.18}$$

In Table 6-1, a summary is given of the within-class graph and between-class graph for the subspace-learning methods, reviewed in this paper.

6.2.1.2. Defining the objective functions

Table 6-2 summarizes the objective functions of the approaches reviewed in the previous section, as well as the constraints used. We can see that SLPP has only one Laplacian matrix defined in its objective function, because it constructs one similarity matrix only, while all the other methods have two matrices: one is based on the intrinsic graph, and the other on the penalty graph.

In general, there are two ways of defining the objective functions with the intrinsic and the penalty matrices. The first one utilizes the Fisher criterion to maximize the ratio between the scattering of the between-class and that of the within-class Laplacian matrices. MFA, MMDA, and CMVM+ employ the Fisher criterion. Although the

Table 6-2. A comparison of the objective functions used by different subspace methods.

	Objective functions	Constraints (s.t.)
LPP [176]/ SLPP [202]	$\max_{\mathbf{A}} \mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}$	$\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A} = \mathbf{I}$
LSDA [25]	$\max_{\mathrm{A}} \mathbf{A}^T\mathbf{X}(a\mathbf{L}_b + (1-\alpha)\mathbf{W}_w)\mathbf{X}^T\mathbf{A}$	$\mathbf{A}^T\mathbf{X}\mathbf{D}_w\mathbf{X}^T\mathbf{A} = \mathbf{I}$
MFA [242]	$\min_{\mathbf{A}} \frac{\mathbf{A}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{A}}{\mathbf{A}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{A}}$	n/a
CMVM [130]	$\max \mathbf{A}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{A}$	$\mathbf{A}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{A} = \mathbf{X}\mathbf{L}_w\mathbf{X}^T$
LPMIP [234]	$\max_{\mathbf{A}} \mathbf{A}^T\mathbf{X}(a\mathbf{L}_b - (1-\alpha)\mathbf{L}_w)\mathbf{X}^T\mathbf{A}$	$\mathbf{A}^T\mathbf{A} - \mathbf{I} = 0$
MMDA [249]	$\max_{\mathbf{A}} \frac{\mathbf{A}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{A}}{\mathbf{A}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{A}}$	n/a
CMVM+ [269]	$\max_{\mathbf{A}} \frac{\mathbf{A}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{A}}{\mathbf{A}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{A}}$	n/a
ILSDA [127]	$\max \mathbf{A}^T(\mathbf{P} - \alpha\mathbf{S}_w)\mathbf{A}$ where $\mathbf{P} = \mathbf{X}(\mathbf{L}_b - \mathbf{L}_w)\mathbf{X}^T$	$\mathbf{A}^T\mathbf{A} = \mathbf{I}$
SDM [143]	$\max \mathbf{S}_b - \alpha\mathbf{S}_w$	n/a
SLPM	$\max \mathbf{A}^T\mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^T\mathbf{A}$	$\mathbf{A}^T\mathbf{A} - \mathbf{I} = 0$

application of the Fisher criterion shows its robustness, it involves taking the inverse of a high-dimensional matrix to solve a generalized eigenvalue problem. To solve this problem, LSDA, LPMIP, and our proposed method define the objective functions as the difference between the intrinsic and the penalty-graph matrices, while MMC and SDM use the difference between the inter-class and the intra-class scatter matrices.

As shown in Table 6-2, ILSDA adopts a similar objective function to LSDA, but with a difference that the within-class scatter matrix is included in the objective function. The within-class scatter matrix $\mathbf{S}_w$ — as used in LDA — indicates the compactness of the data point in each class. ILSDA uses the scatter matrix to project outliers closer to the class centers under consideration. The objective function of ILSDA is defined as follows:

$$\max \mathbf{A}^T(\mathbf{P} - \alpha\mathbf{S}_w)\mathbf{A}, \tag{6.19}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{L}_b - \mathbf{L}_w)\mathbf{X}^T$, as defined in the objective function of LSDA. CMVM, unlike other methods which aim to minimize the within-class spread, intends to maintain the within-class structure for each class by defining a constraint, i.e. $\mathbf{A}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{A} = \mathbf{X}\mathbf{L}_w\mathbf{X}^T$, while increasing the inter-class separability with the following objective function:

$$\max \mathbf{A}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{A}, \tag{6.20}$$

where $\mathbf{L}_w$ and $\mathbf{L}_b$ are the within-class and the between-class Laplacian matrices, respectively.

6.3. Soft Locality Preserving Map (SLPM)

In this section, we introduce the proposed method, Soft Locality Preserving Map (SLPM), with its formulation and connection to the previous works.

6.3.1. Formulation of the SLPM

Similar to other manifold-learning algorithms, two graph-matrices, the between-class matrix $\mathbf{W}_b$ and the within-class matrix $\mathbf{W}_w$, are constructed to characterize the discriminative information, based on the locality and class-label information. Given m data points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m\} \in \mathbb{R}^D$ and their corresponding class labels $\{l(\boldsymbol{x}_1), l(\boldsymbol{x}_2), \dots, l(\boldsymbol{x}_m)\}$, we denote $N_w(\boldsymbol{x}_i, k_w) = \{\boldsymbol{x}_i^{w_1}, \boldsymbol{x}_i^{w_2}, \dots, \boldsymbol{x}_i^{w_{k_w}}\}$ as the set of k_w nearest neighbours with the same class label as $\boldsymbol{x}_i$, i.e. $l(\boldsymbol{x}_i) = l(\boldsymbol{x}_i^{w_1}) = l(\boldsymbol{x}_i^{w_2}) = \cdots = l(\boldsymbol{x}_i^{w_{k_w}})$, and $N_b(\boldsymbol{x}_i, k_b) = \{\boldsymbol{x}_i^{b_1}, \boldsymbol{x}_i^{b_2}, \dots, \boldsymbol{x}_i^{b_{k_b}}\}$ as the set of its k_b nearest neighbours with different class labels from $\boldsymbol{x}_i$, i.e. $l(\boldsymbol{x}_i) \neq l(\boldsymbol{x}_i^{w_j})$, where $j = 1, 2, \dots, k_b$. Then, the inter-class weight matrix $\mathbf{W}_b$ and the intra-class weight matrix $\mathbf{W}_w$ can be defined as below:

$$w_{ij}^{b} = \begin{cases} \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \Big/ t\right), & \boldsymbol{x}_j \in N_b(\boldsymbol{x}_i, k_b), \\ 0, & \text{otherwise.} \end{cases} \tag{6.21}$$

$$w_{ij}^{w} = \begin{cases} \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \Big/ t\right), & \boldsymbol{x}_j \in N_w(\boldsymbol{x}_i, k_w), \\ 0, & \text{otherwise.} \end{cases} \tag{6.22}$$

SLPM is a supervised manifold-learning algorithm, which aims to maximize the between-class separability, while controlling the within-class spread with a control parameter β used in the objective function. Consider the problem of creating a subspace, such that data points from different classes, i.e. represented as edges in $\mathbf{W}_b$, stay as distant as possible, while data points from the same class, i.e. represented as edges in $\mathbf{W}_w$, stay close to each other. To achieve this, two objective functions are defined as follows:

$$\max \tfrac{1}{2} \textstyle\sum_{ij} \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij}^{b}, \tag{6.23}$$

$$\min \tfrac{1}{2} \textstyle\sum_{ij} \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij}^{w}. \tag{6.24}$$

Eq. (23) ensures that the samples from different classes will stay as far as possible from each other, while Eq. (24) is to make samples from the same class stay close to each other after the projection. However, as shown in [127] and [143], small variations in the manifold subspace can lead to overfitting in training. To overcome this problem, we add the parameter β to control the intra-class spread. Note that, the method SDM in [143] uses the within-class scatter matrix $\mathbf{S}_w$ – as defined for LDA – to control the intra-class spread. In our proposed method, we adopt the graph-embedding method, which uses the locality information about each class, in addition to the class information. Hence, the two objective functions Eq. (23) and Eq. (24) can be combined as follows:

$$\max \tfrac{1}{2} \left(\textstyle\sum_{ij} \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij}^{b} - \beta \textstyle\sum_{ij} \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij}^{w} \right)$$

$$= \max\left(J_b(\mathbf{A}) - \beta J_w(\mathbf{A})\right), \tag{6.25}$$

where $\mathbf{A}$ is a projection matrix, i.e. $\mathbf{Y} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m]$. Then, the between-class objective function $J_b(\mathbf{A})$ can be reduced to

$$\begin{aligned} J_b(\mathbf{A}) &= \tfrac{1}{2}\textstyle\sum_{ij}\left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij}^b \\ &= \mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{L}_b\mathbf{X}^{\mathrm{T}}\mathbf{A} \end{aligned} \tag{6.26}$$

where $\mathbf{L}_b = \mathbf{D}_b - \mathbf{W}_b$ is the Laplacian matrix of $\mathbf{W}_b$ and $d_{b_{ii}} = \sum_j w_{ij}^b$ is a diagonal matrix. Similarly, the within-class objective function $J_w(\mathbf{A})$ can be written as

$$\begin{aligned} J_w(\mathbf{A}) &= \tfrac{1}{2}\textstyle\sum_{ij}\left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 w_{ij}^w \\ &= \mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{L}_w\mathbf{X}^{\mathrm{T}}\mathbf{A} \end{aligned} \tag{6.27}$$

where $\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w$ and $d_{w_{ii}} = \sum_j w_{ij}^w$. If J_w and J_b are substituted to Eq. (25), the objective function becomes as follows:

$$\begin{aligned} \max J_T(\mathbf{A}) &= \max\left(J_b(\mathbf{A}) - \beta J_w(\mathbf{A})\right) \\ &= \max(\mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{L}_b\mathbf{X}^{\mathrm{T}}\mathbf{A} - \beta\mathbf{A}^{\mathrm{T}}\mathbf{X}\mathbf{L}_w\mathbf{X}^{\mathrm{T}}\mathbf{A}) \\ &= \max \mathbf{A}^{\mathrm{T}}\mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^{\mathrm{T}}\mathbf{A} \end{aligned} \tag{6.28}$$

which is subject to $\mathbf{A}^{\mathrm{T}}\mathbf{A} - 1 = 0$ so as to guarantee orthogonality. By using Lagrange multiplier, we obtain

$$\mathrm{L}(\mathbf{A}) = \mathbf{A}^{\mathrm{T}}\mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^{\mathrm{T}}\mathbf{A} - \lambda(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathbf{I}). \tag{6.29}$$

By computing the partial derivative of L(**A**), the optimal projection matrix **A** can be obtained, as follows:

$$\frac{\partial \mathrm{L}(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^T\mathbf{A} - \lambda\mathbf{A} = 0, \tag{6.30}$$

i.e. $\mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^T\mathbf{A} = \lambda\mathbf{A}$. The projection matrix **A** can be obtained by computing the eigenvectors of $\mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^T$. The columns of **A** are the *d* leading eigenvectors, where *d* is the dimension of the subspace. LDA, LPP, MFA, and other manifold-learning algorithms, whose objective functions have a similar structure, lead to a

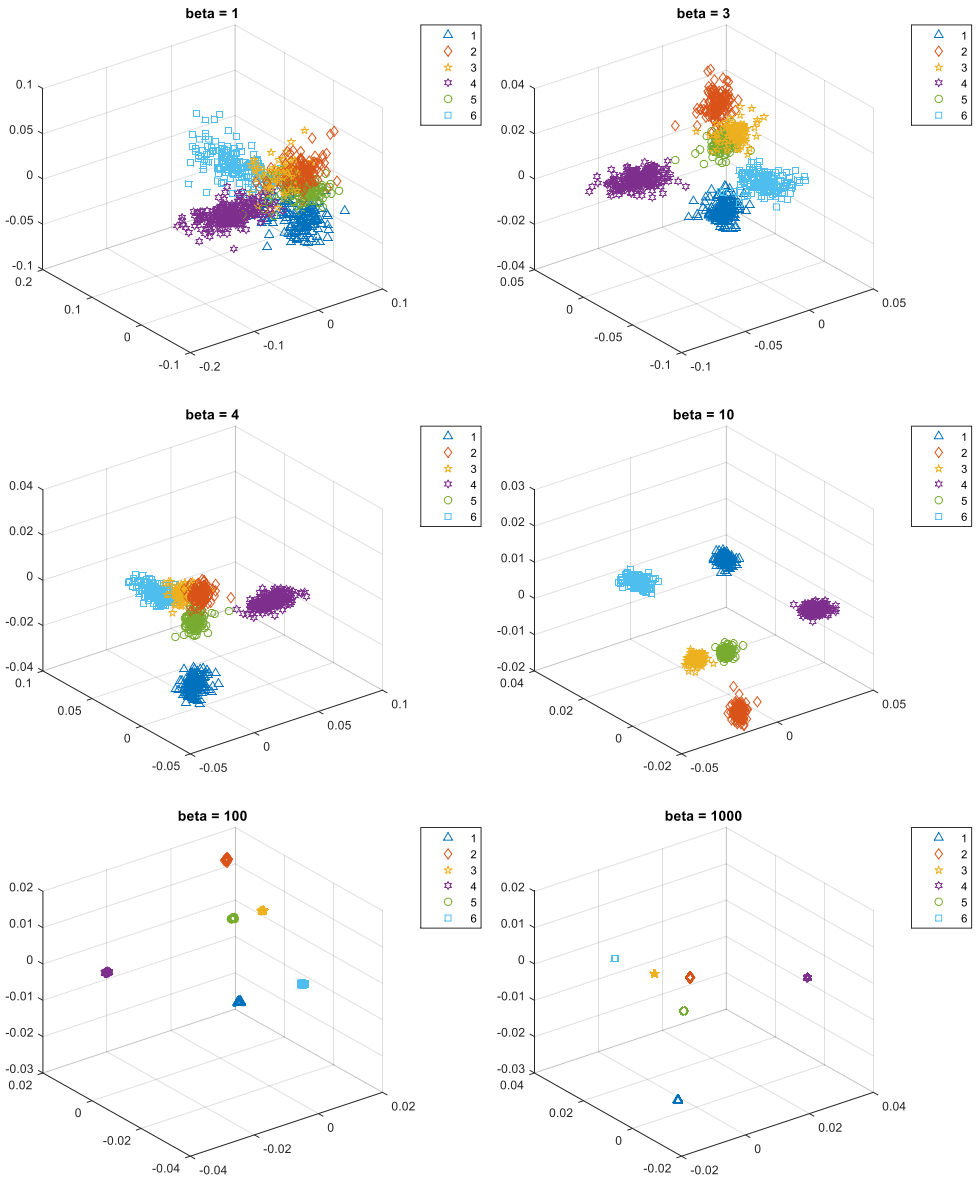

Fig. 6-1. The spread of the respective expression manifolds when the value of β increases from 1 to 1,000: (1) Anger, (2) Disgust, (3) Fear, (4) Happiness, (5) Sadness, and (6) Surprise.

generalized eigenvalue problem. Such methods suffer from the matrix-singularity problem, because the solution involves computing the inverse of a singular matrix. The proposed objective function is designed in such a way as to overcome this singularity problem. However, in our algorithm, PCA is still applied to data, so as to reduce its dimensionality and to reduce noise.

6.3.2. Intra-class spread

As we have mentioned before, the manifold spread of the different classes can affect the generalizability of the learned classifier. To control the spread of the classes, the parameter β is adjusted in our proposed method, like SDM. Fig. 6-1 shows the change in the spread of the classes when β increases. We can see that increasing β will also increase the separability of the data, e.g. the training data is located at almost the same position in the subspace when $\beta = 1{,}000$.

6.3.3. Relations to other subspace learning methods

As discussed in Section 6.2, there have been extensive studies on manifold-learning methods. They share the same core idea, i.e. using locality and/or label information to define an objective function, so that the data can be represented in a specific way after projection.

There are two main differences between SLPM and LSDA. First, LSDA defines their objective function as a subtraction of two objective functions like SLPM. However, LSDA imposes the constraint $\mathbf{A}^T\mathbf{X}\mathbf{D}_w\mathbf{X}^T\mathbf{A} = \mathbf{I}$, which results in a generalized eigenvalue problem. As we mentioned in Section 2, the generalized eigenvalue problem suffers from the computational cost of calculating an inverse matrix. SLPM only determines the orthogonal projections, with the constraint $\mathbf{A}^T\mathbf{A} - \mathbf{I} = 0$. Therefore, SLPM can still be computed by eigenvalue decomposition, without requiring computing any inverse matrix. Second, LSDA finds the neighboring points followed by determining whether the considered neighboring points are of the same class or of different classes. This may lead to an unbalanced and unwanted division of neighboring points, simply because of the fact that a sample point may be surrounded by more samples belonging to the same class than samples with different class labels. In order not to lose locality information in such a case, SLPM defines two parameters

k_1 and k_2, which are the numbers of neighboring points belonging to the same class and different classes, respectively. In other words, the numbers of neighboring points belonging to the same class and different classes can be controlled.

Both SDM and ILSDA also consider the intra-class spread when defining the objective function. SDM controls the level of spread by applying a parameter to the within-class scatter matrix $\mathbf{S}_w$. However, it only uses the label information about the training data – its scatter matrices do not consider the local structure of the data. Our proposed SLPM aims to include the locality information by employing graph embedding in our objective functions. Therefore, SLPM is a graph-based version of SDM. ILSDA uses both the label and neighborhood information represented in the adjacent matrices, and also aims to control the spread of the classes. However, ILSDA achieves this by adding the scatter matrix $\mathbf{S}_w$ to its objective function. In our algorithm, we propose controlling the spread with the within-class Laplacian matrix $\mathbf{L}_w$, without adding a separate element to the objective function.

6.4. Feature Descriptor and Generation

In this section, we will first present the descriptors used for representing facial images for expression recognition, then investigate the use of face images with low-intensity and high-intensity expressions for manifold learning, which represent the corresponding samples at the core and boundary of the manifold for an expression. After that, we will introduce our proposed feature-generation algorithm.

6.4.1. Descriptors

Recent research has shown that local features can achieve higher and more robust recognition performance than by using global features, such as Eigenfaces and Fisherfaces, and intensity values. Therefore, in order to show the robustness of our proposed method, four different commonly used local descriptors for facial expression

1. Extract features from face images: $\mathbf{X}_{desc}$
2. Learn the projection matrix $\mathbf{W}_{pca}$ via PCA
3. Construct the within-class graph matrix $\mathbf{W}_w$ and the between-class similarity matrices $\mathbf{W}_b$
4. Calculate the Laplacian matrices $\mathbf{L}_w$ and $\mathbf{L}_b$
5. Solve the eigenvalue decomposition of $\mathbf{X}(\mathbf{L}_b - \beta\mathbf{L}_w)\mathbf{X}^{\mathrm{T}}$
6. Choose the eigenvectors corresponding to the *d* largest eigenvalues, $\mathbf{W}_{mL}$
7. $\mathbf{Y}_{desc} = \mathbf{W}_{mL}^{\mathrm{T}}\mathbf{W}_{pca}^{\mathrm{T}}\mathbf{X}_{desc}$
8. Add features obtained with either low-intensity images ($\mathbf{Y}^l$) or feature generation ($\overline{\mathbf{Y}}^l$) to form the training data $\mathbf{T}^l$ or $\overline{\mathbf{T}}_l$, respectively
9. Learn the nearest neighbor classifier

Fig. 6-2. The overall flow of our proposed method.

recognition, Local Binary Pattern (LBP) [178, 204], Local Phase Quantization (LPQ) [180], Pyramid of Histogram of Oriented Gradients (PHOG) [18], and Weber Local Descriptor (WLD) [33], are considered in our experiments. These descriptors can represent face images, in terms of different aspects such as intensity, phase, shape, etc., so that they are complementary to each other (please refer to Section 3.3 for a detailed explanation of the descriptors.). As shown in Fig. 6-2, features are extracted using one of the above-mentioned local descriptors, followed by the subspace learning with SLPM and a feature-generation method.

6.4.2. Feature generation

Features in a projected subspace still have a high dimension. A large number of samples for each expression is necessary in order to represent its corresponding manifold accurately. By generating more features located near the manifold boundaries, more accurate decision boundaries can then be determined for accurate facial expression.

Video sequences with face images, changing from neutral expression to a particular expression, are used for learning. Let $f_{i,\theta}$ denote the frame index of the face

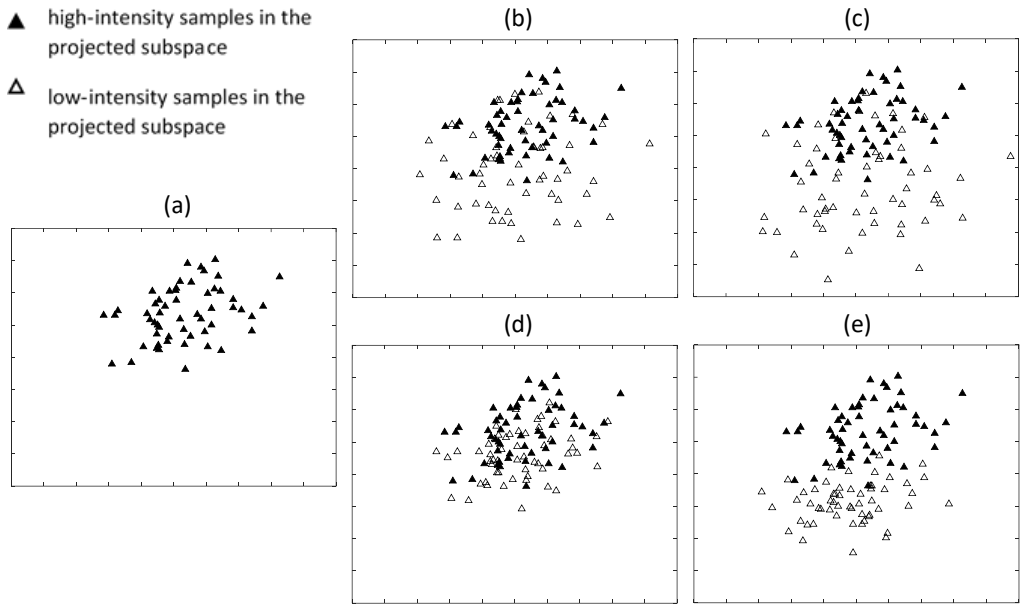

Fig. 6-3. The representation of the feature vectors (FV) of happiness (HA) on the CK+ database, after SLPM: (a) HA, i.e. high-intensity expression samples are applied to SLPM, (b) HA + low intensity FV with $\xi = 0.9$, (c) HA + low intensity FV with $\xi = 0.7$, (d) HA + generated FV with $\theta_{ne} = 0.9$, and (e) HA + generated FV with $\theta_{ne} = 0.7$.

image of expression intensity θ ($0 \leq \theta \leq 1$, $0 =$ neutral expression and $1 =$ the highest intensity of an expression, i.e. the peak expression) of the sequence S_i in a dataset of m video sequences. Let $\boldsymbol{x}_i^{\theta} \in \mathbb{R}^D$ be the feature vector extracted from the $f_{i,\theta}$-th frame of the sequence S_i. The frame index $f_{i,\theta}$ can be calculated as follows:

$$f_{i,\theta} = n_i \times \theta, \tag{6.31}$$

where n_i is the number of frames in the sequence S_i. Therefore, $\{\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_m^1\} \in \mathbb{R}^D$ are the feature vectors extracted from the face images with high-intensity expressions, i.e. the last frames of the m video sequences. Suppose that $\{\boldsymbol{x}_1^{\xi}, \boldsymbol{x}_2^{\xi}, \ldots, \boldsymbol{x}_m^{\xi}\}$ are the feature vectors extracted from the corresponding low-intensity images, and the corresponding frame number in the respective video sequences is $f_{i,\xi}$. In our algorithm,

we use a different set of ξ values, where $0.6 \leq \xi \leq 0.9$, to learn the different expression manifolds.

6.4.2.1. Manifold learning with high and low-intensity training samples

A projection matrix $\mathbf{A}$ that maps the feature vectors $\mathbf{X}^1 = [\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_m^1]$ to a new subspace is first calculated using SLPM. The corresponding projected samples are denoted as $\mathbf{Y}^1 = [\boldsymbol{y}_1^1, \boldsymbol{y}_2^1, \ldots, \boldsymbol{y}_m^1] \in \mathbb{R}^d$ $(d \ll D)$, i.e. $\boldsymbol{y}_i^1 = \mathbf{A}^T \boldsymbol{x}_i^1$. Then, the same projection matrix $\mathbf{A}$ is used to map the low-intensity feature vectors $\mathbf{X}^\xi = [\boldsymbol{x}_1^\xi, \boldsymbol{x}_2^\xi, \ldots, \boldsymbol{x}_m^\xi]$, i.e. $\boldsymbol{y}_i^\xi = \mathbf{A}^T \boldsymbol{x}_i^\xi$, which should lie on the boundary of the corresponding expression manifold. The high-intensity and low-intensity samples in the subspace form a training matrix, denoted as $\mathbf{T}_\xi$, as follows:

$$\mathbf{T}_\xi = [\mathbf{Y}^1 \quad \mathbf{Y}^\xi] = [\mathbf{A}^T\mathbf{X}^1 \quad \mathbf{A}^T\mathbf{X}^\xi], \tag{6.32}$$

where ξ $(0 \leq \xi \leq 1)$ represents the intensity of the low-intensity images. Fig. 6-3(b) and Fig. 6-3(c) demonstrate the training data $\mathbf{T}_\xi$ with two different values of ξ on the CK+ database.

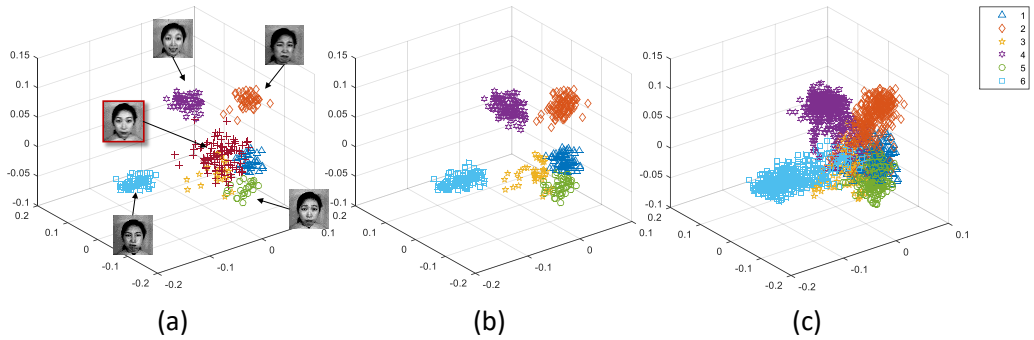

Fig. 6-4. The subspace learned using SLPM, with local descriptors "LPQ", based on the dataset named CK+: (a) the mapped features extracted from high-intensity expression images and neutral face images, (b) the mapped features extracted from high-intensity and low-intensity ($\xi = 0.7$) images, and (c) the mapped features extracted from high-intensity and low-intensity ($\xi = \{0.9, 0.8, 0.7, 0.6, 0.5, 0.4\}$) images.

Conventional manifold-learning methods map training samples, irrespective of how strong the expressing images are, as close as possible after transformation. This results in limited performance in terms of generalization. In our feature-generation algorithm, the subspace learning method, SLPM, is first applied to features extracted from high-intensity expressions. Then, features extracted from low-intensity expressions are mapped to the learned subspace. As observed in Fig. 6-4, features extracted from low-intensity expressions are located farther from the core samples (formed by high-intensity expressions) and near the boundary of the manifolds after the mapping.

The high-intensity samples are used to determine the centroid of an expression manifold, while those low-intensity samples are for representing the manifold boundary. The feature vectors are multi-dimensional, so a large number of low-intensity samples are required to represent the manifold boundary faithfully. However, only a small number of low-intensity images are available from the training video sequences. Furthermore, most of the existing expression databases have static images only. To solve this problem, we propose generating more low-intensity feature vectors for each expression class, so that the manifold learned for each expression class will be more accurate. In this paper, we consider the recognition of six facial expressions, i.e. anger, disgust, fear, happiness, sadness, and surprise. In addition to these facial expressions, we also consider the neutral expression in the proposed feature-generation method.

Let $\left\{\boldsymbol{x}_{s_1}^0, \boldsymbol{x}_{s_2}^0, \ldots, \boldsymbol{x}_{s_p}^0\right\} \in \mathbb{R}^D$ be the set of feature vectors extracted from neutral face images, where $\boldsymbol{x}_{s_i}^0$ is the feature vector of the neutral face image belonging to the subject s_i, and p is the number of the subjects in the dataset. The expression images of the subject s_i are denoted as

$$\mathbf{X}_{s_i}^1 = \left[\boldsymbol{x}_{s_{i,1}}^1, \boldsymbol{x}_{s_{i,2}}^1, \dots, \boldsymbol{x}_{s_{i,r}}^1\right], \tag{6.33}$$

where r is the number of expression images belonging to s_i and $\boldsymbol{x}_{s_{i,j}}^1$ is the feature vector extracted from the jth expression image of s_i. Then, the feature matrix for all the expressions is formed as follows:

$$\mathbf{X}_S^1 = [\mathbf{X}_{s_1}^1, \mathbf{X}_{s_2}^1, \dots, \mathbf{X}_{s_p}^1]. \tag{6.34}$$

The proposed sample-generation method operates in the learned subspace. Thus, the feature vectors extracted from the neutral face images and the expression images are all mapped to the learned subspace using the projection matrix **A** learned from $\mathbf{X}_S^1$, as follows:

$$\mathbf{Y}^1 = \mathbf{A}^T\mathbf{X}_S^1 = \left[\mathbf{Y}_{s_1}^1, \mathbf{Y}_{s_2}^1, \dots, \mathbf{Y}_{s_p}^1\right], \text{ and} \tag{6.35}$$

$$\mathbf{Y}^0 = \mathbf{A}^T\mathbf{X}_S^0 = \left[\boldsymbol{y}_{s_1}^0, \boldsymbol{y}_{s_2}^0, \dots, \boldsymbol{y}_{s_p}^0\right]. \tag{6.36}$$

Equations (6.35) and (6.36) represent the set of feature vectors of high-intensity expressions and neutral expressions of all subjects, respectively, in the subspace.

The proposed feature-generation method generates low-intensity feature vectors based on vector-pairs selected from two different sets: (1) vector-pairs from $\mathbf{Y}_{s_i}^1$, **(**2) vector-pairs from $\mathbf{Y}_{s_i}^1$ and $\boldsymbol{y}_{s_i}^0$. In the following sections, we will describe the feature-generation method with respect to two different vector-pairs.

6.4.2.2. Vector pairs from $\boldsymbol{Y}_{\boldsymbol{s_i}}^{\boldsymbol{1}}$ and $\boldsymbol{y}_{\boldsymbol{s_i}}^{\boldsymbol{0}}$

Let $\overline{\mathbf{Y}}_{s_i}^{\theta_{ne}} = \left[\boldsymbol{y}_{s_{i,1\to0}}^{\theta_{ne}}, \boldsymbol{y}_{s_{i,2\to0}}^{\theta_{ne}}, \dots, \boldsymbol{y}_{s_{i,r\to0}}^{\theta_{ne}}\right]$ be the feature matrix of possible low-intensity expressions with an intensity of θ_{ne} $(0 < \theta_{ne} < 1)$ belonging to the subject s_i, where $\boldsymbol{y}_{s_{i,j\to0}}^{\theta_{ne}}$ is the corresponding low-intensity feature vector generated using $\boldsymbol{y}_{s_{i,j}}^1$ and $\boldsymbol{y}_{s_i}^0$. In the rest of the paper, the arrow "→" indicates the direction of the feature vectors to be generated, with 0 and 1 being a neutral face image and a face image with the highest

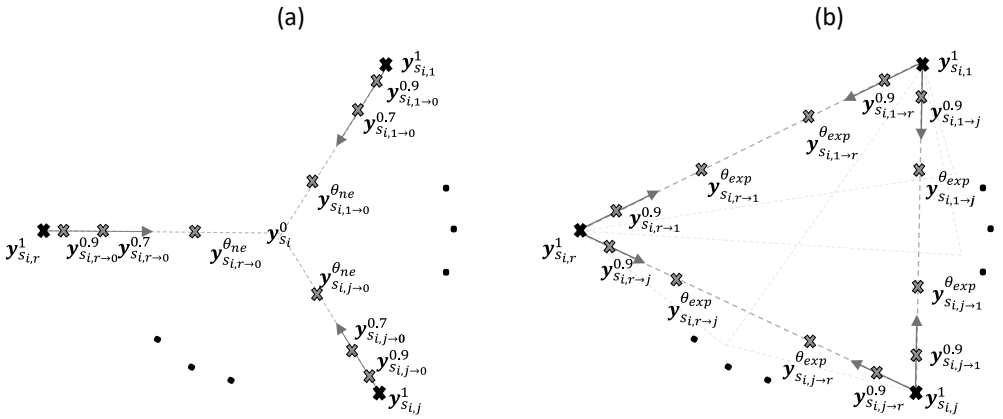

Fig. 6-5. The representation of the sample-generation process based on (a) feature vectors extracted from high-intensity images and neutral-face images, and (b) feature vectors extracted from high-intensity images.

intensity, respectively. $\boldsymbol{y}_{s_{i,j\to 0}}^{\theta_{ne}}$ means that the feature vector is generated in the direction from $\boldsymbol{y}_{s_{i,j}}^{1}$ to $\boldsymbol{y}_{s_i}^{0}$ where $\boldsymbol{y}_{s_{i,j}}^{1}$ is the mapped feature vector extracted from the *j*th expression image of s_i.

A set of feature vectors extracted from an expression video sequence, which starts from a neutral-expression face to the highest intensity of an expression, can be perceived as a path from the reference center, i.e. the neutral manifold, to a particular expression manifold wherein the distance of an expression manifold from the center is directly proportional to the intensity of the expression [30]. Therefore, for databases consisting of only static expression images, the feature matrix of possible low-intensity expressions can be obtained by assuming that the relation between the distance from $\boldsymbol{y}_{s_{i,j\to 0}}^{\theta_{ne}}$ to $\boldsymbol{y}_{s_i}^{0}$ and the expression intensity is linear. As illustrated in Fig. 6-5(a), the low-intensity feature vector $\boldsymbol{y}_{s_{i,j\to 0}}^{\theta_{ne}}$, belonging to s_i, can be computed as follows:

$$\boldsymbol{y}_{s_{i,j\to 0}}^{\theta_{ne}} = \theta_{ne} \cdot \boldsymbol{y}_{s_{i,j}}^{1} + (1-\theta_{ne}) \cdot \boldsymbol{y}_{s_i}^{0}. \tag{6.37}$$

Fig. 6-3(d) and Fig. 6-3(e) outline the training data with the feature generation using neutral images when $\theta_{ne} = 0.9$ and $\theta_{ne} = 0.7$, respectively. As seen in Fig. 6-3, both the absolute low-intensity feature vectors and the possible low-intensity feature vectors generated by linear interpolation have a similar structure.

6.4.2.3. Vector pairs from $\boldsymbol{Y}_{\boldsymbol{s_i}}^{\boldsymbol{1}}$

The respective expression manifolds can be far from each other in the learned subspace. For this reason, more features between expression manifolds are also needed. In the previous section, we proposed the idea that the feature vectors extracted from low-intensity expression images should be distant from the corresponding manifold centre, thus, this can enhance the generalizability of the learned manifold. Using a similar idea, more features that are distant from the manifold centres can be generated using vector-pairs from the feature matrix of high-intensity expressions of the same subject, $\mathbf{Y}_{s_i}^1$, as illustrated in Fig. 6-5(b). A feature vector $\boldsymbol{y}_{s_{i,j\to k}}^{\theta_{exp}}$, which lies on the line from the *j*th expression-vector of s_i, $\boldsymbol{y}_{s_{i,j}}^1$, to the *k*th expression-vector of s_i, $\boldsymbol{y}_{s_{i,k}}^1$, with a weight θ_{exp} ($0 < \theta_{exp} < 1$) can be computed as follows:

$$\boldsymbol{y}_{s_{i,j\to k}}^{\theta_{exp}} = \theta_{exp} \cdot \boldsymbol{y}_{s_{i,j}}^1 + \left(1 - \theta_{exp}\right) \cdot \boldsymbol{y}_{s_{i,k}}^1, \tag{6.38}$$

Suppose that $c_j = l(\boldsymbol{y}_{s_{i,j}}^1)$ and $c_k = l(\boldsymbol{y}_{s_{i,k}}^1)$ are the expression classes of the *j*th and the *k*th expression vectors, respectively, and n_{i,c_j} and n_{i,c_k} are the number of expression-vectors of expression classes c_j and c_k, respectively, belonging to subject s_i. Then, a total of $n_{i,c_j} n_{i,c_k}$ feature vectors can be generated. The feature matrix consisting of the generated features using the pairs from $\mathbf{Y}_{s_i}^1$ can be denoted as follows:

$$\bar{\mathbf{Y}}_{s_i,exp}^{\theta_{exp}} = \left[\boldsymbol{y}_{s_{i,1\to 2}}^{\theta_{exp}}, \boldsymbol{y}_{s_{i,1\to 3}}^{\theta_{exp}}, \dots, \boldsymbol{y}_{s_{i,1\to r}}^{\theta_{exp}}, \dots, \boldsymbol{y}_{s_{i,r\to 1}}^{\theta_{exp}}, \boldsymbol{y}_{s_{i,r\to 2}}^{\theta_{exp}}, \dots, \boldsymbol{y}_{s_{i,r-1\to r}}^{\theta_{exp}}\right]. \tag{6.39}$$

The training matrix, $\mathbf{T}_\theta$, is updated to $\overline{\mathbf{T}}_\theta$, which is used as a static database, as follows:

$$\overline{\mathbf{T}}_\theta = \left[\mathbf{Y}^1 \quad \overline{\mathbf{Y}}_{ne}^{\theta_{ne}} \quad \overline{\mathbf{Y}}_{exp}^{\theta_{exp}}\right], \tag{6.40}$$

where $\overline{\mathbf{Y}}_{ne}^{\theta_{ne}} = \left[\overline{\mathbf{Y}}_{s_{1.ne}}^{\theta_{ne}}, \overline{\mathbf{Y}}_{s_{2,ne}}^{\theta_{ne}}, \dots, \overline{\mathbf{Y}}_{s_{p,ne}}^{\theta_{ne}}\right]$ and $\overline{\mathbf{Y}}_{exp}^{\theta_{exp}} = \left[\overline{\mathbf{Y}}_{s_{1.exp}}^{\theta_{exp}}, \overline{\mathbf{Y}}_{s_{2,exp}}^{\theta_{exp}}, \dots, \overline{\mathbf{Y}}_{s_{p,exp}}^{\theta_{exp}}\right]$.

In our experiments, we vary the θ_{ne} and the θ_{exp} values from 0.7 to 0.9. Fig. 6-2 lists the overall flow of the proposed algorithm.

When a feature vector is generated, it is checked whether or not it is closest to its corresponding manifold class. Furthermore, the feature vectors are generated solely for the pairs of clusters that are in close proximity to each other in the learned subspace.

6.5. Experimental Setup and Results

6.5.1. Experimental setup

In our experiments, four facial-expression databases were used to show the robustness and performances of the proposed methods: 1) Bahcesehir University Multilingual Affective Face Database (BAUM-2) [67], 2) Extended Cohn-Kanade (CK+) [116] database, 3) Japanese Female Facial Expression (JAFFE) [155] database, and 4) Taiwanese Facial Expression Image Database (TFEID) [34].

The BAUM-2 multilingual database consists of short videos extracted from movies. In our experiments, an image dataset, namely BAUM-2i, consisting of images with peak expressions from the videos in BAUM-2, is considered. There are 829 face images from 250 subjects in the BAUM-2i static expression dataset, which express 6 basic emotions. However, only 536 of them, which have their facial-feature points provided, are considered in our experiments. Since the BAUM-2 database was created by extracting images from movies, the images are close to real-life conditions (i.e.

under pose, age, and illumination variations, etc.), and are more challenging than those in acted databases.

The CK+ dataset contains a total of 593 posed sequences across 123 subjects, of which 304 of the sequences have been labelled with one of the six discrete emotions, which are anger, disgust, fear, happiness, sadness, and surprise. Each sequence starts with a neutral face and ends with a frame of peak expression. The last frame of each sequence, and the first frames of the sequences that have unique subject labels, as well as their landmarks provided, are used for recognition. There are a total of 414 face images. Note that some of the first frames are also discarded because the expressed emotions are of low intensity. JAFFE consists of 213 images from 10 Japanese females, which express 6 basic emotions – anger, disgust, fear, happiness, sadness, surprise – and neutral. JAFFE is also a widely used acted database, which means that it was recorded in a controlled environment. The TFEID database contains 268 images, with the six basic expressions and the neutral expression, from 40 Taiwanese subjects. Like CK+ and JAFFE, TFEID is also an acted database.

Each of the above-mentioned databases has its own characteristics. **Table 6-3** shows the number of images for each expression class for the different databases.

Table 6-3. A comparison of the number of images for different expression classes in the databases used in our experiments

	BAUM-2	**CK+**	**JAFFE**	**TFEID**
Anger	80	45	30	34
Disgust	32	59	29	40
Fear	35	25	32	40
Happiness	139	69	31	40
Sadness	68	28	31	39
Surprise	83	82	30	36
Neutral	99	106	30	39
TOTAL	**536**	**414**	**213**	**268**

Although some of the databases also have the contempt expression, only the six basic prototypical facial expressions (i.e. anger, disgust, fear, happiness, sadness, and surprise), as well as the neutral facial expression, are considered in our experiments. Please note that neutral facial expression has been used only for creating feature vectors of low-intensity expressions.

Subspace-learning methods are often applied to feature vectors formed by the pixel intensities of face images. In our method, features are first extracted using the state-of-the-art local descriptors, and then a subspace-learning method is applied for manifold learning and dimensionality reduction. The usual way of using local descriptors is to divide a face image into a number of overlapping or non-overlapping regions, then extract features from these regions, and finally concatenate them to form a single feature vector. In this way, local information, as well as spatial information, can be obtained. Another way of using local descriptors is to consider only the regions that have more salient information about the considered expression classes. Following this idea, features extracted from the eye and mouth regions are used in [220], which showed that features extracted from these regions only can achieve higher recognition rates than those extracted by dividing face images into sub-regions.

In our experiments, face images from the different databases are all scaled to the size of 126×189 pixels, with a distance of 64 pixels between the two eyes. To determine the eye and mouth windows, the facial landmarks, i.e. the eyes and mouth corners, are used. If facial landmarks are not provided for a database, the required facial-feature points are marked manually. The eye window and the mouth window are further divided into 12 and 8 sub-regions, respectively. The nearest neighbour classifier and SVM with linear kernel are used in the experiments.

6.5.2. Experimental results

In this section, we evaluate the performances of our proposed method, using four different descriptors, on the four different databases. We also compare our method with four subspace-learning methods, as well as without using any subspace-learning method.

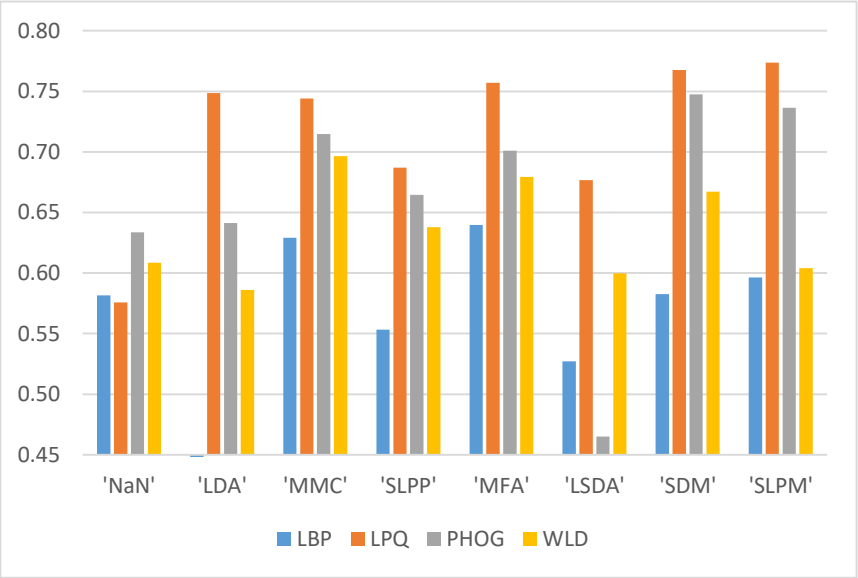

Fig. 6-6. The recognition rates of the different subspace methods, with different local descriptors, based on a combined dataset of BAUM-2, CK+, JAFFE & TFEID.

Firstly, the four acted databases, i.e. BAUM-2, CK+, JAFFE, and TFEID, are combined to form a single dataset, called COMB4, so that we can better measure the general performances of the different subspace-learning methods and the descriptors.

Fig. 6-6 shows that MFA, SDM, and SLPM are the three best subspace-learning methods, which outperform the other subspace-learning methods. The LPQ local descriptor achieves the highest recognition rates, for the different subspace-learning methods, on COMB4. Therefore, the subspace-learning methods, MFA and SDM, and the local descriptor, LPQ, are chosen to further compare the performance of the proposed method on each of the individual datasets. In Fig. 6-6, we can also observe

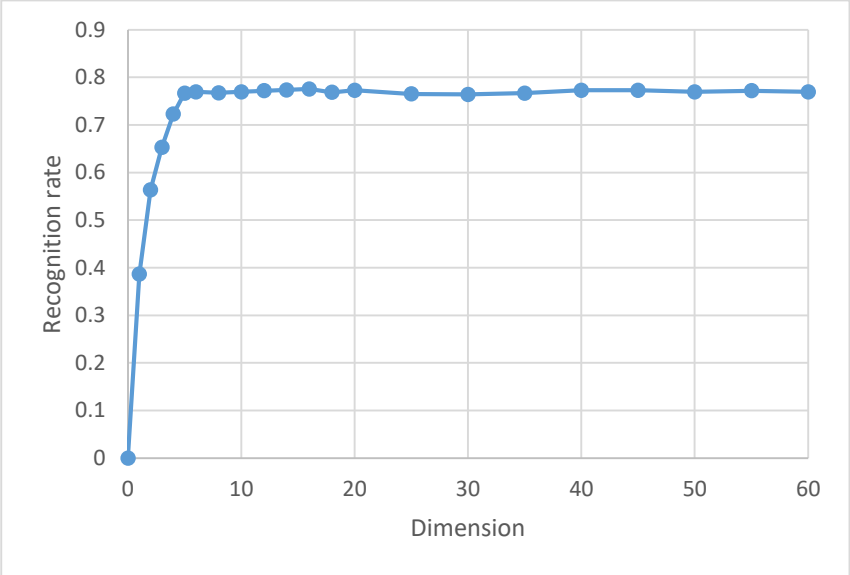

Fig. 6-7. Recognition rates of our proposed method in terms of different dimensions.

that SDM outperforms most of the subspace-learning methods, except SLPM, because the intra-class spread is adjustable. Furthermore, SDM is also computationally simpler than the other compared methods, but it does not incorporate the local geometry of the data. In our proposed method, information about local structure is incorporated into the objective function. Thus, SLPM can achieve higher recognition rates than SDM.

Fig. 6-7 shows the recognition rates of SLPM on COMB4, with the dimensionality of the subspace varied. The results show that SLPM has converged to its highest

Table 6-4. The comparison of recognition rates obtained by using low-intensity images with different l values on the CK+ database, using the LPQ feature.

METHODS	CK+
SLPM	94.81%
SLPM + $\xi = 0.9$	**95.45%**
SLPM + $\xi = 0.8$	94.16%
SLPM + $\xi = 0.7$	93.51%
SLPM + $\xi = 0.6$	91.88%

recognition rate, when the dimensionality is lower than 10. In other words, our method is still very effective even at a low dimensionality. Based on these results, we set the

Table 6-5. The comparison of subspace learning methods on different datasets, with the LPQ descriptor being used with nearest neighbor classifier.

	BAUM-2	**CK+**	**JAFFE**	**TFEID**
MFA	62.01%	93.83%	89.07%	91.70%
SDM	62.01%	93.51%	89.07%	92.58%
SLPM	62.93%	94.81%	90.71%	93.45%
SLPM + $\boldsymbol{\theta_{exp} = 0.9 + \theta_{ne} = 0.9}$	**63.62%**	94.81%	91.26%	93.45%
SLPM + $\boldsymbol{\theta_{exp} = 0.8 + \theta_{ne} = 0.8}$	62.93%	**96.10%**	91.26%	**94.32%**
SLPM + $\boldsymbol{\theta_{exp} = 0.7 + \theta_{ne} = 0.7}$	63.16%	95.13%	**91.80%**	93.89%
SLPM + $\boldsymbol{\theta_{exp} = 0.6 + \theta_{ne} = 0.6}$	62.01%	94.16%	90.71%	93.45%

subspace dimensionality at 11 in the rest of the experiments.

To investigate the effect of the use of images of expression with low intensities, several experiments have been conducted on the CK+ database. As shown in Table 4, the recognition rate is the highest when $\xi = 0.9$. Table 5 and Table 6 show the recognition rates of the three subspace-learning methods, MFA, SDM and SLPM, as well as SLPM, using feature generation with different θ_{exp} and θ_{ne} values, with the LPQ descriptor, on the four different databases using nearest neighbour classifier and

Table 6-6. The comparison of subspace learning methods on different datasets, with the LPQ descriptor being used with SVM classifier.

	BAUM-2	**CK+**	**JAFFE**	**TFEID**
MFA	61.10%	92.21%	**91.26%**	91.70%
SDM	60.18%	92.21%	89.62%	92.58%
SLPM	63.16%	92.53%	89.62%	93.01%
SLPM + $\boldsymbol{\theta_{exp} = 0.9 + \theta_{ne} = 0.9}$	**63.84%**	92.86%	**91.26%**	94.76%
SLPM + $\boldsymbol{\theta_{exp} = 0.8 + \theta_{ne} = 0.8}$	62.47%	93.83%	**91.26%**	**95.20%**
SLPM + $\boldsymbol{\theta_{exp} = 0.7 + \theta_{ne} = 0.7}$	62.24%	**94.48%**	89.07%	94.32%
SLPM + $\boldsymbol{\theta_{exp} = 0.6 + \theta_{ne} = 0.6}$	61.56%	**94.48%**	88.52%	94.32%

SVM classifier, respectively. It can be found that SLPM achieves the best classification performance again, when compared to the other methods. The classification performance is further improved by up to 2%, when feature generation is employed. Furthermore, as observed in **Table 6-5** and **Table 6-7**, the nearest neighbour classifier outperforms the SVM classifier in most of the databases. Lastly, additional experiments were conducted to validate the efficiency of the proposed

Table 6-7. Comparison of the runtimes (in milliseconds) required by the different subspace learning methods (MFA, SDM, and SLPM) on different datasets, with the LPQ descriptor used.

	BAUM-2	**CK+**	**JAFFE**	**TFEID**
MFA	96	69	45	51
SDM	151	133	120	118
SLPM	**65**	**37**	**23**	**25**

subspace learning methods. **Table 6-6** tabulates the runtimes in milliseconds for each of the subspace learning methods. We can see that SLPM is twice as fast as MFA, which solves the generalized eigenvalue problem instead of calculating eigenvalue decomposition like SLPM. SDM is much slower than MFA and SLPM.

6.6. Conclusion

In this chapter, we have proposed a subspace-learning method, named Soft Locality Preserving Map (SLPM), which uses the neighbourhood and class information to construct a projection matrix for mapping high-dimensional data to a meaningful low-dimensional subspace. The difference between the within-class and between-class matrices is used to define the objective function, rather than the Fisher criteria, in order to avoid the singularity problem. Also, a parameter β is added to control the within-class spread, so that the overfitting problem can be solved. The robustness and the generalizability of SLPM have been analysed on four different databases, using four

different state-of-the-art descriptors and two different classifiers, and SLPM has been compared with other subspace-learning methods. Moreover, we have proposed using low-intensity expression images to learn a better manifold for each expression class. By taking advantage of domain-specific knowledge, we have proposed two methods of generating new low-intensity features in the subspace. Our experiment results have shown that SLPM outperforms the other subspace-learning methods, and is a good alternative to performing dimensionality reduction on high-dimensional datasets. Our experiment results, also, have shown that the proposed feature-generation method can further increase the recognition rates.

Chapter 7. A new novel database: Facial Expressions of Comprehension (FEC)

7.1. Introduction

As explained in Section 2.2.6, with the great advancements in computer vision and machine learning techniques, a promising new literature is developing that uses dynamic facial expression data to interpret the facial expressions in the wild. Researchers have studied recognizing pain [10, 114, 154, 195], diagnosing depression [37], automatic estimation of the level of taste liking [51], detection of the presence of ADHD/ASD [105], identifying learners' affective states [122]. Closer to the goals of the study presented in this thesis, Sathik et al. [199] investigated student learning in a classroom setting through the correlation of successful comprehension with positive expressions, and failed comprehension with negative expressions. The authors did not build a model to predict students' comprehension but instead explored the statistical correlation of expressions towards comprehension types.

To date, none have investigated the multidimensionality of comprehension. Expanding the search to related fields brings us to research in psychology, where neuroscience tools matched with computational modelling, and experimentation, have laid a groundwork for the study of memory and learning specifically about language processing [80, 84, 87, 157]. Facial behavior analysis has much to offer this growing literature in that it may well add another tool to the toolbox of identifying underlying states of cognition, much like the introduction of eye-tracking technology did over two decades ago [9, 47, 102, 227]. The current study differs critically from previous work in that we analyze online aspects of comprehension as they occur dynamically across the face. We do not depend on labelled facial configurations of distinct emotions, nor

positions within a continuous model, but rather analyze changes in facial configuration indicative of differing stages of comprehension. To this end, it is necessary to utilize experimental methodologies current in studies of cognitive processing.

One area of investigation that allows for controlled yet natural responses to online stimuli is sentence processing. To fully comprehend a sentence, one must manage multiple different types of relationships, such as the morpho-syntactic and semantic-thematic relationships between the component parts of the sentence, and the relationship between the sentence's resulting meaning and its associations that have been encountered before and stored within long-term semantic memory [128]. Several studies in cognitive neuroscience have examined the neural networks that are active in processing these relationships by investigating the brain responses when each is violated, i.e. syntactic, semantic-thematic, and world knowledge [80, 84, 87, 128, 157]. To study the facial expressions of sentence comprehension and to propose a methodology for future studies, we choose only world knowledge violations in this study.

We introduce a multimodal spontaneous facial expression database, named 'Facial Expressions of Comprehension (FEC)'. The FEC database was created using data obtained in a behavioral experiment in which participants were asked to first read knowledge-based statements, second provide a 'true' or 'false' answer, and finally, receive feedback as either 'correct' or 'false'. During the experiment, the Kinect v2 device was used to record multiple streams of data. Using the FEC database, we analyze the facial behavior not only during successful recall of target information, but during unsuccessful recall as well as the expressions during guessing. The findings can be later used to design a classification system that can predict participants' possible achievements. In addition to these online measures during the reading of true

and false statements, facial expressions are also analyzed during the feedback stage. To analyze the dynamic changes in facial configuration in the feedback stage, we propose Event-Related Intensities (ERIs), which is a new approach to explore the intensity changes in facial expressions caused by an event, e.g. their achievement status.

The organization of the rest of this chapter is as follows. In Section 7.2, the data collection method is explained in detail with respect to participants, stimuli and the experimental design. In Section 7.3, a new facial expression database is introduced and the novel features are listed. In Section 7.4, the methodologies used in our data analysis are explained. Section 7.5 presents the comprehensive data analysis and experiments using facial signals, such as animation units. Finally, Section 7.6 concludes the paper.

7.2. Data Collection

In this section, the experimental design and the data collection process are explained in detail. The constructed database, based on the collected data, is then described in Section 7.3.

7.2.1. Participants

Forty-four healthy volunteers (twenty woman and twenty-four man) aged between 20 and 37 years (mean = 27, SD = 3.89) from 16 different nationalities participated in the study. All participants had normal or corrected-to-normal visual acuity. Self-rated English proficiency was also collected from the participants (mean = 7.61, SD = 1.57, with 10 as a native speaker). Participants gave informed written consent to the experimental procedure. This study was approved by the local ethics committee of The Hong Kong Polytechnic University.

7.2.2. Stimuli

Factually correct and incorrect statements are also in interest of educational studies. There have been a wide range of discussions on the validity of true/false test items in testing students' comprehension and understanding, and the presence of guessing in a true-false test [22, 59, 75]. There exists three status of knowledge: full knowledge, partial knowledge and misinformation. Partial knowledge and misinformation would lead to informed guessing and blind guessing, respectively [59]. The questions are designed in a way that partial knowledge is eliminated as much as possible, since it will cultivate informed guessing, and it is important for our experimental design to keep the difficulty level of questions as distinct as possible.

The experimental material consisted of 100 true or false world-knowledge statements. Selection of the final stimuli began with 240 statements generated from a variety of general knowledge areas: mathematics, science, and technology, as well as art, entertainment, history and geography. Amazon Mechanical Turk was then used to assess each item's difficulty level based on accuracy rate. Four categories were created (25 Baseline TRUE, 25 Baseline FALSE, 25 Range TRUE, and 25 Range FALSE), in which Baseline statements were those that received a high accuracy rate, and Range

Table 7-1. Example statements according to their true/false category.

Statements	Category
There are / 360 degrees / in a circle.	Baseline TRUE
An apple / is larger than / a grape.	Baseline TRUE
The Earth / rotates around / the Moon.	Baseline FALSE
Antarctica / is a province of / France.	Baseline FALSE
Dianne Wiest / won best supporting actress / in 1995.	Range TRUE
The Ig Nobel Prizes / have been awarded / since 1991.	Range TRUE
In 1912, / Jean Sibelius led / his Fourth Symphony premiere.	Range FALSE
Diet Coke / was invented / in 1970.	Range FALSE

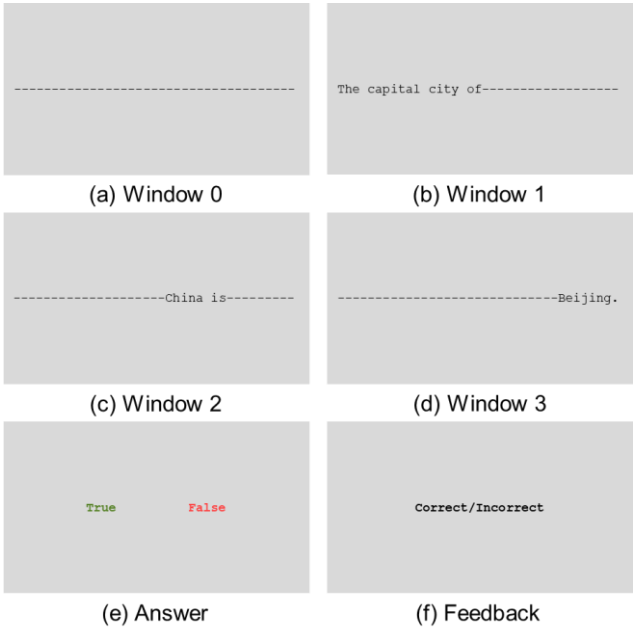

Fig. 7-1. Screenshots of 6 experimental windows seen by participants during each trial.

statements were those that received roughly 50% accuracy, i.e., were the product of guessing. For the purposes of presentation, each statement was segmented into three parts. Table 7-1 presents example stimuli used in the experiment from each category.

7.2.3. Experimental design and procedure

The behavioural experiment was designed using PsychoPy, a stimulus delivery library for the Python programming language [186, 187]. PsychoPy allows for the online recording of participants' responses and response times. The experiment was designed as a self-paced reading task, in which participants control the progression of each experimental window, and therefore its duration, through the press of the space button. Each trial within the experiment consists of 6 windows, as can be seen in Fig. 7-1. In Window 0, the sentence stimuli are masked with dashes that are equal to the length of the sentence if displayed. In Window 1, only the first of three sentence segments is revealed, while the remaining segments remain masked. Windows 2 and 3 similarly

display only the consecutive segments of the sentence while masking the remaining parts. During the Answer, participants are asked to evaluate the veracity of the statement by clicking one of the highlighted texts on the computer screen indicating “True” or “False”. Finally, participants are given feedback in Feedback as either “correct” or “incorrect”. Each trial ends after a 500 millisecond inter-stimulus interval.

During the experiment, participants sat in front of a computer screen in a quiet room. A Kinect 2.0 device, capable of capturing high-resolution RGB videos, was placed on top of the computer screen to save the relevant information during the experiment. All animation units, as well as RGB, depth, and infrared streams obtained from the Kinect 2.0 device, were saved to a hard drive during the experiment.

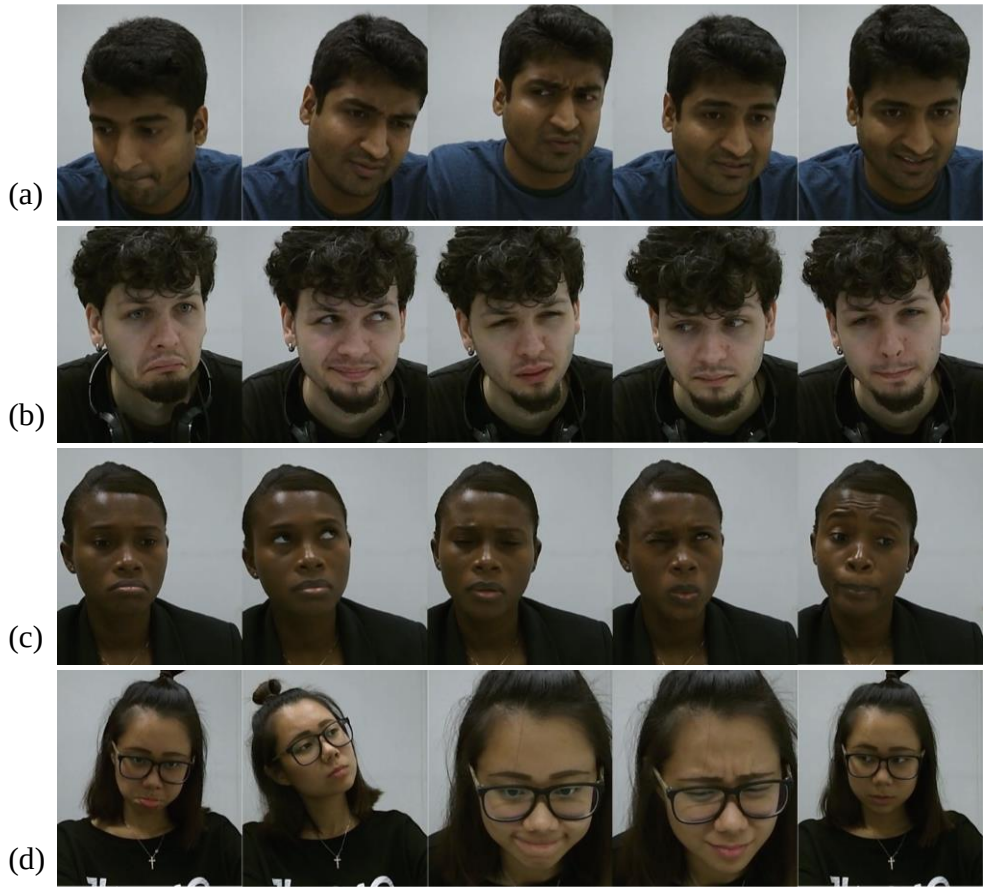

Fig. 7-2. Images selected from the video recordings in the FEC database for (a) subject 1, (b) subject 10, (c) subject 23, and (d) subject 43.

7.3. A Facial Expression Database: Facial Expressions of Comprehension (FEC)

7.3.1. Database organization

The data collected from the Kinect 2.0 device is divided into segments, where each segment has the Animation Units and the frames belonging to only one trial answered by one participant, i.e. per trial and per participant. It means that one participant would have 100 RGB video shots, 100 depth video shots, etc. corresponding to 100 stimuli.

We created the FEC database, which consists of 3,576 unique segments from 44 participants and 100 stimuli. Because of an experimental flaw while saving the data using the Kinect 2.0 device, not all the participants have 100 unique segments, although their responses to all of the statements were saved using the PsychoPy experiment. Therefore, we cleaned up the database by deleting these segments where the participants got distracted by external factors. Each unique segment has the RGB, depth and infrared video shots, as well as 17 animation units per frame provided by the Kinect 2.0 device. The average length of the videos is 9.1990 seconds (STD: 3.6962 seconds). In Fig. 7-2, some example frames from the facial clips in the FEC database are shown.

7.3.2. Annotation of video clips

The labels that each video clip has are limited to the difficulty category of the statement (baseline or range), the veracity of the statement (true or false), and the achievement status of the participants (correct or incorrect). Each frame is annotated with a window name with respect to the computer screen at which they look, as illustrated in Fig. 7-1. In addition, the metadata are generated, including the subject's age, gender, the total years of education, self-rated English proficiency, and the

information of whether the video has occlusion, and whether the participant in the corresponding video is talking.

7.3.3. Features of the database

The novelties of the FEC database can be highlighted as follows:

1. The FEC database consists of facial video clips whose expressions are induced by our cognitive/affective experimental setup. Although the participants knew that their videos were recorded during the experiment, we can assume that the expressions performed by the participants are spontaneous, since their main focus is not performing naturalistic expressions but answering questions.
2. The FEC database consists of facial expressions induced by both a knowledge stimulus, i.e. recalling the information, and an affective stimulus, i.e. learning the achievement status. To the best of our knowledge, FEC database is the first database that has two types of induced expressions.
3. The video clips were recorded in a room, which gave a lot of control over illumination and other environmental conditions.
4. Instead of selecting expressive video shots from a long video recording and labelling them with action units or with emotion labels like the other facial expression databases, we present the raw video recordings that are cut into several segments with respect to question numbers. This means that we provide the videos both with and without a specific expression with high intensity.
5. The video clips in the FEC database do not contain facial expressions temporally standardized as onset-apex-offset or onset-apex, since the videos are not manually selected or cut to fit one of those temporal sequences. We believe that this property of the database can promote new methodologies in the analysis of the facial behavior in videos.

6. The FEC database is a multicultural database that includes participants from 16 different nationalities, which, to the best of our knowledge, is a feature that does not exist in other facial expression databases.
7. The FEC database is a multi-modal database that has video clips from different streams: RGB, depth, and infrared.
8. The FEC database is richly annotated, using the attributes listed in the previous section.
9. The 2D positions of the tracked facial landmarks on the face at each frame are also provided. To identify the facial landmark points, we use the two state-of-the-art methods in [12] and [103].

7.4. Our Proposed Method

7.4.1. Measuring facial signals

The facial activity of a participant in a video can be represented as a multivariate time series, e.g. facial signals, where each time series refers to the intensities of one of the AUs or the AnUs over time. In this paper, we utilize two different measurements of facial activity to analyse the experimental data: Action Unis (AUs) in FACS and Animation Units (AnUs) obtained from the Kinect v2 device.

Facial Action Coding System (FACS): One of the most well-known approaches to facial expression recognition is the Facial Action Coding System (FACS) [62] developed by Ekman and Friesen based on the muscular and anatomical basis of facial movement. FACS can be thought of as a dictionary of facial muscle movements, i.e. action units (AUs), where any facial expression theoretically can be decomposed as a combination of one or more AUs. There are 32 AUs: 9 AUs belonging to the upper-face, 18 AUs to the lower-face, and 5 AUs have no particular region [185]. With the enhancements in the computer vision techniques, AU detection or recognition from a

single frame or videos with the use of appearance or shape information have received broad attention by researchers [12, 103]. Another part of the AU research is to identify the intensity of the occurring AUs in the range of [0,1], where an AU occurs in its highest intensity at 1. Two state-of-the-art methods, OpenFace [12] and the method described in [103], are used to extract AUs from the videos provided in the FEC database.

Animation Units from the Kinect v2 Device: The Kinect v2 device provides Animation Units (AnUs) based on face-shape deformations and depth information. Each AnU is expressed as a numeric weight, where the values of three AnUs, i.e. jaw slide right, left eyebrow lower, and right eyebrow lower, vary in the range $[-1, 1]$ and the others in $[0,1]$. As mentioned in [104], the AnUs cannot detect the facial behaviour caused only by appearance changes, but they are considered reliable since they are

Table 7-2. Animation Units and corresponding Action Units.

Part of Face	Animation Units	Action Units
Upper Face	Right eyebrow Lowerer	Brow Lowerer (AU4)
	Left eyebrow Lowerer	
	Left cheek Puff	Cheek Raiser (AU6)
	Right cheek Puff	
	Left eye Closed	Blink (AU45)
	Right eye Closed	
Lower Face	Lip Corner Puller Left	Lip Corner Puller (AU12)
	Lip Corner Puller Right	
	Lip Corner Depressor Left	Lip Corner Depressor (AU15)
	Lip Corner Depressor Right	
	Lip Stretcher Right	Lip Stretcher (AU20)
	Lip Stretcher Left	
	Jaw Open	Jaw Drop (AU26)
	Lip Pucker	--
	Jaw Slide Right	--
	Lower-lip Depressor Left	--
	Lower-lip Depressor Right	--

obtained from the RGBD data. Table 7-2 presents the AnUs and their corresponding AUs. Please note that AnUs do not exactly correspond to the AUs, but are just analogous to each other.

7.4.2. Event-Related Intensities (ERIs)

Face morphology is a facial configuration at a given time, e.g. static images. Facial dynamics, on the other hand, refers to the temporal evolution of a facial configuration from one to another in terms of timing, duration, speed of activation and deactivation, e.g. videos [159]. Several studies have shown that face dynamics are useful for a higher-level interpretation of facial signals [8, 64], and the detection of the temporal phases of a facial activity, such as neutral, onset, apex and offset, has been widely studied [46, 226, 253, 254, 257]. More studies, recently, have been investigating the detection of an anomaly in a video sequence represented as a facial signal, such as pain detection [195]. However, to the best of our knowledge, there exists no methodology to interpret the facial activity caused by a stimulus.

Our behavioral experiment design includes two different types of stimuli: knowledge stimuli and affective stimuli. These two stimuli do not just differ in terms of the neural activity required to process stimuli, but also in the temporal effects on facial activity. Knowledge stimuli consist of continuous activity of the brain and face, because the person of interest is required to read a sentence where the affective stimulus is a short stimulus whose effect on facial activity can be observed in a short period of time, following the presentation of the stimulus, i.e. feedback.

Responses, which are the direct results of a thought process or perception, have been extensively studied by the experimental psychologists and neuroscientists, e.g. Event-Related Potentials (ERPs) [162, 213]. ERPs are voltage fluctuations measured using electroencephalography (EEG), i.e. electrical potentials generated by the brain.

ERPs reflect the stages of information processing, i.e. neural activity, stimulated by sensory-related operations, memory-related operations, affective operations, etc. [126]. The ERPs are time-locked waveforms usually obtained by averaging signals over multiple trials. Most of the ERP studies explore electrophysiological responses that occur up to and within 1500 ms following a stimulus, because the changes of the event-related voltage happen in the brain in a short time of periods. For example, P300, an ERP component at 300 ms (300-500 ms), is related to the neural activity involving memory processing [69, 110].

Inspired by the ERPs in the EEG studies, we propose Event-Related Intensities (ERIs) that refer to the intensity changes of facial configuration caused by an event. If the facial activity signals in a video, represented by either AnU or AU, is assumed to be analogous to EEG waveforms, the ERIs can be defined as the facial dynamics that happened following a presentation of stimulus, e.g. their achievement as correct or incorrect. Unlike ERPs that investigate the voltage potentials in the waveform in a small range, e.g. around 200 ms, any time after the stimulus representation, ERIs are proposed to be the representation of all the intensity changes that happen in the first 1500 ms following a stimulus.

Let us define the time series belonging to the *i*-th AnU, anu_i, in a video as a set of real values $\mathrm{X}_{anu_i} = \left\{x_{anu_i}^{(1)}, x_{anu_i}^{(2)}, \dots, x_{anu_i}^{(n_i)}\right\}$, where $x_{anu_i}^{(t)} \in \mathbb{R}$ is the *t*-th element in the time series and n_i is the number of frames in the video. When $t = n_f$ is the time index of the beginning of the feedback window, the average of the previous f elements is, first, calculated as follows:

$$x_{anu_i}^{avg,n_f} = \frac{1}{f}\sum_{t=n_f-f}^{n_f-1} x_{anu_i}^{(k)}. \tag{7.1}$$

Then, all the following elements of the time series, starting from $t = n_f$, are normalized using $x_{anu_i}^{avg,n_f}$ as follows:

$$\bar{x}_{anu_i}^{(t)} = x_{anu_i}^{(t)} - x_{anu_i}^{avg,n_f}, \tag{7.2}$$

where $t = n_f, n_f + 1, \dots, n_i$. The resulting set of values $\overline{\mathrm{X}}_{anu_i} = \left\{\bar{x}_{anu_i}^{(n_f)}, \bar{x}_{anu_i}^{(n_f+1)}, \dots, \bar{x}_{anu_i}^{(n_i)}\right\}$ is called the ERI belonging to the *i*-th AnU.

7.4.3. Dynamics of head motion

Visual cues, such as head movement and facial expressions, play a significant role in affective expression and human communication. Among the various visual cues, facial expressions and gestures have been studied extensively. In the last decade, dynamics of head motion have started to receive more attention. For example, rigid head motion was studied with respect to speech, such as the acoustic perception of speech [171], recognition of the speaker identity [89], and expressive speech animation [23], in addition to interaction between distressed couples and between a mother and her infant [86], and emotional expression of infants [85]. However, understanding the dynamics of head motion, related to the expression of affective and mental states, is still an open research area.

In addition to the data analysis with facial signals, i.e. ERIs, we also analyse the rigid head motion caused by the achievement status in terms of pitch, yaw and roll. Angles of pitch, yaw and roll in radians and the rigid shape parameters of the Point Distribution Model (PDM), such as scale, translation in terms of x and y coordinates and rotation in terms of x, y and z coordinates, are obtained using OpenFace [12] for each frame belonging the feedback window separately. The dimension of the features referring to the angles of pitch, yaw and roll is three for each frame where the dimension of the features obtained from the parameters of the PDM is six. These

features are used to investigate the correlation between rigid head motion and the feedback (i.e. achievement status). Similar to ERIs, the head motions are normalized to consider the changes, only after the participants are shown their achievement status.

7.4.4. Sentence comprehension

At the cognitive level, the construction of an interpretation of a sentence requires not only combining the meaning of words and phrases followed by computing their thematic and syntactic relations, but also using world knowledge [112]. At the brain level, sentence comprehension activates a network of neurons whose activation areas and degrees differ with the type and the complexity of the sentence. The psycholinguists and neuroscientist have been extensively investigated sentence comprehension in the past with EEG and fMRI [79, 84, 128]. Recently, the studies have been started using eye tracking technology [9, 47, 102, 227]. This study aims to introduce a new tool to understand the sentence comprehension through facial dynamics and vice versa.

In the first part of the behavioural experiment design, we manipulated comprehension by altering the veracity of the statements to investigate the facial dynamics during online sentence comprehension, in case of world-knowledge violations. We also controlled the difficulty of the world-knowledge to explore the facial dynamics when the participants lack the required knowledge. It impels us to analyse participants' facial behaviours during online sentence comprehension in terms of two aspects: 1) whether we can identify that they knew the question of interest or simply guessed, i.e. knowing face versus guessing face, and 2) whether we can identify the veracity of the statements through their facial dynamics.

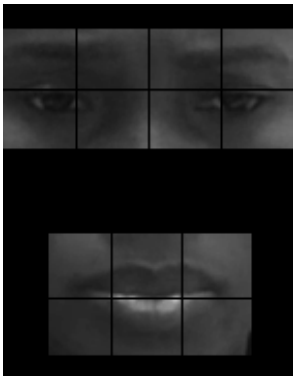

Fig. 7-3. The eye and the mouth regions used in our experiments to extract the LPQ-TOP features.

To achieve the analysis, the discriminant features are extracted using two different sources that represent the facial dynamics in different perspectives: (1) the facial video clips, and (2) the animation units.

The faces in the videos are aligned based on the eye positions and are cropped to 140×180 pixels. Then, the eye and the mouth regions are established as shown in Fig. 7-3. The eye and mouth regions are divided into 2×4 and 2×3 subregions, respectively, followed by extraction and concatenation of the local features from those subregions. The local features are based on a spatiotemporal descriptor, called "Local Phase Quantization Three Orthogonal Planes (LPQ-TOP) [108]." LPQ-TOP, which is a spatiotemporal extension of LPQ to three orthogonal planes (TOP), is selected, because in [221], we observed that LPQ is an effective and discriminative feature of facial expression and has achieved excellent performance in facial expression recognition. Facial expression for comprehension can be viewed as an expression in between micro and macro-expressions. TOP is an effective feature-extraction method for micro-expression recognition.

The animation units, obtained by the Kinect v2 device, are also used to extract discriminative features of facial dynamics. To do so, first, we formulated the AnUs of

a frame sequence as a signal image by vertically appending raw facial signals to each other. Given 17 AnUs representing different parts of the face in a video clip and the number of frames with respect to the window considered as l_{w_i}, the size of the signal image for the video clip under consideration is $17 \times l_{w_i}$. Since the video clips might have different lengths, this will cause an inconsistent comparison of the facial dynamics. Therefore, temporal normalization with cubic interpolation is applied to ensure that all the texture images are of the same length, i.e. L_{w_i}. We set L_{wA} and L_{w23}, the length of the sequence with respect to the Window Answer and the Windows 2 + 3, respectively, as 40 and 100, which were determined by averaging the number of frames of the respective windows over all the video clips. To extract the discriminant information of the signal images based on AnUs with the size of $17 \times L_{w_i}$, the signal images are divided into 2×2 subregions, where the uniform Local Binary Pattern (LBP) [4] is applied to each subregion, so as to concatenate them to form a feature vector, i.e. AnU-LBP.

To analyse the effect of each window, the spatiotemporal local features are extracted from different window configurations, e.g. W23 as Windows 2 and 3 etc., by discarding Window 1 because of two reasons: 1) the residual of the facial configuration caused by the feedback often observed in the beginning of the next trial, and 2) the part of the sentence given in Window 1 does not affect the truth condition of the statement.

Table 7-3. Participants' average accuracies and STDs.

Questions	Mean Acc.	Standard dev.
All	0.7023	0.0630
Baseline	0.8900	0.0845
Range	0.5145	0.0751

7.5. Experiments and Discussion

Each participant was confronted with 100 true/false statements. Table 7-3 shows their average accuracies and standard deviations with respect to all statements, Baseline and Range. As tested and observed in Amazon Mechanical Turk, Baseline statements achieved higher accuracies and Range statements were answered with a correction rate of 51% only by the participants of the experiment. One reason for Baseline statements not reaching one hundred percent accuracy unlike our first goal, is the participants' English proficiency, since not all of the participants were native English speakers – the mean of self-rated English proficiency being 7.61.

7.5.1. Experiments on ERIs

Before the behavioral experiment, the participants were not clarified whether they were allowed to read the sentences aloud or how they could react to a correct or an incorrect outcome since we did not want to pressure them with any restriction. At the end, there were two noticeable groups of trials: 1) the trials with no visible mouth movement caused by talking, and 2) the trials with visible mouth movement caused by talking. Since the effects of mouth movement caused by talking would be the concern of another study, we conducted our experiments, especially the ones with the ERIs, in two groups: 1) trials without talking, and 2) all trials. The AnUs are also divided into two groups, Lower Face and Upper Face, to explore the dynamics of the upper and lower face behaviors, with and without talking.

After data preprocessing, three sets of variables are specified: 1) the facial signals, i.e. ERIs, as outcome variables, 2) the labels, such as category, veracity of the statements and feedback as predictor variables, and 3) the subjects and the stimuli as random variables. To investigate the correlation between the outcome and the predictor variables, the Linear Mixed Effects (LME) model is adapted. LME is a

popular method that allows researchers to investigate the effects and the interaction of the predictor variables, i.e. fixed effects, on a particular outcome variable, while controlling the interclass correlation of the random effects [27, 163]. The lme4 package [14] in the R environment is used in our experiments. After the model is learned, a *t*-test is applied to determine if the predictor variables are statistically significant for explaining the outcome variable. In other words, a *t*-test is utilized to measure how likely the experiment is to be repeatable or what is the probability of the results happening by chance. In hypothesis testing the *t*-score or *t*-value is the ratio between the difference between the means of two groups in a fixed variable and the standard deviations within the groups as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{{s_1}^2/_{n_1} + {s_2}^2/_{n_2}}} \tag{7.3}$$

where $\bar{x}_i$, s_1, and n_1 are the mean value, standard deviation and the number of the samples belonging the ith group, respectively.

For the sake of brevity, we only present the *t*-values and the *p*-values obtained in the *t*-test. In terms of the two measurements, the higher the absolute value of a *t*-value, or the lower *p*-value, and the greater the evidence against the null hypothesis, i.e. there is greater difference between their respective distributions.

Table 7-4 presents the *t*-values of the predictor variables on the facial signals of the upper and lower faces using the two groups of trials, i.e. trials without talking and all trials. Indeed, we found a significant effect of the feedback variable. It is expected because of the fact that the facial signals, fed to the LME model, are the signals obtained from the feedback window, and their facial behaviour would be mostly affected by the feedback they received. However, what is interesting to find is the fact that the difficulty of the questions has no significant impact on participants' facial behaviours except lip stretching and cheek puff, unlike our first assumption. It

suggests that the participants are likely to react similarly, if they answer a question incorrectly, regardless of the difficulty of the questions.

We can also observe that the facial signals of the upper face are more correlated with the feedback. This aligns with the Buck's proposition, which states that the eyes are more likely to produce spontaneous affects on the face than the lower face [20].

Table 7-5 presents the *t* values of the predictor variables on each facial signal separately. As observed in Table 7-5, parts of the face defined by the AnUs have varying correlations with the feedback they received. Also, the highest correlation that we observe is "eye closed", which refers to the eyes' openness.

Table 7-4. Statistical results on ERIs of face and head based on trial categories.

Sample type	AnU	Variable	*t* value	*p* value	Signif.
Trials without talking	AnU Upper Face	(Intercept)	2.664	0.00895	**
		value - true	-0.327	0.74434	
		category - range	1.265	0.20879	
		feedback - incorrect	**-11.908**	**< 2e-16**	***
	AnU Lower Face	(Intercept)	-1.772	0.0805	.
		value - true	-0.179	0.8586	
		category - range	-1.395	0.1662	
		feedback - incorrect	0.989	0.3228	
	Head Movements – Pitch, Yaw, Roll	(Intercept)	-0.698	0.487	
		value - true	0.6	0.55	
		category - range	-1.613	0.11	
		feedback - incorrect	**7.635**	**2.26E-14**	***
	Rigid Face Shape Parameters+	(Intercept)	-1.095	0.277	
		value - true	-0.162	0.871	
		category - range	0.891	0.375	
		feedback - incorrect	**-4.789**	**1.68e-06**	***
All trials	AnU Upper Face	(Intercept)	2.543	0.0125	*
		value - true	-0.544	0.5875	
		category - range	0.356	0.7226	
		feedback - incorrect	**-14.386**	**<2e-16**	***
	AnU Lower Face	**(Intercept)**	-3.546	0.000567	***
		value - true	0.98	0.32957	
		category - range	-0.272	0.786211	
		feedback - incorrect	0.064	0.948958	
	Head Movements – Pitch, Yaw, Roll	(Intercept)	0.392	0.6958	
		value - true	-0.259	0.7958	
		category - range	1.366	0.1749	
		feedback - incorrect	**-3.023**	**0.0025**	**
	Rigid Face Shape Parameters+	(Intercept)	-2.702	0.00783	**
		value - true	2.2	0.03017	*
		category - range	1.344	0.182	
		feedback - incorrect	-0.475	0.63474	

Table 7-5. Statistical results on ERIs of face based on AnUs.

AnU	Variable	*t* value	*p* value	Signif.
AnU1 Jaw Open	(Intercept)	0.102	0.919	
	value - true	-1.294	0.199	
	category - range	-1.641	0.104	
	feedback - incorrect	1.552	0.121	
AnU2 Lip Pucker	(Intercept)	-2.114	0.0382	*
	value - true	-0.669	0.5051	
	category - range	-1.77	0.0799	.
	feedback - incorrect	**4.61**	**4.04E-06**	***
AnU3 Jaw Slide Right	(Intercept)	-0.343	0.7322	
	value - true	-2.255	0.0264	*
	category - range	0.356	0.7225	
	feedback - incorrect	**-3.945**	**7.99E-05**	***
AnU4 - AnU5 Lip Stretcher	(Intercept)	1.046	0.2978	
	value - true	-1.476	0.1433	
	category - range	-2.541	0.0126	*
	feedback - incorrect	**4.948**	**7.52E-07**	***
AnU6 - AnU7 Lip Corner Puller	(Intercept)	-0.151	0.88	
	value - true	0.629	0.531	
	category - range	-1.614	0.11	
	feedback - incorrect	0.528	0.597	
AnU8 - AnU9 Lip Corner Depressor	(Intercept)	-2.551	0.013	*
	value - true	0.768	0.444	
	category - range	0.223	0.824	
	feedback - incorrect	**-3.914**	**9.09E-05**	***
AnU10 - AnU11 Cheek Puff	(Intercept)	-2.215	0.03115	*
	value - true	0.956	0.34146	
	category - range	2.994	0.00348	**
	feedback - incorrect	**5.117**	**3.11E-07**	***
AnU12 - AnU13 Eye Closed	**(Intercept)**	3.779	0.000361	***
	value - true	-0.096	0.923783	
	category - range	0.51	0.610892	
	feedback - incorrect	**-12.801**	**< 2e-16**	***
AnU14 - AnU15 Eyebrow Lowerer	(Intercept)	-1.476	0.1431	
	value - true	-0.504	0.6156	
	category - range	0.893	0.3741	
	feedback - incorrect	-1.748	0.0805	.
AnU16 - AnU17 Lower-lip Depressor	(Intercept)	-0.295	0.768	
	value - true	0.355	0.724	
	category - range	0.36	0.719	
	feedback - incorrect	1.48	0.139	

7.5.2. Experiments on sentence comprehension

In the behavioral experiment, we assume that all the trials belonging to the Baseline class, and answered correctly by the participants, represent the class “knowing face”. Similarly, all the trials belonging to the Range class should be part of the class

"guessing face". The stimuli representing false world-knowledge in the Baseline are also discarded to prevent the facial dynamic caused by the veracity of the statements.

In the first set of experiments, the features extracted from all trials across all participants are used to learn the feature subspace of knowing face and guessing face, using Marginal Fisher Analysis (MFA) [242]. MFA, which is a general framework for manifold learning and dimensionality reduction, constructs two adjacency graphs to represent the within-class and the between-class geometry of the data and uses the Fisher criterion.

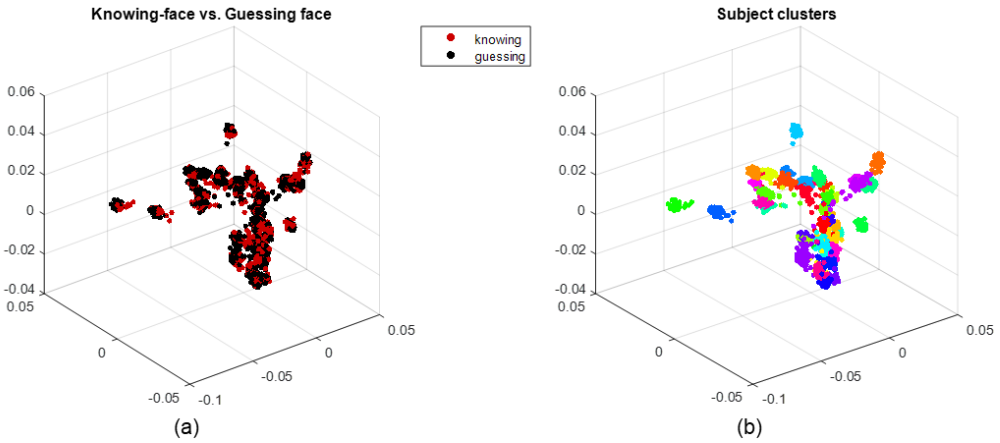

Fig. 7-4. First three dimensions of new subspace learned by the label "guessing face" and "knowing face". (a) plotted using the labels as guessing face and knowing face, and (b) plotted using subject labels.

As we could observe in Fig. 7-4, although the new subspace is learned based on the knowing-face and guessing-face labels, the subspace learned is mostly clustered based on participants. Fig. 7-4(a) represents the new feature space learned by labels as knowing face and guessing face and Fig. 7-4(b) shows the trials with participant labels. From this graph, we can deduce that, the facial dynamics are indeed more person-specific than universal. It is also because of the fact that the participants were from 17 different countries with different cultural background and language.

Table 7-6. Results of online sentence comprehension.

AUC (%)		AnU-LBP	LPQ-TOP	AnU-LBP & LPQ-TOP
Knowing face vs. Guessing face	**w23**	53.65	55.37	**60.05**
	wA	55.73	63.34	**64.79**
True vs. False	**w23**	40.75	51.79	**52.34**
	wA	61.91	80.18	**80.95**

Because of the fact that facial dynamics in our study are not generalizable across participants, a leave-one-trial-out (LOTO) classification scheme is adopted, per person, to investigate the predictability of a knowing face and the veracity of the statements. LOTO scheme trains each fold using the $n_s - 1$ trials belonging to one person and tests the model with one trial, where n_s is the number of trials belonging to the *s*-th participant. Furthermore, the Support Vector Machine (SVM) classifier is adopted since there is a limited number of trials per participant.

The first two rows in Table 7-6 shows the AUCs (area under the curves) of the first approach of online-sentence comprehension. As observed in **Table 7-6**, the highest precision is obtained by using features from both the LPQ-TOP and the AnU-LBP during the Window Answer.

In the second set of experiments, all trials belonging to the Baseline class and answered correctly by the participants are further divided into two classes: representing the veracity of the statements as "true" and "false". Here, we aim to detect the statements with world-knowledge violations using facial dynamics.

As observed in the last two rows of **Table 7-6**, the AUCs of the second set of experiments are much higher than the first set of experiments. This suggests that we can detect the world-knowledge violation in a statement better than detecting a knowing face. Furthermore, the best AUCs are obtained by both the LPQ-TOP and

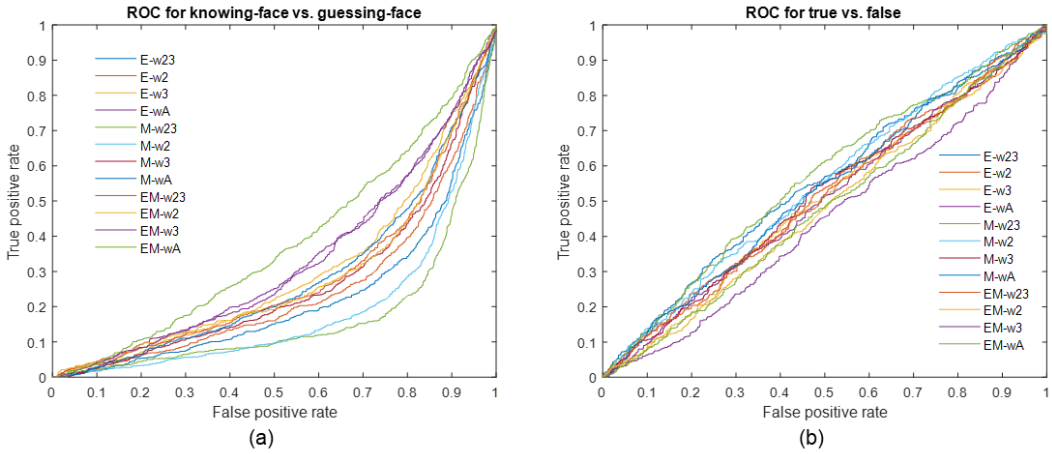

Fig. 7-5. ROC for the results on online sentence comprehension.

the AnU-LBP features extracted from the Window Answer, which suggests that the Window Answer carries more facial cues with respect to online-sentence comprehension, and the combination of LPQ-TOP and AnU-LBP can achieve better performance than either of them alone.

Fig. 7-5 shows the ROCs of the experiments that were conducted, using the features extracted during different stages of online-sentence comprehension.

It is possible that the LPQ-TOP descriptor and the animation units obtained by the Kinect v2 device cannot catch all the changes, especially the micro-expression, that is important to detect guessing faces and world-knowledge violations. A possible future work is to enhance the algorithm to become more sensitive to micro-facial dynamics.

7.6. Conclusion

Researchers in different disciplines, such as computational linguistics and computational neuroscience, are often not aware of the advances in recognizing facial information. This disconnection between disciplines limits the interdisciplinary multimodal studies to understand human facial behavior. This study motivates and fosters the interdisciplinary study by attempting to bring together studies, results and

questions from different disciplines by focusing on the computational analysis of human behavior in an experimental setting, specifically facial behavior which can practically provide methodological support to investigate people's facial behavior and mental states.

We collected 100 general knowledge questions from a wide variety of topics, including mathematics, history, sports, art, etc. A total of 44 participants joined our behavioral experiment, where they were asked to answer the collected questions as true or false in front of a computer, while the Kinect v2 device was recording their faces. After each response, the participants were shown their achievements as correct or incorrect. Two stages of the behavioral experiment gave us the chance to investigate physical behavior induced by both a cognitive process and an emotional state.

From the videos obtained during the behavioral experiment, a new facial expression database was collected. A new concept, namely event-related intensities (ERI), has been proposed, which is later utilized to analyze the event related changes in facial configuration. Then, several classifiers are trained to predict the state of online sentence comprehension.

The results show that it is a new promising area to understand facial behavior linked with the cognitive and emotional processes and can give a chance to answer questions that are asked in different disciplines.

Chapter 8. Conclusion and Future Works

In this thesis, the concepts and the development of facial expression recognition (FER) is first introduced along with the existing works in several aspects of a FER system, which serve as a foundation of the works presented. Our research focuses on four areas: facial feature extraction, feature fusion, dimensionality reduction, and comprehension recognition through facial expressions, where we have given a brief review on the well-known algorithms in these areas.

A systematic review and analysis of current histogram-based local feature descriptors, which have been applied for FER, are provided in Chapter 3. The weaknesses and strengths of the existing descriptors, as well as their underlying connections, have also been discussed and analysed. Then, a comprehensive evaluation of the performances of different descriptors for facial-expression recognition has been conducted and presented. The robustness of the respective local descriptors is tested under different conditions, such as varying image resolutions and number of sub-regions, and the classifiers. The highest classification accuracies are obtained mostly by LGBPHS and LPQ. This shows that Gabor wavelets and phase information are important features for representing expression-specific information. Deep neural-network-based methods indeed can achieve excellent classification accuracies on FER. However, both deep methods and LGBPHS suffer from time and space complexities. According to the comprehensive analysis shown in this research, the best local descriptor for FER, by considering the feature length, computational cost, and the classification accuracy simultaneously, is LPQ.

We propose a new approach for FER by fusing the features extracted from the eye and the mouth windows of a face using LPQ and PHOG, which can achieve good performance in FER. The respective features from the two windows are fused by

projecting them into a coherent subspace, which is used in the second level of classification, where the features from the eye and mouth windows have their correlation maximized. Based on the training feature vectors, the SVM is employed to learn a binary classifier for each of the emotions. Experiment results have shown that our method, with the LPQ feature, outperforms the PHOG feature; and that our method can achieve greater accuracy than other, state-of-the-art FER methods.

We have proposed a classification method, which utilizes the adaptive descriptor selection algorithm to further increase the performance of a facial expression recognition system. The adaptive descriptor selection algorithm determines the best two descriptors for each expression class, out of four commonly-used descriptors such as LBP, LPQ, PHOG and WLD, so as to fuse them by DCC to obtain more discriminant features for the considering expression class. In our experiments, four expression classes are considered for evaluating the performance of the proposed classification method. The LS-SVM is employed based on the features projected to a coherent subspace to learn a binary classifier for each of the expression classes. Experiment results have shown that the proposed classification method can achieve higher recognition rate than any of the individual descriptors.

When dealing with the dimensionality reduction problem in Chapter 6, we have proposed a subspace learning method, named Soft Locality Preserving Map (SLPM), which uses the neighbourhood and class information to construct a projection matrix. The main contribution of the proposed method is adding a parameter β to control the within-class spread, so that the overfitting problem can be solved. Also, we have proposed using low-intensity expression images to learn a better manifold for each expression class. In case of the limited number of low-intensity expression images, two methods of generating new low-intensity features in the subspace are proposed.

Our experiment results have shown that SLPM outperforms the other subspace-learning methods, and is a good alternative to performing dimensionality reduction on high-dimensional datasets. Our experiment results, also, have shown that the proposed feature-generation method can further increase the recognition rates. Moreover, as observed in Table 6-5 and Table 6-6, the nearest neighbour classifier following manifold learning methods, such as SLPM, often achieves higher recognition rates than the SVM classifier. The reason behind this is due to the fact that SLPM causes the expression manifolds to be well separated. This results in overfitting for the SVM classifier, while the same amount of separation helps the nearest neighbour classifier to work more effectively.

Researchers in different disciplines, such as computational linguistics and computational neuroscience, are often not aware of the advances in recognizing facial information. This disconnection between disciplines limits the interdisciplinary multimodal studies to understand human facial behavior. With the study presented in Chapter 7, we aim to motivate and foster the interdisciplinary study by attempting to bring together studies, results and questions from different disciplines by focusing on the computational analysis of human behavior in an experimental setting, specifically facial behavior which can practically provide methodological support to investigate people's facial behavior and mental states. Two stages of the behavioral experiment gave us the chance to investigate physical behavior induced by both a cognitive process and an emotional state. From the videos obtained during the behavioral experiment, a new facial expression database is collected. A new concept, namely event-related intensities (ERI), is proposed, which is later utilized to analyze the event related changes in facial configuration. The results show that it is a new promising

area to understand facial behavior linked with the cognitive and emotional processes and can give a chance to answer questions that are asked in different disciplines.

With the increasing popularity and the success of deep neural-network-based methods, there has been a shift in FER into adopting deep neural-network-based approaches to any part of a FER system that varies from feature extraction and dimensionality reduction to feature fusion. The success of deep neural-network-based methods has empowered us to look beyond the basic questions of FER and to adopt these methods to understand human facial behavior in the wild. Thus, our future direction focuses mainly on the improvement of the study presented in Chapter 7, i.e. the FEC study.

There are several shortcomings of the FEC study, such as its lack of other streams of data that can give discriminative information about underlying cognitive processes, such as eye movement and brain signals, and its lack of homogenous groups of native language speakers. In the future, we aim to expand the current research of the FEC study in three ways: 1) adding an additional data stream, through either eye tracking or EEG, 2) utilizing garden path statements, i.e. a statement that leads participants to certain assumptions, then either fulfils their expectations, or violates them with a word or phrase that is incorrect based on world-knowledge, semantic understanding of the expresses context, or the syntactic plan of the phrase, and 3) conducting the experiment with an homogenous group of native speakers.

Reference

[1] (2011). *LS-SVMlab: a MATLAB/C toolbox for Least Squares Support Vector Machines*. Available: https://www.esat.kuleuven.be/sista/lssvmlab/

[2] B. Abboud, F. Davoine, and M. Dang, "Facial expression recognition and synthesis based on an appearance model," *Signal processing: image communication,* vol. 19, no. 8, pp. 723-740, 2004.

[3] F. Ahmed, "Gradient directional pattern: a robust feature descriptor for facial expression recognition," *Electronics letters,* vol. 48, no. 19, pp. 1203-1204, 2012.

[4] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 28, no. 12, pp. 2037-2041, 2006.

[5] T. Ahsan, T. Jabid, and U.-P. Chong, "Facial expression recognition using local transitional pattern on Gabor filtered facial images," *IETE Technical Review,* vol. 30, no. 1, pp. 47-52, 2013.

[6] P. D. Allison, *Logistic regression using SAS: Theory and application*. SAS Institute, 2012.

[7] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 356-361: IEEE.

[8] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological science,* vol. 16, no. 5, pp. 403-410, 2005.

[9] T. Armstrong and B. O. Olatunji, "Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis," *Clinical psychology review,* vol. 32, no. 8, pp. 704-723, 2012.

[10] A. B. Ashraf *et al.*, "The painful face–pain expression recognition using active appearance models," *Image and vision computing,* vol. 27, no. 12, pp. 1788-1796, 2009.

[11] V. Balntas, L. Tang, and K. Mikolajczyk, "Binary online learned descriptors," *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 3, pp. 555-567, 2018.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-10: IEEE.

[13] F. Bashar, A. Khan, F. Ahmed, and M. H. Kabir, "Robust facial expression recognition based on median ternary pattern (MTP)," in *2013 International*

Conference on Electrical Information and Communication Technology (EICT), 2014, pp. 1-5: IEEE.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *NIPS*, 2001, vol. 14, pp. 585-591.

[17] G. Benitez-Garcia, T. Nakamura, and M. Kaneko, "Facial Expression Recognition Based on Local Fourier Coefficients and Facial Fourier Descriptors," *Journal of Signal and Information Processing*, vol. 8, no. 03, p. 132, 2017.

[18] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 401-408: ACM.

[19] K. Brady *et al.*, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 97-104: ACM.

[20] R. Buck, *The communication of emotion*. guilford press, 1984.

[21] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[22] R. F. Burton, "Misinformation, partial knowledge and guessing in true/false tests," *Medical Education*, vol. 36, no. 9, pp. 805-811, 2002.

[23] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075-1086, 2007.

[24] D. Cai and X. He, "Orthogonal locality preserving indexing," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 3-10: ACM.

[25] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality Sensitive Discriminant Analysis," in *IJCAI*, 2007, pp. 708-713.

[26] J. Cao, H. Wang, P. Hu, and J. Miao, "PAD model based facial expression analysis," in *International Symposium on Visual Computing*, 2008, pp. 450-459: Springer.

[27] V. Carey and Y.-G. Wang, "Mixed-effects models in S and S-PLUS," ed: Taylor & Francis, 2001.

[28] S. Chakraborty, S. Singh, and P. Chakraborty, "Local Gradient Hexa Pattern: A Descriptor for Face Recognition and Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[30] Y. Chang, C. Hu, and M. Turk, "Manifold of facial expression," in *AMFG*, 2003, pp. 28-35.

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

[32] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 508-513: ACM.

[33] J. Chen *et al.*, "WLD: A robust local image descriptor," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1705-1720, 2010.

[34] L.-F. Chen and Y.-S. Yen, "Taiwanese facial expression image database," *Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan*, 2007.

[35] W.-P. Choi, S.-H. Tse, K.-W. Wong, and K.-M. Lam, "Simplified Gabor wavelets for human face recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 1186-1199, 2008.

[36] W. Chu, "Facial expression recognition based on local binary pattern and gradient directional pattern," in *), IEEE International Conference on Green Computing and Communications (GreenCom), 2013 IEEE Internet of Things and IEEE Cyber, Physical and Social Computing (iThings/CPSCom)*, 2013, pp. 1458-1462: IEEE.

[37] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII'09)*, 2009, pp. 1-7: IEEE.

[38] C. Conati and X. Zhou, "Modeling students' emotions from cognitive appraisal in educational games," in *International Conference on Intelligent Tutoring Systems*, 2002, pp. 944-954: Springer.

[39] A. Constantinidis, M. C. Fairhurst, and A. F. R. Rahman, "A new multi-expert decision combination algorithm and its application to the detection of circumscribed masses in digital mammograms," *Pattern Recognition*, vol. 34, no. 8, pp. 1527-1537, 2001.

[40] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* no. 6, pp. 681-685, 2001.

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

[42] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory,* vol. 13, no. 1, pp. 21-27, 1967.

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

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

[45] A. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing,* 2016.

[46] F. De la Torre, J. Campoy, Z. Ambadar, and J. F. Cohn, "Temporal segmentation of facial behavior," in *IEEE 11th International Conference on Computer Vision (ICCV 2007),* 2007, pp. 1-8: IEEE.

[47] V. Demberg and F. Keller, "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity," *Cognition,* vol. 109, no. 2, pp. 193-210, 2008.

[48] A. Dhall, "Collecting large, richly annotated facial-expression databases from movies," 2012.

[49] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* 2011, pp. 878-883: IEEE.

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

[51] H. Dibeklioglu and T. Gevers, "Automatic Estimation of Taste Liking through Facial Expression Dynamics," *IEEE Transactions on Affective Computing,* 2018.

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

[53] W. Ding *et al.*, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 506-513: ACM.

[54] N. P. Doshi and G. Schaefer, "A comprehensive benchmark of local binary pattern algorithms for texture retrieval," in *21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 2760-2763: IEEE.

[55] E. Douglas-Cowie *et al.*, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," *Affective computing and intelligent interaction*, pp. 488-500, 2007.

[56] S. Du, Y. Yan, and Y. Ma, "Local spiking pattern and its application to rotation-and illumination-invariant texture classification," *Optik-International Journal for Light and Electron Optics*, vol. 127, no. 16, pp. 6583-6589, 2016.

[57] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-aware local binary feature learning for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[58] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Learning rotation-invariant local binary descriptor," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3636-3651, 2017.

[59] R. L. Ebel, "The case for true-false test items," *The School Review*, vol. 78, no. 3, pp. 373-389, 1970.

[60] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 467-474: ACM.

[61] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169-200, 1992.

[62] P. Ekman and W. Friesen, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto: Consulting Psychologists*, 1978.

[63] P. Ekman and K. G. Heider, "The universality of a contempt expression: A replication," *Motivation and emotion*, vol. 12, no. 3, pp. 303-308, 1988.

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

[65] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[66] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*: Springer, 2005, pp. 181-200.

[67] C. E. Erdem, C. Turan, and Z. Aydin, "BAUM-2: a multilingual audio-visual affective face database," *Multimedia Tools and Applications,* vol. 74, no. 18, pp. 7429-7459, 2015.

[68] S. Fadaei, R. Amirfattahi, and M. R. Ahmadzadeh, "Local derivative radial patterns: A new texture descriptor for content-based image retrieval," *Signal Processing,* vol. 137, pp. 274-286, 2017.

[69] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke, "Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks," *Electroencephalography and clinical neurophysiology,* vol. 78, no. 6, pp. 447-455, 1991.

[70] K.-C. Fan and T.-Y. Hung, "A novel local pattern descriptor—local vector pattern in high-order derivative space for face recognition," *IEEE transactions on image processing,* vol. 23, no. 7, pp. 2877-2891, 2014.

[71] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 445-450: ACM.

[72] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing,* vol. 49, no. 12, pp. 3136-3144, 2001.

[73] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics,* vol. 7, no. 2, pp. 179-188, 1936.

[74] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems,* vol. 42, no. 3, pp. 143-166, 2003.

[75] M. F. Fritz, "Guessing in a true-false test," *Journal of Educational Psychology,* vol. 18, no. 8, p. 558, 1927.

[76] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in *IJCAI,* 2005, vol. 5, pp. 1048-1053.

[77] P. D. Gader and M. A. Khabou, "Automatic feature generation for handwritten digit recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, no. 12, pp. 1256-1261, 1996.

[78] L. Gang, Z. Yong, L. Yan-Lei, and D. Jing, "Three dimensional canonical correlation analysis and its application to facial expression recognition," in *Intelligent Computing and Information Science*: Springer, 2011, pp. 56-61.

[79] S. M. Garnsey, M. K. Tanenhaus, and R. M. Chapman, "Evoked potentials and the study of sentence comprehension," *Journal of psycholinguistic research,* vol. 18, no. 1, pp. 51-60, 1989.

[80] W. Groen *et al.*, "Semantic, factual, and social language comprehension in adolescents with autism: an FMRI study," *Cerebral Cortex,* vol. 20, no. 8, pp. 1937-1945, 2009.

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

[82] S. Guo and Q. Ruan, "Facial expression recognition using local binary covariance matrices," in *4th IET International Conference on Wireless, Mobile & Multimedia Networks (ICWMMN 2011),* 2011, pp. 237-242: IET.

[83] A. Gupta, R. Jaiswal, S. Adhikari, and V. Balasubramanian, "DAISEE: Dataset for Affective States in E-Learning Environments," *arXiv,* pp. 1-22, 2016.

[84] P. Hagoort, L. Hald, M. Bastiaansen, and K. M. Petersson, "Integration of word meaning and world knowledge in language comprehension," *science,* vol. 304, no. 5669, pp. 438-441, 2004.

[85] Z. Hammal, J. F. Cohn, C. Heike, and M. L. Speltz, "What can head and facial movements convey about positive and negative affect?," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII),* 2015, pp. 281-287: IEEE.

[86] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head movement dynamics during play and perturbed mother-infant interaction," *IEEE transactions on affective computing,* vol. 6, no. 4, pp. 361-370, 2015.

[87] S. Harris, S. A. Sheth, and M. S. Cohen, "Functional neuroimaging of belief, disbelief, and uncertainty," *Annals of neurology,* vol. 63, no. 2, pp. 141-147, 2008.

[88] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2016, pp. 770-778.

[89] H. Hill and A. Johnston, "Categorizing sex and identity from the biological motion of faces," *Current biology,* vol. 11, no. 11, pp. 880-885, 2001.

[90] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology,* vol. 24, no. 6, p. 417, 1933.

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

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

[93] Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE transactions on pattern analysis and machine intelligence,* vol. 17, no. 1, pp. 90-94, 1995.

[94] S. Z. Ishraque, A. H. Banna, and O. Chae, "Local Gabor directional pattern for facial expression recognition," in *2012 15th International Conference on Computer and Information Technology (ICCIT)*, 2012, pp. 164-167: IEEE.

[95] M. S. Islam, "Gender Classification using Gradient Direction Pattern," *Science International*, vol. 25, no. 4, 2013.

[96] M. S. Islam, "Local gradient pattern-A novel feature representation for facial expression recognition," *Journal of AI and Data Mining*, vol. 2, no. 1, pp. 33-38, 2014.

[97] M. S. Islam and S. Auwatanamongkol, "Facial Expression Recognition using Local Arc Pattern," *Trends in Applied Sciences Research*, vol. 9, no. 2, p. 113, 2014.

[98] T. Jabid and O. Chae, "Local Transitional Pattern: A Robust Facial Image Descriptor for Automatic Facial Expression Recognition," in *Proc. International Conference on Computer Convergence Technology, Seoul, Korea*, 2011, pp. 333-44.

[99] T. Jabid and O. Chae, "Facial Expression Recognition Based on Local Transitional Pattern," *International Information Institute (Tokyo). Information*, vol. 15, no. 5, p. 2007, 2012.

[100] T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (LDP)–A robust image descriptor for object recognition," in *2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 482-487: IEEE.

[101] R. E. Jack, O. G. Garrod, and P. G. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," *Current biology*, vol. 24, no. 2, pp. 187-192, 2014.

[102] R. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," *Mind*, vol. 2, no. 3, p. 4, 2003.

[103] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-8: IEEE.

[104] S. Jaiswal, M. Valstar, A. Gillott, and D. Daley, "Automatic detection of ADHD and ASD from expressive behaviour in RGBD data," *arXiv preprint arXiv:1612.02374*, 2016.

[105] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley, "Automatic detection of ADHD and ASD from expressive behaviour in RGBD data," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 762-769: IEEE.

[106] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard, "Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions

and Body Language," in *International Conference on Intelligent Virtual Agents*, 2016, pp. 64-74: Springer.

[107] J. Jia, Z. Wu, S. Zhang, H. M. Meng, and L. Cai, "Head and facial gestures synthesis using PAD model for an expressive talking avatar," *Multimedia Tools and Applications,* vol. 73, no. 1, pp. 439-461, 2014.

[108] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* 2011, pp. 314-321: IEEE.

[109] X.-Y. Jing, D. Zhang, and J.-Y. Yang, "Face recognition based on a group decision-making combination approach," *Pattern Recognition,* vol. 36, no. 7, pp. 1675-1678, 2003.

[110] R. Johnson, A. Pfefferbaum, and B. S. Kopell, "P300 and long-term memory: Latency predicts recognition performance," *Psychophysiology,* vol. 22, no. 5, pp. 497-507, 1985.

[111] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, "Deep temporal appearance-geometry network for facial expression recognition," *arXiv preprint arXiv:1503.01532,* 2015.

[112] M. A. Just, P. A. Carpenter, T. A. Keller, W. F. Eddy, and K. R. Thulborn, "Brain activation modulated by sentence comprehension," *Science,* vol. 274, no. 5284, pp. 114-116, 1996.

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

[114] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," *Advances in visual computing,* pp. 368-377, 2012.

[115] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 10, pp. 1925-1937, 2011.

[116] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceesings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition,* 2000, pp. 46-53: IEEE.

[117] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images," *Pattern Recognition Letters,* vol. 34, no. 10, pp. 1159-1168, 2013.

[118] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces,* vol. 10, no. 2, pp. 173-189, 2016.

[119] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1005-1018, 2007.

[120] S. J. Kirsh and J. R. Mounts, "Violent video game play impacts facial emotion recognition," *Aggressive behavior,* vol. 33, no. 4, pp. 353-358, 2007.

[121] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[122] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion," in *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, 2001, pp. 43-46: IEEE.

[123] I. Kotsia, S. Zafeiriou, N. Nikolaidis, and I. Pitas, "Texture and shape information fusion for facial action unit recognition," in *Advances in Computer-Human Interaction, 2008 First International Conference on*, 2008, pp. 77-82: IEEE.

[124] R. L. Kristensen, Z.-H. Tan, Z. Ma, and J. Guo, "Binary pattern flavored feature extractors for Facial Expression Recognition: An overview," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* 2015, pp. 1131-1137: IEEE.

[125] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[126] J. D. Kropotov, *Functional Neuromarkers for Psychiatry: Applications for Diagnosis and Treatment*. Academic Press, 2016.

[127] H.-W. Kung, Y.-H. Tu, and C.-T. Hsu, "Dual subspace nonnegative graph embedding for identity-independent expression recognition," *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 3, pp. 626-639, 2015.

[128] G. R. Kuperberg, T. Sitnikova, and B. M. Lakshmanan, "Neuroanatomical distinctions within the semantic system during sentence comprehension: evidence from functional magnetic resonance imaging," *Neuroimage,* vol. 40, no. 1, pp. 367-388, 2008.

[129] Z. Lei, T. Ahonen, M. Pietikäinen, and S. Z. Li, "Local frequency descriptor for low-resolution face recognition," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* 2011, pp. 161-166: IEEE.

[130] B. Li, D.-S. Huang, C. Wang, and K.-H. Liu, "Feature extraction using constrained maximum variance mapping," *Pattern Recognition,* vol. 41, no. 11, pp. 3287-3294, 2008.

[131] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks,* vol. 17, no. 1, pp. 157-165, 2006.

[132] J. Li, N. Sang, and C. Gao, "Face recognition with Riesz binary pattern," *Digital Signal Processing,* vol. 51, pp. 196-201, 2016.

[133] S. Li, D. Gong, and Y. Yuan, "Face recognition using Weber local descriptors," *Neurocomputing,* vol. 122, pp. 272-283, 2013.

[134] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013),* 2013, pp. 1-6: IEEE.

[135] Z. Li, J.-i. Imai, and M. Kaneko, "Face and expression recognition based on bag of words method considering holistic and local image features," in *2010 International Symposium on Communications and Information Technologies (ISCIT),* 2010, pp. 1-6: IEEE.

[136] J.-J. J. Lien, *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity*. University of Pittsburgh, 1998.

[137] C. Liu, "A Bayesian discriminating features method for face detection," *IEEE transactions on pattern analysis and machine intelligence,* vol. 25, no. 6, pp. 725-740, 2003.

[138] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," *Pattern Recognition,* vol. 62, pp. 135-160, 2017.

[139] L. Liu, S. Lao, P. W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Median robust extended local binary pattern for texture classification," *IEEE Transactions on Image Processing,* vol. 25, no. 3, pp. 1368-1381, 2016.

[140] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conference on Computer Vision*, 2014, pp. 143-157: Springer.

[141] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749-1756.

[142] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805-1812.

[143] R. Liu and D. F. Gillies, "Overfitting in linear feature extraction for classification of high-dimensional image data," *Pattern Recognition,* vol. 53, pp. 73-86, 2016.

[144] S. Liu, Y. Tian, C. Peng, and J. Li, "Facial expression recognition approach based on least squares support vector machine with improved particle swarm optimization algorithm," in *2010 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2010, pp. 399-404: IEEE.

[145] S. Liu, Y. Zhang, and K. Liu, "Facial expression recognition under partial occlusion based on Weber Local Descriptor histogram and decision fusion," in *2014 33rd Chinese Control Conference (CCC)*, 2014, pp. 4664-4668: IEEE.

[146] 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 Recognition*, vol. 61, pp. 610-628, 2017.

[147] J. Lu, V. Erin Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for face recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3721-3729.

[148] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5356-5368, 2015.

[149] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[150] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2041-2056, 2015.

[151] J. Lu, Y. Zhao, and J. Hu, "Enhanced Gabor-based region covariance matrices for palmprint recognition," *Electronics letters*, vol. 45, no. 17, pp. 880-881, 2009.

[152] K. Lu and X. Zhang, "Facial expression recognition from image sequences based on feature points and canonical correlations," in *2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI)*, 2010, vol. 1, pp. 219-223: IEEE.

[153] Z. Lubing and W. Han, "Local gradient increasing pattern for facial expression recognition," in *2012 19th IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 2601-2604: IEEE.

[154] P. Lucey *et al.*, "Automatically detecting pain in video through facial action units," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 664-674, 2011.

[155] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200-205: IEEE.

[156] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters,* vol. 70, pp. 80-86, 2016.

[157] J. F. Marques, N. Canessa, and S. Cappa, "Neural differences in the processing of true and false sentences: Insights into the nature of 'truth'in language comprehension," *Cortex,* vol. 45, no. 6, pp. 759-768, 2009.

[158] A. M. Martínez and A. C. Kak, "Pca versus lda," *IEEE transactions on pattern analysis and machine intelligence,* vol. 23, no. 2, pp. 228-233, 2001.

[159] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, 2017.

[160] K. Mase, "Recognition of facial expression from optical flow," *IEICE transactions (E),* vol. 74, pp. 3474-3483, 1991.

[161] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks,* vol. 16, no. 5-6, pp. 555-559, 2003.

[162] G. McCarthy and C. C. Wood, "Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models," *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section,* vol. 62, no. 3, pp. 203-208, 1985.

[163] C. E. McCulloch and J. M. Neuhaus, *Generalized linear mixed models*. Wiley Online Library, 2001.

[164] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in *Proceedings of the Cognitive Science Society*, 2007, vol. 29, no. 29.

[165] D. McDuff, R. El Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 512-518: IEEE.

[166] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing,* vol. 3, no. 1, pp. 5-17, 2012.

[167] A. Mehrabian and J. A. Russell, *An approach to environmental psychology*. the MIT Press, 1974.

[168] T. Mohammad and M. L. Ali, "Robust facial expression recognition based on local monotonic pattern (LMP)," in *2011 14th International Conference on Computer and Information Technology (ICCIT)*, 2011, pp. 572-576: IEEE.

[169] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-10: IEEE.

[170] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *arXiv preprint arXiv:1708.03985,* 2017.

[171] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological science,* vol. 15, no. 2, pp. 133-137, 2004.

[172] K. P. Murphy and S. Russell, "Dynamic bayesian networks: representation, inference and learning," 2002.

[173] L. Nanni, A. Lumini, and S. Brahnam, "Local binary patterns variants as texture descriptors for medical image analysis," *Artificial intelligence in medicine,* vol. 49, no. 2, pp. 117-125, 2010.

[174] L. Nanni, A. Lumini, and S. Brahnam, "Survey on LBP based texture descriptors for image classification," *Expert Systems with Applications,* vol. 39, no. 3, pp. 3634-3641, 2012.

[175] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing,* vol. 2, no. 2, pp. 92-105, 2011.

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

[177] Y.-H. Oh, A. C. Le Ngo, J. See, S.-T. Liong, R. C.-W. Phan, and H.-C. Ling, "Monogenic Riesz wavelet representation for micro-expression recognition," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 1237-1241: IEEE.

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

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

[180] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International conference on image and signal processing*, 2008, pp. 236-243: Springer.

[181] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[182] A. Ortony and T. J. Turner, "What's basic about basic emotions?," *Psychological review,* vol. 97, no. 3, p. 315, 1990.

[183] J. Päivärinta, E. Rahtu, and J. Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Scandinavian Conference on Image Analysis*, 2011, pp. 360-369: Springer.

[184] Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 7, pp. 989-993, 2008.

[185] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3505-3513, 2009.

[186] J. W. Peirce, "PsychoPy—psychophysics software in Python," *Journal of neuroscience methods*, vol. 162, no. 1, pp. 8-13, 2007.

[187] J. W. Peirce, "Generating stimuli for neuroscience using PsychoPy," *Frontiers in neuroinformatics*, vol. 2, p. 10, 2009.

[188] X. Peng, Z. Xia, L. Li, and X. Feng, "Towards facial expression recognition in the wild: a new database and deep recognition system," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 93-99.

[189] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197-219, 1984.

[190] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556-567, 2014.

[191] R. Ptucha and A. Savakis, "Manifold based sparse representation for facial understanding in natural images," *Image and Vision Computing*, vol. 31, no. 5, pp. 365-378, 2013.

[192] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 3, pp. 252-264, 1991.

[193] A. R. Rivera, J. R. Castillo, and O. Chae, "Local directional texture pattern image descriptor," *Pattern Recognition Letters*, vol. 51, pp. 94-100, 2015.

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

[195] P. Rodriguez *et al.*, "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," *IEEE Transactions on Cybernetics*, 2017.

[196] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[197] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science,* vol. 290, no. 5500, pp. 2323-2326, 2000.

[198] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological review,* vol. 110, no. 1, p. 145, 2003.

[199] M. Sathik and S. G. Jonathan, "Effect of facial expressions on student's comprehension recognition in virtual educational environments," *SpringerPlus,* vol. 2, no. 1, p. 455, 2013.

[200] A. Savran *et al.*, "Bosphorus database for 3D face analysis," *Biometrics and identity management,* pp. 47-56, 2008.

[201] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815-823.

[202] C. Shan, S. Gong, and P. W. McOwan, "Appearance manifold of facial expression," in *International Workshop on Human-Computer Interaction*, 2005, pp. 221-230: Springer.

[203] C. Shan, S. Gong, and P. W. McOwan, "A comprehensive empirical study on linear subspace methods for facial expression analysis," in *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 2006, pp. 153-153: IEEE.

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

[205] L. Shen, M. Wang, and R. Shen, "Affective e-learning: Using" emotional" data to improve learning in pervasive learning environment," *Journal of Educational Technology & Society,* vol. 12, no. 2, p. 176, 2009.

[206] S. Shojaeilangari, W.-Y. Yau, J. Li, and E.-K. Teoh, "Feature extraction through binary pattern of phase congruency for facial expression recognition," in *2012 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, 2012, pp. 166-170: IEEE.

[207] S. Shojaeilangari, Y. W. Yun, and T. E. Khwang, "Person independent facial expression analysis using Gabor features and genetic algorithm," in *2011 8th International Conference on Information, Communications and Signal Processing (ICICS)*, 2011, pp. 1-5: IEEE.

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

[209] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing,* no. 1, pp. 17-28, 2016.

[210] T. Song, H. Li, F. Meng, Q. Wu, and J. Cai, "LETRIST: Locally Encoded Transform Feature Histogram for Rotation-Invariant Texture Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[211] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437-2448, 2005.

[212] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476-3483.

[213] S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187-1188, 1965.

[214] M. Suwa, "A preliminary note on pattern recognition of human emotional expression," in *Proc. of The 4th International Joint Conference on Pattern Recognition*, 1978, pp. 408-410.

[215] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293-300, 1999.

[216] V. Takala, T. Ahonen, and M. Pietikäinen, "Block-based methods for image retrieval using local binary patterns," *Image analysis*, pp. 13-181, 2005.

[217] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

[218] Z. Tősér, L. A. Jeni, A. Lőrincz, and J. F. Cohn, "Deep learning for facial action unit detection under large head poses," in *Computer Vision–ECCV 2016 Workshops*, 2016, pp. 359-371: Springer.

[219] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653-1660.

[220] C. Turan and K.-M. Lam, "Region-based feature fusion for facial-expression recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5966-5970: IEEE.

[221] C. Turan and K.-M. Lam, "Histogram-based Local Descriptors for Facial Expression Recognition (FER): A comprehensive Study," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 331-341, 2018.

[222] C. Turan, K.-M. Lam, and X. He, "Soft Locality Preserving Map (SLPM) for Facial Expression Recognition," *arXiv preprint arXiv:1801.03754*, 2018.

[223] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Computer Vision–ECCV 2006*, pp. 589-600, 2006.

[224] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 2006, pp. 149-149: IEEE.

[225] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, p. 65.

[226] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 28-43, 2012.

[227] T. Van Gog and K. Scheiter, "Eye tracking as a tool to study and enhance multimedia learning," ed: Elsevier, 2010.

[228] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988-999, 1999.

[229] M. Verma and B. Raman, "Local tri-directional patterns: A new texture feature descriptor for image retrieval," *Digital Signal Processing*, vol. 51, pp. 62-72, 2016.

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

[231] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," *Human–computer interaction*, pp. 6-18, 2007.

[232] R. Walecki, V. Pavlovic, B. Schuller, and M. Pantic, "Deep Structured Learning for Facial Action Unit Intensity Estimation," *arXiv preprint arXiv:1704.04481*, 2017.

[233] C. Wan, Y. Tian, H. Chen, and S. Liu, "Based on local feature region fusion of facial expression recognition," in *2010 2nd International Conference on Advanced Computer Control (ICACC)*, 2010, vol. 1, pp. 202-206: IEEE.

[234] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection," *IEEE Transactions on Neural Networks*, vol. 19, no. 4, pp. 571-585, 2008.

[235] S. Wang, J. Lu, X. Gu, H. Du, and J. Yang, "Semi-supervised linear discriminant analysis for dimension reduction and classification," *Pattern Recognition*, vol. 57, pp. 179-189, 2016.

[236] Q. Wei, B. Sun, J. He, and L. Yu, "BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels," *Signal Processing: Image Communication*, vol. 59, pp. 168-181, 2017.

[237] W. K. Wong and H. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognition*, vol. 45, no. 1, pp. 186-197, 2012.

[238] X. X. Xia, Z. L. Ying, and W. J. Chu, "Facial Expression Recognition Based on Monogenic Binary Coding," in *Applied Mechanics and Materials*, 2014, vol. 511, pp. 437-440: Trans Tech Publ.

[239] J. Xie, "Face recognition based on Curvelet transform and LS-SVM," in *Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), Huangshan, PR China*, 2009, pp. 140-143.

[240] X. Xu and Z. Mu, "Feature fusion method based on KCCA for ear and profile face based multimodal recognition," in *2007 IEEE International Conference on Automation and Logistics*, 2007, pp. 620-623: IEEE.

[241] M. Xue, W. Liu, and L. Li, "Person-independent facial expression recognition via hierarchical classification," in *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2013, pp. 449-454: IEEE.

[242] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 1, pp. 40-51, 2007.

[243] W.-J. Yan *et al.*, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

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

[245] B.-Q. Yang, T. Zhang, C.-C. Gu, K.-J. Wu, and X.-P. Guan, "A novel face recognition method based on iwld and iwbc," *Multimedia Tools and Applications*, vol. 75, no. 12, p. 6979, 2016.

[246] M. Yang, L. Zhang, S. C.-K. Shiu, and D. Zhang, "Monogenic binary coding: An efficient local feature extraction approach to face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1738-1751, 2012.

[247] M. Yang, L. Zhang, L. Zhang, and D. Zhang, "Monogenic binary pattern (MBP): A novel feature extraction and representation model for face recognition," in *2010 20th International Conference on Pattern Recognition (ICPR'10)*, 2010, pp. 2680-2683: IEEE.

[248] S. Yang and B. Bhanu, "Facial expression recognition using emotion avatar image," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011, pp. 866-871: IEEE.

[249] W. Yang, C. Sun, and L. Zhang, "A multi-manifold discriminant analysis method for image feature extraction," *Pattern Recognition*, vol. 44, no. 8, pp. 1649-1657, 2011.

[250] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *8th IEEE International Conference on Automatic Face & Gesture Recognition (FG'08)*, 2008, pp. 1-6: IEEE.

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

[252] J. Yu and B. Bhanu, "Evolutionary feature synthesis for facial expression recognition," *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1289-1298, 2006.

[253] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, and M. Pantic, "Learning slow features for behaviour analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2840-2847.

[254] L. Zafeiriou, S. Zafeiriou, and M. Pantic, "Deep Analysis of Facial Behavioral Dynamics," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1988-1996: IEEE.

[255] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39-58, 2009.

[256] Z. Zeng, L. Song, Q. Zheng, and Y. Chi, "A new image retrieval model based on monogenic signal representation," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 85-93, 2015.

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

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

[259] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 57-68, 2007.

[260] J. Zhang, J. Yu, J. You, D. Tao, N. Li, and J. Cheng, "Data-driven facial animation via semi-supervised local patch alignment," *Pattern Recognition*, vol. 57, pp. 1-20, 2016.

[261] L. Zhang, L. Zhang, Z. Guo, and D. Zhang, "Monogenic-LBP: A new approach for rotation invariant texture classification," in *2010 17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 2677-2680: IEEE.

[262] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face

representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV 2005)*, 2005, vol. 1, pp. 786-791: IEEE.

[263] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 699-714, 2005.

[264] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438-3446.

[265] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915-928, 2007.

[266] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using volume local binary patterns," in *Dynamical Vision*: Springer, 2007, pp. 165-177.

[267] X. Zhao *et al.*, "Peak-piloted deep network for facial expression recognition," in *European Conference on Computer Vision*, 2016, pp. 425-442: Springer.

[268] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and least squares support vector machines," in *Advances in Electronic Engineering, Communication and Management Vol. 2*: Springer, 2012, pp. 707-712.

[269] Z. Zhao, J. Han, Y. Zhang, and L.-f. Bai, "A New Supervised Manifold Learning Algorithm," in *International Conference on Image and Graphics*, 2015, pp. 240-251: Springer.

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

[271] Z. Zhong, G. Shen, and W. Chen, "Facial Emotion Recognition Using PHOG and a Hierarchical Expression Model," in *2013 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, 2013, pp. 741-746: IEEE.

[272] X. Zhou, W. Zheng, and M. Xin, "Improving CCA via spectral components selection for facial expression recognition," in *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2012, pp. 1696-1699: IEEE.

[273] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2879-2886: IEEE.