OULU 2014
C 509

ACTA UNIVERSITATIS OULUENSIS

C
TECHNICA

Xiaohua Huang

METHODS FOR FACIAL EXPRESSION RECOGNITION WITH APPLICATIONS IN CHALLENGING SITUATIONS

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING,
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING;
UNIVERSITY OF OULU,
INFOTECH OULU

ACTA UNIVERSITATIS OULUENSIS
C Technica 509

XIAOHUA HUANG

METHODS FOR FACIAL EXPRESSION RECOGNITION WITH APPLICATIONS IN CHALLENGING SITUATIONS

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in Auditorium IT116, Linnanmaa, on 11 December 2014, at 12 noon

UNIVERSITY OF OULU, OULU 2014

Acta Univ. Oul. C 509, 2014

Supervised by
Professor Matti Pietikäinen
Associate Professor Guoying Zhao

Reviewed by
Professor Xiaoyi Feng
Doctor Caifeng Shan

ISBN 978-952-62-0655-4 (Paperback)
ISBN 978-952-62-0656-1 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2014

Huang, Xiaohua, Methods for facial expression recognition with applications in challenging situations.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering, Department of Computer Science and Engineering; Infotech Oulu

Acta Univ. Oul. C 509, 2014
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

In recent years, facial expression recognition has become a useful scheme for computers to affectively understand the emotional state of human beings. Facial representation and facial expression recognition under unconstrained environments have been two critical issues for facial expression recognition systems.

This thesis contributes to the research and development of facial expression recognition systems from two aspects: first, feature extraction for facial expression recognition, and second, applications to challenging conditions.

Spatial and temporal feature extraction methods are introduced to provide effective and discriminative features for facial expression recognition. The thesis begins with a spatial feature extraction method. This descriptor exploits magnitude while it improves local quantized pattern using improved vector quantization. It also makes the statistical patterns domain-adaptive and compact. Then, the thesis discusses two spatiotemporal feature extraction methods. The first method uses monogenic signal analysis as a preprocessing stage and extracts spatiotemporal features using local binary pattern. The second method extracts sparse spatiotemporal features using sparse cuboids and spatiotemporal local binary pattern. Both methods increase the discriminative capability of local binary pattern in the temporal domain.

Based on feature extraction methods, three practical conditions, including illumination variations, facial occlusion and pose changes, are studied for the applications of facial expression recognition. First, with near-infrared imaging technique, a discriminative component-based single feature descriptor is proposed to achieve a high degree of robustness and stability to illumination variations. Second, occlusion detection is proposed to dynamically detect the occluded face regions. A novel system is further designed for handling effectively facial occlusion. Lastly, multi-view discriminative neighbor preserving embedding is developed to deal with pose change, which formulates multi-view facial expression recognition as a generalized eigenvalue problem. Experimental results on publicly available databases show that the effectiveness of the proposed approaches for the applications of facial expression recognition.

Keywords: computer vision, facial expression recognition, feature extraction, local binary pattern, machine learning

Huang, Xiaohua, Menetelmiä kasvonilmeiden tunnistukseen ja sovelluksia haastaviin tilanteisiin.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta, Tietotekniikan osasto; Infotech Oulu
Acta Univ. Oul. C 509, 2014
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Kasvonilmeiden tunnistamisesta on viime vuosina tullut tietokoneille hyödyllinen tapa ymmärtää affektiivisesti ihmisen tunnetilaa. Kasvojen esittäminen ja kasvonilmeiden tunnistaminen rajoittamattomissa ympäristöissä ovat olleet kaksi kriittistä ongelmaa kasvonilmeitä tunnistavien järjestelmien kannalta.

Tämä väitöskirjatutkimus myötävaikuttaa kasvonilmeitä tunnistavien järjestelmien tutkimukseen ja kehittymiseen kahdesta näkökulmasta: piirteiden irrottamisesta kasvonilmeiden tunnistamista varten ja kasvonilmeiden tunnistamisesta haastavissa olosuhteissa.

Työssä esitellään spatiaalisia ja temporaalisia piirteenirrotusmenetelmiä, jotka tuottavat tehokkaita ja erottelukykyisiä piirteitä kasvonilmeiden tunnistamiseen. Ensimmäisenä työssä esitellään spatiaalinen piirteenirrotusmenetelmä, joka parantaa paikallisia kvantisoituja piirteitä käyttämällä parannettua vektorikvantisointia. Menetelmä tekee myös tilastollisista malleista monikäyttöisiä ja tiiviitä. Seuraavaksi työssä esitellään kaksi spatiotemporaalista piirteenirrotusmenetelmää. Ensimmäinen näistä käyttää esikäsittelynä monogeenistä signaalianalyysiä ja irrottaa spatiotemporaaliset piirteet paikallisia binäärikuvioita käyttäen. Toinen menetelmä irrottaa harvoja spatiotemporaalisia piirteitä käyttäen harvoja kuusitahokkaita ja spatiotemporaalisia paikallisia binäärikuvioita. Molemmat menetelmät parantavat paikallisten binärikuvioiden erottelukykyä ajallisessa ulottuvuudessa.

Piirteenirrotusmenetelmien pohjalta työssä tutkitaan kasvonilmeiden tunnistusta kolmessa käytännön olosuhteessa, joissa esiintyy vaihtelua valaistuksessa, okkluusiossa ja pään asennossa. Ensiksi ehdotetaan lähi-infrapuna kuvantamista hyödyntävää diskriminatiivistä komponenttipohjaista yhden piirteen kuvausta, jolla saavutetaan korkea suoritusvarmuus valaistuksen vaihtelun suhteen. Toiseksi ehdotetaan menetelmä okkluusion havainnointiin, jolla dynaamisesti havaitaan peittyneet kasvon alueet. Uudenlainen menetelmä on kehitetty käsittelemään kasvojen okkluusio tehokkaasti. Viimeiseksi työssä on kehitetty moninäkymäinen diskriminatiivisen naapuruston säilyttävään upottamiseen pohjautuva menetelmä käsittelemään pään asennon vaihtelut. Menetelmä kuvaa moninäkymäisen kasvonilmeiden tunnistamisen yleistettynä ominaisarvohajotelmana. Kokeelliset tulokset julkisilla tietokannoilla osoittavat tässä työssä ehdotetut menetelmät suorituskykyisiksi kasvonilmeiden tunnistamisessa.

Asiasanat: kasvonilmeiden tunnistaminen, konenäkö, koneoppiminen, LBP-menetelmä, piirteiden ilmaisu

To my parents and wife

Acknowledgements

The research work related to this thesis has been carried out at Center for Machine Vision Research, Department of Computer Science and Engineering, University of Oulu, Finland, during the years 2010-2014.

I would like to express my deepest appreciation and thanks to my supervisors, Prof. Matti Pietikäinen and Associate Prof. Guoying Zhao, for their unreserved support and guidance. They gave me all the freedom to grow as an independent researcher, and at the same time, continuing to provide help when it was needed. Your advice on both research as well as my career have been priceless. This thesis would not have been possible without them.

I would like to give my heartfelt appreciation to my parents, who brought me up with their love and encouragement me to pursue my dream. I would like to express my special appreciation to my beloved wife, Xiuyan, who accompanied me with her love, unlimited patience, understanding, helping encouragement. Without her unconditional support, I would never be able to accomplish this work. I would also like to thank my sister for her faith and support over years.

I would like to acknowledge the co-authors of the papers, Prof. Wenming Zheng, Prof. Stan Li and Dr. Xiaopeng Hong, for their valuable comments and discussions. I want also to thank Dr. Yazhou Liu, Dr. Yinghao Cai, Dr. Jie Chen and Dr. Ziheng Zhou and the planning officers of our group, Hannakaisa Aikio and Hannu Rautio, for providing me much help and advice during my research life. Thanks to members of the Center for Machine Vision Research, past and present, created a wonderful research environment.

The research work has been carried out at Senior Research Fellow Start-up package from Infotech Oulu and project “Computer Vision for Continuous Emotional State Analysis” funded by Academy of Finland. In addition, this doctoral thesis has been financially supported by Infotech Oulu Doctoral Program and the Nokia Foundation. The importance of these grants and funding is gratefully acknowledged.

I would like to gratefully acknowledge the reviewers Prof. Xiaoyi Feng from Northwestern Polytechnical University, China and Dr. Caifeng Shan from Philips Research, Eindhoven, the Netherlands, for their valuable and extensive review comments which helped to improve the final outcome. I would also like to thank Prof. Joni

Kämäräinen from Tampere University of Technology, and Prof. Veikko Surakka from University of Tampere, for serving as the opponents in the defense. I wish also to thank Dr. Pertti Väyrynen for the language revision. With regard to the dissertation publishing, my thanks go to Dr. Vili-Petteri Kellokumpu for helping me with Finnish translations and providing useful information.

Finally, I would like to thank my friends for their support and encouragement.

Oulu, December 2014

List of symbols

$\vec{x}$	*Vector*
$\mathbf{X}$	*Matrix*
I	*An image*
$\Re^d$	*The d-dimensional feature space*
χ^2	*Chi-Square distance of histogram*
$L(x)$	*A logistic function with* $L(x) = 1$ *if x is true and* $L(x) = 0$ *otherwise*
C	*Number of classes*
P	*Number of neighbors*
R	*Radius of neighborhood*
T	*Time length of a video clip*
θ	*Orientation of filters*
$g_{x,y}$	*Intensity value of the pixel* (x,y)
φ	*Number of dominant orientations*
(x,y)	*Cartesian plane coordinates*
2-D	*Two-dimensional*
$\|\cdot\|_2$	L^2*-norm*
$\|\cdot\|_1$	L^1*-norm*
$*$	*Convolution operation*
i, j, m	*Scalar index variable*
$\vert\cdot\vert$	*Absolute value*
$sign(\cdot)$	*The sign of a real number*
$Pr(\vec{x})$	*Probability of* $\vec{x}$
V	*Voting number of classifier*

List of abbreviations

AAM	*Active Appearance Model*
ASM	*Active Shape Model*
AU	*Action Unit*
AUC	*Area under the ROC-curve*
BDA	*Bayes Discriminant Analysis*
CK+	*The Extended Cohn-Kanade Database*
CLBP	*Completed Local Binary Pattern*
CLQP	*Completed Local Quantized Pattern*
CMFD	*Component-based Multiple Feature Descriptor*
CSFD	*Component-based Single Feature Descriptor*
COPE	*Infant Classification of Pain Expressions Database*
CRF	*Conditional Random Field*
DBN	*Dynamic Bayesian Network*
DisCSFD	*Discriminative Component-based Single Feature Descriptor*
DisSFD	*Discriminative Sparse Feature Descriptor*
DNPE	*Discriminative Neighbor Preserving Embedding*
DoM	*Difference of Magnitude*
DoO	*Difference of Orientation*
DoS	*Difference of Sign*
FACS	*Facial Action Coding System*
GMM	*Gaussian Mixture Model*
HCI	*Human-Computer Interaction*
HMMs	*Hidden Markov Models*
HOG	*Histogram of Oriented Gradients*
ICA	*Independent Component Analysis*
KLT	*Kanade-Lucas-Tomasi*
LBP	*Local Binary Pattern*
LDA	*Linear Discriminant Analysis*
LGBP	*Local Gabor Binary Pattern*
LMMBP	*Local Monogenic Magnitude Binary Pattern*
LMRBP	*Local Monogenic Real Binary Pattern*

LMIBP	*Local Monogenic Imaginary Binary Pattern*
LPP	*Local Preserving Projection*
LPQ	*Local Phase Quantization*
LQP	*Local Quantized Pattern*
LTP	*Local Ternary Pattern*
LUT	*Look-Up Table*
LXP	*Local XOR operator*
MCF	*Multi-Classifier Fusion*
MKL	*Multiple Kernel Learning*
MVDNPE	*Multi-view Discriminative Neighbor Preserving Embedding.*
NIR	*Near-Infrared*
NMF	*Non-negative Matrix Factorization*
NPE	*Neighbor Preserving Embedding*
OD	*Occlusion Detection*
PCA	*Principle Component Analysis*
PQDC	*Phase-Quadrant Demodulation Coding*
RBF	*Radial Basis Function*
SFD	*Sparse Feature Descriptor*
SIFT	*Scale-Invariant Feature Transform*
SRC	*Sparse Representation Classifier*
STGabor	*SpatioTemporal Gabor filters for motion processing*
STLMBP	*SpatioTemporal Local Monogenic Binary Pattern*
STLMIBP	*SpatioTemporal Local Monogenic Imaginary Binary Pattern*
STLMMBP	*SpatioTemporal Local Monogenic Magnitude Binary Pattern*
STLMRBP	*SpatioTemporal Local Monogenic Real Binary Pattern*
SVM	*Support Vector Machine*
TOP	*Three Orthogonal Planes*
VIS	*Visible light*
VLPQ	*Volume Local Phase Quantization*
WL	*Weight Learning*

List of original articles

This thesis is based on the following articles, which are referred to in the text by their Roman numerals (I–VII):

I Huang X & Zhao G & Hong X & Pietikäinen M & Zheng W (2013) Texture description with completed local quantized patterns. In: Image Analysis, SCIA 2013 Proceedings, Lecture Notes in Computer Science, 7944:1-10.

II Huang X & Zhao G & Zheng W & Pietikäinen M(2012) Spatiotemporal local monogenic binary patterns for facial expression recognition. IEEE Signal Processing Letters, 19(5):243-246.

III Huang X & Zhao G & Pietikäinen M & Zheng W (2010) Dynamic facial expression recognition using boosted component-based spatiotemporal features and multi-classifier fusion. In: Advanced Concepts for Intelligent Vision Systems, ACIVS 2010 Proceedings, Lecture Notes in Computer Science, 6475:312-322.

IV Zhao G & Huang X & Taini M & Li SZ & Pietikäinen M (2011) Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619.

V Huang X & Zhao G & Pietikäinen M & Zheng W (2011) Expression recognition in videos using a weighted component-based feature descriptor. In: Image Analysis, SCIA 2011 Proceedings, Lecture Notes in Computer Science, 6688:569-578.

VI Huang X & Zhao G & Zheng W & Pietikäinen M (2012) Towards a dynamic expression recognition system under facial occlusion. Pattern Recognition Letters, 33(16): 2181-2191.

VII Huang X & Zhao G & Pietikäinen M (2013) Emotion recognition from facial images with arbitrary views. Proc. the British Machine Vision Conference (BMVC 2013): 76.1-76.11.

The author is the first author in publications I-III, V-VII and the second author in publication IV. The writing and experiments of the papers (including Section 5 and related experiments in publication IV) were work of the present author, while valuable comments and discussion were given by the co-authors.

Contents

Abstract

Tiivistelmä

Acknowledgements 9

List of symbols 11

List of abbreviations 13

List of original articles 15

Contents 17

1 Introduction 19

1.1 Background 19

1.2 Facial expression recognition databases 21

1.3 The contributions of the thesis 23

1.4 Summary of original articles 25

1.5 Organization of the thesis 26

2 Spatial and temporal feature extraction for facial expression recognition 27

2.1 Introduction 27

2.2 Related work 28

2.3 Local binary pattern and variants 35

2.3.1 Basic of LBP and LBP-TOP 35

2.3.2 Implementations and variants 37

2.3.3 Challenges and solutions 38

2.4 Completed local quantized patterns 39

2.4.1 Completed local patterns extraction 39

2.4.2 Learning statistical dominant patterns 40

2.4.3 Implementation 41

2.4.4 Experiments 42

2.5 Spatiotemporal monogenic binary patterns 43

2.5.1 2-D multi-scale monogenic signal analysis 43

2.5.2 Encoding procedure 45

2.5.3 Experiments 46

2.6 Sparse spatiotemporal features and multi-classifier fusion 48
2.6.1 Sparse feature descriptor 49
2.6.2 Discriminative information enhancement 50
2.6.3 Multi-classifier fusion 51
2.6.4 Experiments 52
2.7 Discussion 53
3 Facial expression recognition in uncontrolled conditions 57
3.1 Background and motivation 57
3.2 Related works 61
3.3 Component-based feature descriptor 68
3.3.1 Component-based single feature descriptor 69
3.3.2 Component-based multiple feature descriptor 70
3.3.3 Weight learning 71
3.4 Discriminative component-based spatiotemporal features with NIR imaging for illumination variations 72
3.4.1 NIR imaging 72
3.4.2 Discriminative component-based features 72
3.4.3 Experiments 73
3.5 Occlusion detection for dynamic occlusions 74
3.5.1 Construction of occlusion detection 75
3.5.2 Experimental analysis 76
3.6 Multi-view neighborhood preserving embedding 78
3.6.1 Multi-view discriminative framework 79
3.6.2 Implementation 81
3.6.3 Experiments 82
3.7 Discussion 84
4 Summary 87
4.1 Methods and contributions 87
4.2 Limitation and future work 89
References 91
Original articles 99

1 Introduction

1.1 Background

The study of Darwin (1872) showed that facial muscle movement and the tone of the speech are two major ways for expressing the common emotions of human beings when communicating. In addition, Mehrabian (1968) indicated that the facial expression of the speaker contributes 55% to the effect of the spoken message, which is more than the verbal part (7%) and the vocal part (38%). Therefore, the face tends to be the most visible form of emotion communication. It makes facial expression recognition a widely used scheme for measuring the emotional state of human beings. The first study of Suwa *et al.* (1978) towards automatic facial expression recognition using computers was taken in 1978, but it has been increasingly developed since the 1990s. Today, facial expression recognition systems have great potential for different applications, such as human-computer interaction (HCI)[1] and medical analysis (Kaltwang *et al.* 2012). It is worth investigating the nature of facial expression and design of new systems.

The facial expressions are categorized into some basic emotions (*e.g.*, neutral, happiness, anger, disgust, sadness, fear, and surprise), as shown in Figure 1, which were presented by Darwin (1872). Therefore it is possible to make the relevant training and test materials of facial expressions available. Most of the existing efforts on facial expression recognition have been made to classify these basic facial expressions. They have taken much effort on measurement of facial geometric, description of facial appearance and its motion, also including classification of facial expressions. According to Tian *et al.* (2005, 2011), it is well known that the general framework for facial expression recognition is like the one presented in Figure 2. The system in general consists of face acquisition, facial expression extraction and representation, and facial expression classification. Specifically, the stage of face acquisition attempts to find the face area from the input images or videos and sometimes align all faces into a reference model by using facial landmarks, the procedure of facial expression extraction and representation aims to extract the features for well representing facial expressions, and finally, the classification step assigns the input pattern to a specific category.

Facial expression representation plays an important role in facial expression recogni-

[1] http://fox44.com/news/around-world/japanese-company-introduces-emotional-robot

tion. It can be viewed as generating good features for well describing the appearance, structure and motion of facial expressions. More specifically, facial expression features attempt to effectively describe the facial muscle or facial motion for static or dynamic facial images. They are usually elementary characteristics, *e.g.*, shape, color and texture. Alternatively, they can also refer to the result of Fisher criterion of minimizing within-class variations of facial expressions while maximizing between-class variations. According to the recent progress in face descriptors, facial geometric and appearance methods are two widely used schemes to extract the features of interest of facial expressions. The most representative ones are shape feature (Jain *et al.* 2011), Gabor filter (Buciu *et al.* 2003) and local binary pattern (LBP) (Feng *et al.* 2005a,b, Shan *et al.* 2009). They are utilized in the analysis of still images and dynamic image sequences for facial expression recognition. However, shape features with effective computational cost are not robust to low-resolution, while Gabor and LBP have large dimension of feature for facial images.

Another limitation of the existing facial expression recognition methods is such that recent researches have attempted to recognize facial expressions from data collected in a highly controlled environment. In practice, it is not always easy to get face images with good quality in the real-world environment. Conversely, the face images with poor quality, possibly at low resolution or bad illumination in visual surveillance, provide more challenges to facial expression recognition. Recently, the Emotion Recognition in the wild challenge and workshop (EmotiW 2013[2], 2014[3]) first explored the performance of emotion recognition methods to work in the wild. This further demonstrated that a facial expression analysis system should be able to automatically recognize facial expressions at lower resolution and handle the full range of head motion.

Over the years, much effort has been made on the development of geometric and appearance features for facial expression recognition. The task is far from being solved, although many interesting applications have been developed by posting constraints on the environments. This thesis starts from the spatial and temporal analysis of features and proposes new methods for facial expression recognition.

[2] http://cs.anu.edu.au/few/emotiw

[3] http://cs.anu.edu.au/few/emotiw2014.html

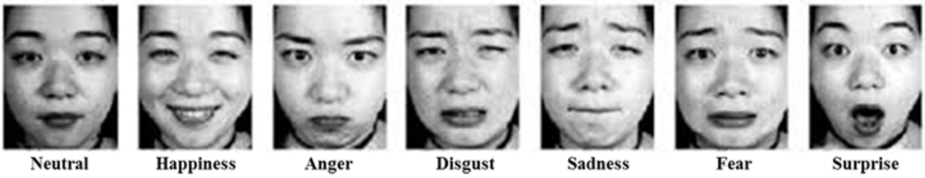


Fig 1. According to Lyons *et al.* (1998), seven basic emotions are commonly used in facial expression recognition.

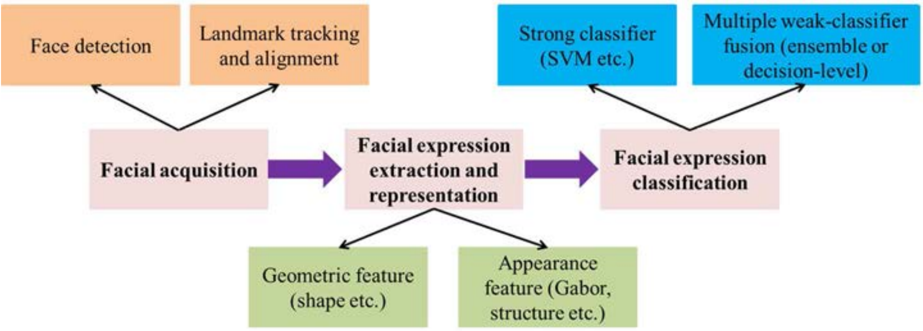

Face detection
Landmark tracking and alignment
Strong classifier (SVM etc.)
Multiple weak-classifier fusion (ensemble or decision-level)
Facial acquisition
Facial expression extraction and representation
Facial expression classification
Geometric feature (shape etc.)
Appearance feature (Gabor, structure etc.)

Fig 2. Classical pipeline of facial expression recognition.

1.2 Facial expression recognition databases

Having enough labeled data of facial expressions is a prerequisite of automatic facial expression recognition. Most of the existing studies on facial expression recognition have been based on the data sets of deliberately expressed emotions, elicited by asking the participants to perform a series of emotional expressions in front of a camera. As far as the databases used in this thesis are concerned, the following databases need to be mentioned. The details of other databases can be referred to Zeng *et al.* (2009).

The Cohn-Kanade database

This database consists of 100 university students, ranging in age from 18 to 30 years. Sixty-five percent were female, fifteen percent African-American, and three percent Asia or Latino. The subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units (AUs) and combinations of AUs, six of which were based on descriptions of prototypical emotions of anger, disgust, fear, happiness, sadness, and surprise. (Kanade *et al.* 2000).

The extended Cohn-Kanade database

This database (CK+) is the extension of the Cohn-Kanade Database. It has been further augmented to include 593 sequences from 123 subjects for seven expressions (additional 107 sequences, 26 subjects and contempt expression), which makes it more challenging than the original database. (Lucey *et al.* 2010).

Oulu-CASIC NIR&VIS database

This database in Paper IV consists of 80 subjects between 23 and 58 years old. 73.8% of the subjects were males, captured with two imaging systems, near-infrared (NIR) and visible light (VIS). It contains six basic emotions (*i.e.*, anger, disgust, fear, happiness, sadness, and surprise). All the images were taken under the normal, weak and dark illumination conditions. The normal illumination condition means that good lighting is used. The weak illumination condition means only a computer monitor is on and the subject sits in front of the computer during dynamic facial expression process. The dark illumination condition in its turn means lighting condition is close to darkness.

Infant Classification of Pain Expressions database

The infant classification of pain expressions database (COPE) contains 204 face images of 26 neonates and involves five neonatal expressions of these, 67 are rest, 18 cry, 23 air stimulus, 36 friction, and 60 pain. Photographs were taken of the infants at baseline rest and while experiencing several noxious stimuli: bodily disturbance, an air stimulus on the nose, friction on the external lateral surface of the heel, and the pain of a heel stick. (Brahnam *et al.* 2007).

BU-3DFE database

BU-3DFE database consists of 100 subjects (56% female and 44% male), ranging age from 18 years to 70 years old, with a variety of ethnic ancestries. Each subject performed seven expressions in front of the 3D face scanner. With the exception of the neutral expression, each of the six prototypic expressions (happiness, disgust, fear, angry, surprise and sadness) includes four levels of intensity. Therefore, there are 25 instant 3D expression models for each subject, resulting in a total of 2,500 3D facial expression models in the database. (Yin *et al.* 2006). Associated with each expression shape model, is a corresponding facial texture image captured at two views. As a result, the database consists of 2,500 two views texture images and 2,500 geometric shape models.

Multi-PIE database

The CMU Multi-PIE face database contains images from 337 subjects. Subjects were predominantly male (70%). 60% of subjects were European Americans, 35% Asian and 3% African Americans. The average age of the subjects was 28 years old. Data was captured during four sessions over a six months period. In each session, subjects were instructed to display various facial expressions (neutral, smile, surprise, squint, disgust and scream). Subjects were imaged under fifteen view points and nineteen illumination conditions while displaying a range of facial expressions. (Gross *et al.* 2010).

1.3 The contributions of the thesis

The contributions of the thesis are related to facial expression representation and potential methods to handle the problem of serious conditions. A brief overview of the thesis and its contributions are shown in Figure 3. The motivation of this thesis comes from two aspects: (1) recent feature descriptors that cannot efficiently describe facial expressions and (2) the practical conditions from which facial expression recognition suffers. Regarding the previously mentioned research objectives, the main contributions can be summarized from two different aspects.

The first main contribution is three texture based methods for facial expression recognition. The first description is based on still images by exploiting local quantized pattern (LQP) (Hussain & Triggs 2012, Hussain *et al.* 2012) and completed information. The complementary difference based on magnitude and orientation can provide more sufficient information than only using the sign-based one. Moreover, the revised vector quantization makes the LQP to have much low computational cost. It also allows such a descriptor to deeply explore the large-scale spatial neighbors or irregular neighbor topology and create a flexible codebook for various applications. The second description is based on a two-layer representation for dynamic facial images. It presents an alternative way to combine LBP with orientation-sensitive filters from dynamic images. The last one is to marginally utilize the geometric information and feature selection to compress the local binary pattern from three orthogonal planes (LBP-TOP) (Zhao &

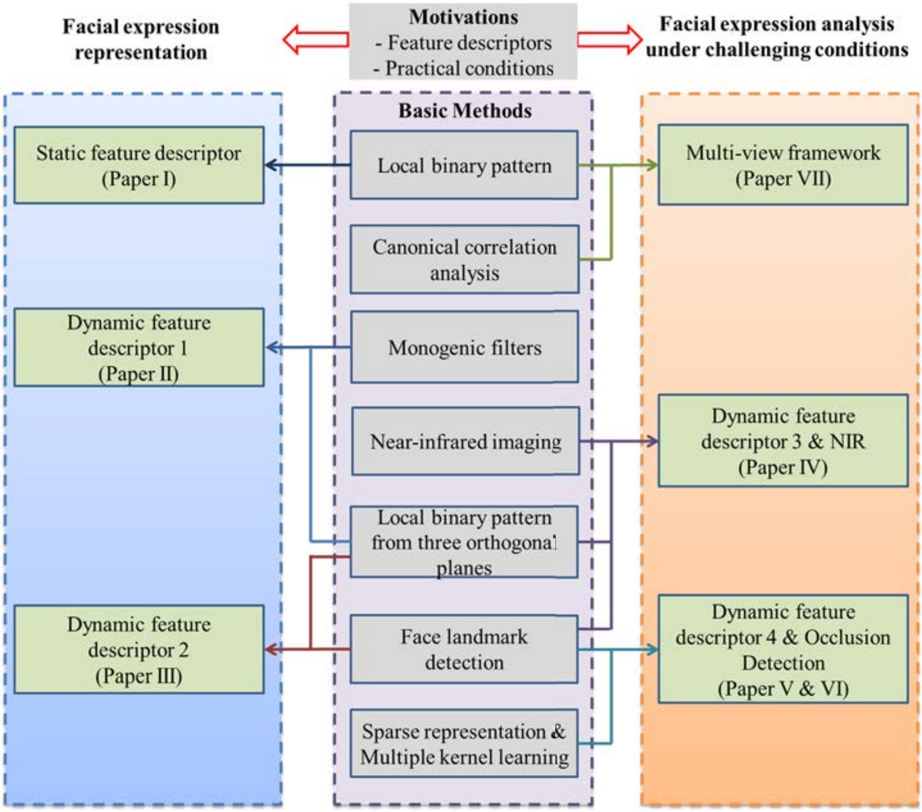

Facial expression representation
Motivations
- Feature descriptors
- Practical conditions
Facial expression analysis under challenging conditions
Basic Methods
Static feature descriptor (Paper I)
Local binary pattern
Multi-view framework (Paper VII)
Canonical correlation analysis
Dynamic feature descriptor 1 (Paper II)
Monogenic filters
Near-infrared imaging
Dynamic feature descriptor 3 & NIR (Paper IV)
Local binary pattern from three orthogonal planes
Dynamic feature descriptor 2 (Paper III)
Face landmark detection
Dynamic feature descriptor 4 & Occlusion Detection (Paper V & VI)
Sparse representation & Multiple kernel learning

Fig 3. Description of the outline of the thesis. The gray areas in the middle indicate the motivations and basic methods of this thesis, and the shaded ones the contributions of this thesis.

Pietikäinen 2007). It presents a new way to combine the geometric and appearance features.

The second main contribution is three new methods to resolve the problems caused by illumination variations, facial occlusion and pose changes. The first method presents the combination of NIR imaging technique and the component-based method to reduce the influence of illumination. In addition, an improvement to the component-based method is proposed via feature selection method. The second method is further based on the component-based method and sparse representation to resolve the difficulties by facial occlusion. The major work includes the multiple component-based features and occlusion detection (OD). The new system can work well for facial expression recognition in both normal and occluded conditions. According to multi-view theorem and manifold learning, the third method is to regard the facial image under arbitrary view as a multi-view correlation and construction problem. This method provides a new way to tackle facial expression recognition under various views.

1.4 Summary of original articles

This thesis concentrates on the new descriptors presented in Papers I-III and the methods for illumination variations, facial occlusion and pose changes presented in Papers IV-VII. The main inspiration comes from (1) the applications of spatial and dynamic texture descriptors for facial expression recognition, especially using LBP and LBP-TOP, (2) hybrid feature description and the application of machine learning in facial expression recognition.

The idea of Paper I is to extend LQP via the completed information and revised vector quantization. LQP makes the statistical patterns domain-adaptive and stable to different applications. In addition, it allows the utilization of irregular spatial sampling structure. Two improvements are proposed for LQP. One is to exploit completed information, including the sign-based, magnitude-based and orientation-based differences. The other is to modify the objective function by introducing weights and pre-define initialization. The proposed descriptor in Paper I performed well and effectively in facial expression recognition and texture classification.

The method in Paper II can be regarded as parallel to Paper I. Both of them investigate the role of the orientation. Paper II is motivated by the multi-layer representation for face recognition in still image which combined the Gabor or Haar and LBP as in Zhang *et al.* (2005), Roy & Marcel (2009). In their work, LBP features extracted from Haar (Roy & Marcel 2009) or Gabor (Zhang *et al.* 2005) representation performed more robust to illumination variations or pose changes than the original LBP. Following their work, the monogenic filters used in Paper II target to extract the magnitude and orientation information and yield redundant information of facial images. Additionally, the efficient computation is considered. For encoding magnitude and orientation, a form similar to LBP is utilized to exploit the neighboring information from magnitude and orientation.

Paper III describes the utilization of geometric information and appearance information for dynamic images. Geometric information is considered to select the regions of interest, while the dynamic appearance features are extracted from these regions. This method can yield the influence of less important regions, *e.g.*, face boundary. In addition, the schemes of feature selection and multiple classifier fusion are used to boost the final performance of the facial expression recognition system.

Paper IV further extends the method presented in Paper III on NIR facial images. NIR imaging technique makes the obtained facial images invariant to illumination changes. Based on NIR images, the discriminative component-based LBP-TOP is proposed for describing the spatiotemporal features of six facial regions. The completed system is verified to obtain satisfying results in different illuminations.

Paper V re-visits the component-based method of Paper IV. It introduces two dynamic features (LBP-TOP and edge histograms) for eyes, nose and mouth. According to the facial action coding system (FACS), it is well known that eyes, nose and mouth act different roles for facial expressions. We further develop the weight learning method to assign the optimal weight to each facial region. Paper VI is the extension of the work done in Paper V. Sparse representation is used to perform occlusion detection on the dynamic images. An improvement in normal and occlusion conditions is shown.

The last work presented in Paper VII aims to design a reliable framework to reduce the influence of pose changes. The multi-set canonical correlation analysis and discriminative criterion are two basic schemes in this paper. It starts from the development of discriminative neighboring preserving embedding based on graph and discriminative criterion. Then, the correlation of each facial expression in arbitrary view is considered in multi-set canonical correlation analysis. It regards the emotion recognition from facial images with arbitrary views as an optimization problem.

1.5 Organization of the thesis

This thesis is organized as follows: in this chapter, the background of the research topics, the objectives, the research problem, motivations and contributions are all briefly discussed.

Chapter 2 presents a literature overview of the research relevant to the facial expression recognition problem with its background, objectives and challenges, followed by the presentation of three description methods.

Chapter 3 expands the scope of the facial expression recognition problem in real-world application, followed by two variants of the component-based method, and presented three respective methods based on video analysis and image analysis for dealing with illumination variations, facial occlusion and pose changes.

Chapter 4 summarizes the main work of the thesis and draws conclusions.

2 Spatial and temporal feature extraction for facial expression recognition

This chapter firstly presents the state-of-the-art feature extraction methods in still images and video sequences for describing facial expressions. This is followed by the introductions of the methods originally presented in Papers I, II and III.

2.1 Introduction

Psychologists postulate that facial expressions have a consistent and meaningful structure for inferring inner affective states of human beings (Ekman 1993, Ekman & Davidson 1994). Facial expression therefore has become an important element in recognition of human emotions. However, recognition of facial expressions is a complex task as physiognomies of faces vary from one individual to another quite considerably due to age, ethnicity, gender, facial hair. Since the 1990s, most of the efforts have been made to develop theorems and methods for automatic facial expression recognition. Some of these works have attempted to extract the features of interest from still and dynamic facial images for representing facial expressions:

(1) For still images, the research has considered potential applications, for example, ‘Smile detection’ of Sony camera and pain detection for medical image analysis (Nanni *et al.* 2010). In addition, it is easy to exploit the appearance features on still images. In recent years, local binary pattern (LBP), a simple and effective operator, has been widely applied to face recognition and facial expressions recognition (Feng *et al.* 2005a,b, Shan *et al.* 2009). Though it can achieve some promising results on facial expression recognition, there is still some space to further exploit LBP variants for facial expression recognition.

(2) Recent research (Raducanu & Dornaika 2008, Shan & Braspenning 2010, Krumhuber *et al.* 2013) has suggested that still images may not clearly reveal subtle changes in faces. It is importantly pointed out that some dynamic extensions of LBP have been proposed for facial expression recognition (Zhao & Pietikäinen 2007, Pfister *et al.* 2011, Almaev & Valstar 2013). The generic framework of Zhao & Pietikäinen (2007) is shown to be easily implemented for LBP variants in dynamic images. In addition, this framework has low computational cost. Indeed, it is possible to develop a new dynamic

LBP for facial expression recognition.

Therefore, this thesis revisits the facial expression recognition based on LBP and its variants in this chapter. It starts from a novel spatial descriptor based on LBP variants and LQP for still images in Section 2.4. Then, two new spatiotemporal features based on LBP-TOP are discussed in Sections 2.5 and 2.6, respectively.

2.2 Related work

Face representation has been studied intensively for automatic facial expression recognition over the past decades, and a variety of approaches have been presented based on still and dynamic facial images. There are many categorization classes for facial feature representation. More generally, they can be categorized into geometric and appearance-based approaches and hybrid feature method, which are briefly summarized in Tables 1, 2 and 3. The other categorizations can be referred to Fasel & Luettin (2003), Tian *et al.* (2005, 2011), Zeng *et al.* (2009), Whitehill *et al.* (2013).

Geometric-based feature method

Geometric features are driven as more efficient methods, most of which have shown good performance in facial expression recognition (Kanaujia & Metaxas 2006, Shin & Chun 2008, Jain *et al.* 2011, Rudovic *et al.* 2013), if facial landmarks can be accurately detected and tracked. Specifically, geometric features are in general defined as (1) one can use position of facial feature points as visual information (Zhang *et al.* 1998, Rudovic *et al.* 2013); (2) one can measure the geometrical displacement of facial feature points (Kotsia & Pitas 2007, Kotsia *et al.* 2008a); and (3) one can form a geometric graph representation of the faces (Tian *et al.* 2002, Zhang & Ji 2005). These efforts have been made to interpret the facial structure of various facial expressions.

(1) Using position of geometric points is a simple way to directly measure the contour of faces. A representative one can be found in Zhang *et al.* (1998). In this study, they simply used 34 fiducial points to represent the facial geometry of still image and formulate them into a feature vector. For still images, Rudovic *et al.* (2013) used 39 facial landmark points to describe facial expression in various views. This kind of methods has recently received more attention, especially in dynamic facial images (Kanaujia & Metaxas 2006, Shin & Chun 2008, Jain *et al.* 2011). The work presented in Shin & Chun (2008) used eighteen major feature points defined in MPEG 4

Table 1. Summary of geometric-based feature methods.

Reference	Features	Tracking method	Dynamic	Classifier
Zhang *et al.* (1998)	34 facial feature point	Manual labeling	No	Two-layer perceptron
Tian *et al.* (2002)	15 parameters of geometric features	Multi-state models	Yes	Three-layer neutral network
Zhang & Ji (2005)	Geometric deformation feature of AUs	Kalman filtering	Yes	Dynamic Bayesian networks
Kanaujia & Metaxas (2006)	78 facial feature points	Modified active shape model	Yes	Conditional random fields (CRF)
Kotsia & Pitas (2007)	Geometric deformation feature	Kanade-Lucas-Tomasi (KLT) tracker	Yes	Support vector machine (SVM)
Kotsia *et al.* (2008a)	Geometric deformation feature	-	No	SVM
Shin & Chun (2008)	18 facial feature points	Dense optical flow	Yes	Hidden Markov models
Zafeirious & Petrou (2010)	Geometric deformation feature	KLT tracker	No	Sparse representation
Jain *et al.* (2011)	68 facial points	Generalized proscrustes analysis	Yes	Latent-dynamic CRF
Rudovic *et al.* (2013)	39 facial feature points	Active appearance model	No	SVM

and then applied the dense optical flow method to track the feature points for sequential frames. In Jain *et al.* (2011), more than eighteen facial points were located by using Generalized Procrustes analysis and then formed as a 136 dimensional feature vectors for each facial image. In their work, they used this feature vector to describe the geometric structure of each frame over time in a facial expression video clip. Thanks to the temporal model classifier, for example, Hidden Markov models (HMMs) and Dynamic Bayesian networks (DBN), these features can be simply modeled for the dynamics of facial expression.

(2) More general formulation for geometric features is to quantify the facial

movement. It is addressed by measuring the displacement of facial points between one facial image and the reference image. Specifically, the reference image is always chosen from the facial image with the neutral expression. As said that, this kind of deformation is analogous to human observations of facial activities. (Zhang & Ji 2005, Kotsia & Pitas 2007, Kotsia *et al.* 2008a, Zafeirious & Petrou 2010). The common procedure of this formulation is done as (1) the grids are tracked in consecutive frames over time via the grid-tracking and deformation, (2) the difference of node coordinates is calculated by comparing the neutral frame with the greatest expression-intensity one, and (3) these differences are fed into the classification stage.

(3) The complicated approach is to extract the shape of each facial component, such as eyes, brows, cheeks, and lips, to describe the facial representation. But this idea has mostly been applied to recognize the AUs (Tian *et al.* 2002, Zhang & Ji 2005), since it should conduct the parametric setup according to FACS (Ekman & Friesen 1978).

Geometric features have the advantage of low dimension and simplicity. Nevertheless, all methods for constructing the geometric features suffer from the problems caused by the variation of lighting and non-rigid motion. Additionally, they are sensitive to the error of image registration and motion discontinuities. Therefore, it is difficult to design a deterministic physical model of facial expressions that can exactly better represent facial geometrical properties and muscle activities for all facial expressions.

Appearance-based feature method

Some of the studies have suggested that the appearance-based features are more stable to image spatial transforms than the geometric features, especially for inaccurate mis-alignment and images with low-resolution. More specifically, for an image, it can be characterized by the appearance-based features in terms of (1) variation of pixel intensity or (2) low-level feature in the face. In the past decade, there are numerous approaches on the appearance-based features for facial expression recognition.

(1) For still images, Gabor (Zhang *et al.* 1998, Tian *et al.* 2002) and local binary pattern (LBP) (Feng *et al.* 2005a,b, Shan *et al.* 2009, Moore & Bowden 2011) are two most representative ones for facial expression recognition. Gabor feature is related to the perception in human visual system. It consists of a sinusoid carrier signal modulated by a Gaussian, each of which determines the frequency the filter is tuned to. Some of the most successful facial expression recognition systems to date have utilized the Gabor energy filters. One reason for their success may stem from the fact that they are robust

Table 2. Summary of appearance-based feature methods.

Reference	Features	Dynamic	Classifier
Yacoob & Davis (1996)	Optical flow	Yes	A rule based system
Zhang *et al.* (1998)	Gabor	No	Two-layer perceptron
Tian *et al.* (2002)	Gabor	Yes	Neutral network
Buciu *et al.* (2003)	Independent component analysis (ICA) and Gabor	No	Maximum correlation classifier
Feng *et al.* (2005a,b)	Local binary pattern (LBP)	No	Linear programming
Shan *et al.* (2005)	LBP	No	SVM
Littlewort *et al.* (2006)	Gabor	Yes	SVM
Yesin *et al.* (2006)	Optical flow	Yes	Hidden Markov models
Zhao & Pietikäinen (2007)	Local binary pattern from three orthogonal planes	Yes	SVM
Shan *et al.* (2009)	LBP and Boosted-LBP	No	SVM
Jabid *et al.* (2010)	Local directional pattern	No	SVM
Wu *et al.* (2010)	Gabor motion energy filters	Yes	Linear SVM
Moore & Bowden (2011)	Variants of LBP	No	SVM
Jun *et al.* (2011)	Compact LBP	No	Nearest neighbor classifier
Sánchez *et al.* (2011)	Differential optical flow	Yes	SVM
Long *et al.* (2012)	Spatiotemporal features based on ICA	Yes	SVM
Almaev & Valstar (2013)	Local Gabor binary pattern from three orthogonal planes	Yes	SVM
Feng *et al.* (2013)	LBP on key points	No	SVM
Jiang *et al.* (2014)	Local Phase Quantization from three orthogonal planes	Yes	SVM

to contrast polarity and image alignment errors. Using spatial Gabor energy filters as the feature type, Littlewort *et al.* (2006) achieved good performance reported on the Cohn-Kanade dataset when classifying the seven basic emotions. Zhang *et al.* (1998) efficiently applied Gabor wavelet to 34 facial points, which was more efficient than processing the whole image. It also showed that appearance features based on facial points can preserve the low dimension and have promising performance.

On the other hand, LBP is simple to implement, fast to compute and has led to high accuracy in texture-based recognition tasks. In recent years significant progress has been made in using LBP for facial expression recognition (Feng *et al.* 2005a,b, Feng *et al.* 2013). Perhaps, the most important property of the LBP operator in real-world applications is its invariance against monotonic gray level changes caused by illumination variations, for example. Another property of equal importance is its computational simplicity, which makes it possible to analyze images in challenging real-time settings. Shan *et al.* (2005, 2009) preliminarily applied LBP for representing facial expressions. They used the simple LBP for representing salient micro-patterns of face images. They also showed that they are more discriminative and efficient than Gabor features. Following them, there are many efforts attempting to directly utilize variants of LBP for estimating intensity of facial expressions (Chang *et al.* 2013), facial AUs (Jiang *et al.* 2011, Yuce *et al.* 2013), multi-view facial expression recognition (Moore & Bowden 2011) and 3D facial AUs detection (Bayramoglu *et al.* 2013). Some studies have attempted to enhance the representative ability of LBP to further improve the performance (Jabid *et al.* 2010, Jun *et al.* 2011). Jabid *et al.* (2010) proposed a new variant of LBP for recognizing facial expressions. Interestingly, they computed the edge response values in all directions at each pixel point, and then generated a code according to the relative magnitude's strength. Jun *et al.* (2011) obtained a compact LBP through the maximization of mutual information between features and class labels. More applications of LBP in facial expression recognition are referred to Smith & Windeatt (2010), Jun *et al.* (2011), Huang *et al.* (2011), Valstar *et al.* (2011), Majumder *et al.* (2013), Yu *et al.* (2013), Yuce *et al.* (2013).

(2) For video sequences, a well-known method is by Yacoob & Davis (1996) who applied dense optical flow to facial expression recognition or facial actions in the 1990s. The procedure of using dense optical flow is to compute the motion in the rectangular regions for estimating the activity of face region. It can catch the smooth flow and global information. Additionally, it can get the accurate time derivatives using more than two frames. Due to its advantage, Lien *et al.* (1998) proposed a spatial-temporal descriptor integrating dense optical flow, feature point tracking and high gradient component analysis and then used HMMs to recognize fifteen AUs. Other than that, motion pattern of facial expression was represented by using the horizontal and vertical components of optical flow (Yesin *et al.* 2006, Sánchez *et al.* 2011). However, optical flow is sensitive to image misalignment error. Recently, image filters proposed and texture descriptor for still images have been very attractive techniques for recognizing dynamic facial expression.

Some representative ones of image filters are Haar features (Yang *et al.* 2007), Gabor wavelet representation (Wu *et al.* 2010) as well as independent component analysis (ICA) (Long *et al.* 2012). Gabor representations have been used to design the temporal descriptor. Wu *et al.* (2010) made Gabor motion energy filters as a biologically inspired representation for dynamic facial expressions. ICA is also a common method to decompose facial expressions into independent non-Gaussian signals. A new recent application of ICA was presented in Long *et al.* (2012), where they employed ICA to learn spatiotemporal filters from natural videos, and then constructed feature representations for input videos based on learned filters. The combination of more than two image filters has also been investigated. For example, the combination of ICA and Gabor filters proposed by Buciu *et al.* (2003) was used to classify seven categories of facial expressions.

For the texture descriptor, a simple yet very efficient texture operator, LBP, has been extended to dynamic images (Zhao & Pietikäinen 2007). In their work, the LBPs were used to describe the temporal motion and the texture of appearance to achieve an effective dynamic facial expression description. Another latest extension was presented in Almaev & Valstar (2013) that used LBP to encode the templates of multi-scale and multi-orientation Gabor filters, named as local Gabor binary pattern from three orthogonal planes (LGBP-TOP), for achieving good results on emotion recognition in unrestricted conditions. Moreover, an interesting method proposed by Jiang *et al.* (2011), Jiang *et al.* (2014) used local phase quantization to describe the temporal information for facial actions. Recent studies on LBP have demonstrated that the dynamic LBP more easily and powerfully describes the temporal variation of facial expressions than a complicated temporal model using DBN or HMMs.

Hybrid feature method

It is known that geometric-based and appearance-based features have their respective special properties and limitations, *e.g.*, geometric-based features have effectiveness in computation while they are sensitive to noise; in contrast, appearance-based features are robust to image mis-alignment but it takes much time in computation. Therefore, the fusion of these features has become an active research topic (Shan *et al.* 2009, Meng *et al.* 2011, Chen *et al.* 2013, Zavaschi *et al.* 2013). Among these methods, decision-level and feature-level fusions are two common ways to fuse multiple feature sets.

The decision-level fusion methods aim to explore the utilization of the classifiers

Table 3. Summary of hybrid feature methods.

Reference	Combination method	Feature	Dynamic	Classifier
Kotsia *et al.* (2008b)	Fusion RBFs	Shape and texture features	No	Distance classifier and SVM
Gajsek *et al.* (2010)	Score-level fusion	Audio and video features	Yes	Maximum correlation classifier and SVM
Meng *et al.* (2011)	Objective function construction	Motion history histogram and motion change frequency	Yes	SVM-2K
Chen *et al.* (2013)	Feature concatenating	Motion histogram images from HOG and Image-HOG	Yes	Gaussian SVM
Ouyang & Sang (2013)	Classifier combination	HOG and LBP	No	Sparse representation
Sikka *et al.* (2013)	Multiple kernel learning (MKL)	Bag of words, GIST, audio feature and LBP-TOP	Yes	-
Zavaschi *et al.* (2013)	Ensemble of classifiers	Gabor and LBP	No	Decision tree
Zhang *et al.* (2013)	Hessian MKL	LBP and HOG	No	-

to ensemble the decision of all feature sets. The probability or voting of classifiers is usually used. For example, Ouyang & Sang (2013) used a classifier combination strategy to fuse the results of histograms of oriented gradient (HOG) and LBPs with sparse representation classifiers. Gajsek *et al.* (2010) used weighted sum-rule fusion to combine audio and video features at the matching score level for obtaining a good performance in multi-modal emotion recognition.As another example, Kotsia *et al.* (2008b) fused the scores of shape information and texture features using radial basis function (RBF) neural network. Besides the probability, the implementation of the decision tree or the design of the objective function of classifiers are other ways to obtain a stable decision for all feature sets. The new application of the decision tree found in Zavaschi *et al.* (2013) was to employ a tree from the binary number of pair-class classifier to learn the decision of fusion of Gabor features and LBPs. For the design of the objective function of classifiers, the plausible work can be seen in Meng *et al.*

(2011), where they used two-view support vector machine (SVM) to fuse the dynamic geometric information from motion history histogram and motion change frequency of LBPs for facial expression recognition.

Feature-level fusion is another alternative technique to fuse multiple features. Among the feature-level fusion methods, a simple one is to concatenate all feature sets into a new feature vector. Its intuitive advantage is the simple computational cost, but it may ignore the correlation and importance of each feature set. Additionally, the more feature sets, the higher dimension of the new feature. Some alternative ways are to use machine learning techniques for fusing multiple feature sets. They not only preserve the low dimension, but also possibly explore the power of each feature set. For example, Chen *et al.* (2013) alternatively used a bag of words based representation for combining motion histogram images from HOG and image-HOG for multi-modal emotion recognition. A useful and powerful technique, multiple kernel learning (MKL) method, has recently been applied for multi-modal emotion recognition, because it can flexibly rank all feature sets according to their importance. Additionally, it has a practical and theoretical framework based on SVMs. The newest utilization of MKL was described in the work of Sikka *et al.* (2013). They proposed to use MKL to optimally combine the feature kernels of a bag of words, GIST feature, audio feature, as well as LBP-TOP features for emotion recognition in the wild. The work in Zhang *et al.* (2013) presented a new framework for the MKL problem by expanding the HessianMKL algorithm into multi-class SVM with one-against-one rule. This framework was further utilized to recognize seven facial expressions by combining three kernel functions and two image representations.

2.3 Local binary pattern and variants

Three new feature extraction methods in Sections 2.4, 2.5, and 2.6 are proposed on the basis of LBP and LBP-TOP.

2.3.1 Basic of LBP and LBP-TOP

The local binary pattern operator was presented in a generic form by Ojala *et al.* (2002). Consider a monochrome image $I(x,y)$ and let g_c denote the gray level of an arbitrary pixel (x,y). Moreover, let g_m denote the gray value of a sampling point in an evenly spaced circular neighborhood of P sampling points and radius R around point (x,y). The

LBP operator is defined as

$$h_{P,R}(x_c,y_c) = \sum_{m=0}^{P-1} L((g_m - g_c) \geq 0)2^m, \tag{1}$$

where $L(x)$ is a logical function with $L(x) = 1$ if x is true and $L(x) = 0$ otherwise. For obtaining the compact feature for facial expression, facial image is separated into several blocks. For each block, the LBP operator with specific sampling points and radius is applied. The histograms of all blocks are concatenated into one feature.

The essence of LBP-TOP is such that it applies LBP (Ojala *et al.* 2002, Ahonen *et al.* 2006) separately on three orthogonal planes (XY, XT and YT) which intersect in the center pixel. All histograms can describe effectively appearance, horizontal motion and vertical motion from an image sequence (Zhao & Pietikäinen 2007). For LBP-TOP, it is possible to change the radii in axes X, Y and T, which are marked as R_x, R_y and R_t, respectively. Meanwhile, the numbers of neighboring points in XY, XT and YT planes are denoted as P_{XY}, P_{XT} and P_{YT}. Using these notations, LBP-TOP features can be denoted as LBP-TOP$_{P_{XY},P_{\mathrm{XT}},P_{\mathrm{YT}},R_{\mathrm{X}},R_{\mathrm{Y}},R_{\mathrm{T}}}$.

When calculating LBP-TOP$_{P_{\mathrm{XY}},P_{\mathrm{XT}},P_{\mathrm{YT}},R_{\mathrm{X}},R_{\mathrm{Y}},R_{\mathrm{T}}}$ distribution for an assumed X× Y × T dynamic texture, only the center part of the dynamic texture can be taken into account because a large enough neighborhood cannot be used on the borders of this 3D space. A histogram of the dynamic texture can be defined as

$$H_{i,j} = \sum_{x,y,t} L(h_j(x,y,t,P_j,R_j) = i), \tag{2}$$

where $i = 0,\ldots,N_j - 1$, $j = 0,1,2$, N_j is the number of different labels generated by the LBP operator in the jth plane ($j = 0$: XY, $j = 1$: XT and $j = 2$: YT), and $h_j(x,y,t)$ expresses the LBP code of the center pixel (x,y,t) in the jth plane:

$$h_j(x,y,t,P_j,R_j) = \sum_{m=0}^{P_j-1} L((g_{m,j} - g_{c,j}) \geq 0)2^m, \tag{3}$$

and $g_{m,j}$ denotes the intensity value of neighboring pixel, $g_{c,j}$ is the intensity value of the center pixel.

To acquire a coherent description, even though the videos to be compared are of different spatial and temporal sizes, the histograms must be normalized by using L1-normalization. Regarding the LBP-TOP histogram, a description of the video can be effectively obtained based on the LBP codes from the three orthogonal planes. These three histograms are concatenated into one feature histogram in order to build a global description of a video, including both the spatial and temporal features.

2.3.2 Implementations and variants

There are a large number of LBP and LBP-TOP variants to date that are applied in various fields, for example, face recognition and texture classification.

For details of LBP variants, we refer to Pietikäinen *et al.* (2011), Huang *et al.* (2011). Here, we take two recent techniques more related to our thesis into account. These are the completed local binary pattern (CLBP) (Guo *et al.* 2010) and the local quantized pattern (LQP) (Hussain & Triggs 2012, Hussain *et al.* 2012). The first one can be viewed as a completed modeling of the LBP operator. The image local differences were decomposed into two complementary components: the sign and the magnitudes and two operators, CLBP-Sign and CLBP-Magnitude were proposed to code them. As well, the center pixels represented the image gray level and they were converted into a binary code by global thresholding. All were combined as complementary information to improve the texture classification. LQP can be seen as the generalization of the LBP operator. The binary patterns of the more complicated spatial structures were encoded by vector quantization procedure. It allowed local pattern features to have many more pixels and quantization levels without sacrificing simplicity and computational efficiency.

LBP-TOP was originally proposed as an effective video descriptor for texture classification. It has become attractive since the temporal patterns are regarded as two motion texture images and it has low computational cost independent of temporal classifiers, such as HMMs. Since then, LBP-TOP has inspired much research on new local variants for different applications, including not only facial expression recognition and texture classification (Zhao & Pietikäinen 2007), but also human action recognition (Kellokumpu *et al.* 2011), face recognition (Lei *et al.* 2008), micro-expression analysis (Li *et al.* 2013) and crowd density estimation (Yang *et al.* 2011).

In recent years, there are many representative extensions of LBP-TOP (Nanni *et al.* 2011, Pfister *et al.* 2011, Zhao *et al.* 2012, Chan *et al.* 2012, Ruiz-Hernandez & Pietikäinen 2013). For example, local ternary pattern from three orthogonal planes (LTP-TOP) proposed by Nanni *et al.* (2011) quantized intensity differences of neighboring pixels and center pixel into three levels to increase the robustness against noise. Completed local binary pattern from three orthogonal planes (CLBP-TOP) presented in Pfister *et al.* (2011) combined intensity difference and magnitude quantization of neighboring pixels and center pixel to increase the robustness against noise. Local ordinal contrast patterns presented in Chan *et al.* (2012) used a pairwise ordinal contrast measurement of pixels from a circular neighborhood starting at the center pixel to increase the robustness against intensity noise. Zhao *et al.* (2012) proposed histogram Fourier LBP-TOP for rotation-invariant video description. A re-parameterization of the second local ordinal Gaussian jet presented in Ruiz-Hernandez & Pietikäinen (2013) was used to encode LBP for more robust and reliable representation.

2.3.3 Challenges and solutions

LBP variants

CLBP and LQP are two variants of LBP in recent years for texture classification and face recognition. However, CLBP and LQP suffer from some in itself limitations:

(1) Large neighbor sets may provide an increase in discriminative power (Ojala *et al.* 2002). However, the hand-crafted coding for LBP may limit the various types of diverse structures and the depths of the pixel comparisons. One reason is that the size of histograms increases exponentially with the spatial support of the pattern and the number of quantization levels. For example, with a local pattern including 24 pixel-comparisons, the table has 2^{24} entries. (Hussain & Triggs 2012). However, CLBP uses a similar coding method to LBP (Ojala *et al.* 2002). Therefore, this makes it difficult for CLBP to handle irregularly deep spatial structure, especially for large local pattern neighborhoods.

(2) The LQP can encode the features of the irregular spatial structure in a simple and interesting way. Unfortunately, it missed some information like the one gained by CLBP. Their complementarity may thus provide a superior ability and generalized performance for LBP.

To tackle their bottleneck, a model of completed quantized pattern is presented in Paper I to fully consider the intensity and orientation of each pixel, to be introduced in Section 2.4.

Variants of LBP-TOP

For the implementation of LBP-TOP, there are two major limitations.

(1) The first limitation is brought by the single layer, where only the intensity is used to generate the occurrence histogram. An advantage of LBP is its invariance to monotonic gray-level changes. However, this advantage does not hold for some cases, such as illumination or pose change. Thus, the concept of multi-layer was introduced in LGBP-

TOP which was learnt from multi-scale and multi-orientation Gabor images (Almaev & Valstar 2013). But the main problem of LGBP-TOP is with the computational cost due to the use of Gabor filters. That is, the more Gabor images are considered, the more computational time it takes. Additionally, the representation capability of Gabor filter is not optimal as the maximum bandwidth of Gabor filter is limited to approximately one octave. To overcome these limitations, a two-layer strategy is presented in Paper II to learn more compact patterns, introduced in Section 2.5. This method simultaneously considers the robustness, discriminative power, and representation capability of features in a two-layer feature model.

(2) Another limitation of LBP-TOP in facial expression recognition is that it is only based on the whole face regions, while not all face regions have equally important roles in facial expression recognition. To solve this limitation, a sparse spatiotemporal feature descriptor is proposed in Paper III to embed the geometric information in LBP-TOP. It has much higher discriminative power and lower dimension than LBP-TOP as it explores the spatial and temporal features from the regions of interest. This method is introduced in Section 2.6.

2.4 Completed local quantized patterns

This section presents a general extension of LQP proposed in Paper I, named as completed local quantized pattern (CLQP). CLQP resolves two problems of LQP: (1) LQP only exploits the difference of sign and (2) the vector quantization is not effective. It inherits the merits of LQP and CLBP. It covers a large local neighborhood with an "economic" number of codes. It also allows us to learn the domain-adaptive codebook for reflecting the most representative patterns on the database. The framework of CLQP is presented in Figure 4.

2.4.1 Completed local patterns extraction

Given an image I, the local pattern at the spatial coordinate (x,y) can be formulated as follows:

$$\vec{x} = [l(g_{x,y}, g_{x_1,y_1}), l(g_{x,y}, g_{x_2,y_2}), \ldots, l(g_{x,y}, g_{x_P,y_P})], \tag{4}$$

where $l(g_{x,y}, g_{x_i,y_i})$ is the formula of values of two coordinates, $g_{x,y}$ is the intensity or other low-level feature of the pixel (x,y), and (x_i,y_i) is the neighbor sampling points of (x,y). The construction method of $l(g_{x,y}, g_{x_i,y_i})$ includes the difference of sign, magnitude and orientation.

(1) Difference of sign and magnitude (DoS and DoM): Given the intensity value g_c of the center pixel (x,y) and g_p of its neighbor (x_i,y_i), their difference can be calculated a $d_i = g_{x_i,y_i} - g_{x,y}$. DoS can be represented by $l_s(g_{x_i,y_i}, g_{x,y}) = L((g_{x_i,y_i} - g_{x,y}) \geq 0)$ and the DoM can be presented as $l_m(g_{x_i,y_i}, g_{x,y}) = L(|g_{x_i,y_i} - g_{x,y}| \geq \delta)$, where δ is the mean value of all magnitude and $|\cdot|$ is an absolute operator.

(2) Difference of orientation (DoO): The basic idea is to encode the relationship between dominant orientations of neighboring pixels in the image. It consists of three stages: the estimation of orientation angle of each pixel, the calculation of dominant orientation, and the operation of neighboring orientations. The estimation procedure of the orientation angle of each pixel can be found in Gizatdinova & Surakka (2006) and Paper I.

Given the orientation angle $\theta(x,y)$ and its angles of neighbors $\theta(x_i,y_i)$, the orientation angles are quantified by applying the quantification function

$$q = mod(\lfloor \frac{\theta(x,y)}{\frac{2\pi}{\varphi}} + 0.5 \rfloor, \varphi), \tag{5}$$

where *mod* is the modulo operation, and φ is the number of dominant orientations.

The dominant orientation bins of (x,y) and (x_i,y_i) can be obtained, here they are denoted as $q_{x,y}$ and q_{x_i,y_i}. Their operation is calculated as follows:

$$l_o(g_{x_i,y_i}, g_{x,y}) = q_{x,y} \bigoplus q_{x_i,y_i} = \begin{cases} 0, q_{x,y} = q_{x_i,y_i} \\ 1, q_{x,y} \neq q_{x_i,y_i} \end{cases}. \tag{6}$$

2.4.2 Learning statistical dominant patterns

Given training images, according to Equation 4, the local patterns of sign (or magnitude/orientation) are obtained. Here, they are denoted as $\vec{x}_i (i = 1, \dots, N)$, where N is the total number of pixels from these images. Codebook learning using k-means clustering usually requires much time and huge memory. One main reason is that local patterns that occur several times are calculated repeatedly in clustering. It therefore leads to great redundancy in calculation. To address this problem, we propose a revised vector quantization. In particular, we can obtain $\overline{N}$ unique local patterns $\hat{\mathbf{X}} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{\overline{N}}]$ and the number of occurrences of local pattern $\sigma = [\sigma_1, \sigma_2, \dots, \sigma_{\overline{N}}]$, where $\overline{N} \ll N$, *e.g.*, $\overline{N} = 256$ for 8-point sampling. The objective function of k-means clustering method is

Table 4. Comparison among methods on Infant COPE database, where $*$ represents our method (I, published by permission of Springer).

Method	AUC
EQP$_{\beta}$ (Nanni *et al.* 2010)	0.922
LTP$_{P,R}^{riu2}$ (Nanni *et al.* 2010)	0.77
LPQ (Nanni *et al.* 2010)	0.923
ENS (Nanni *et al.* 2010)	0.923
$dis(S+M)_{N,R}$ (Guo *et al.* 2012)	0.929
CLQP$_{S,M,O}*$	**0.935**

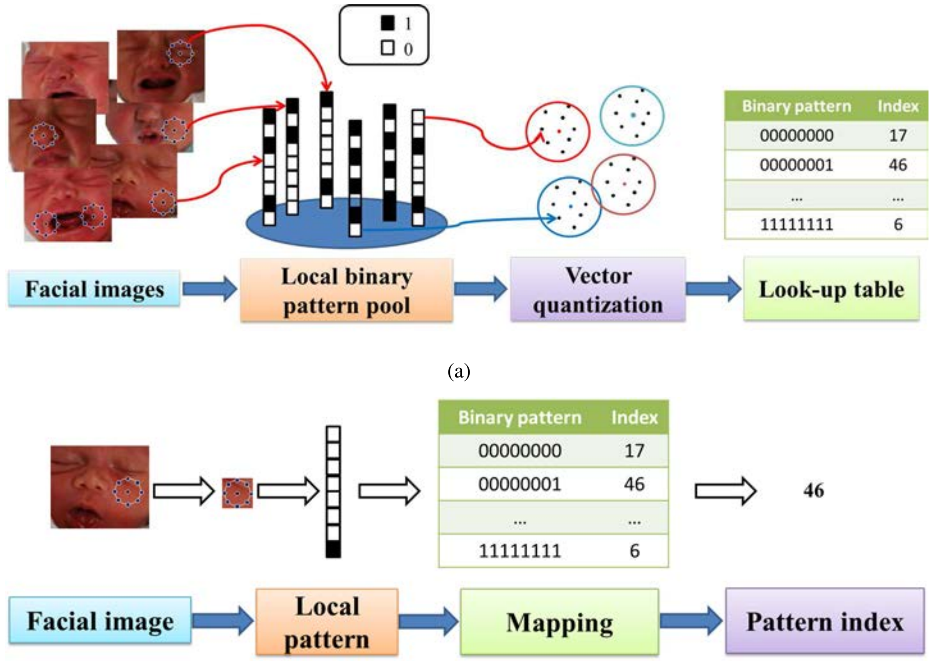
consequently re-written as

$$\hat{\psi}^* = \arg\min_{\hat{\psi}} \sum_{j=1}^{K} \sum_{\hat{x}_i \in \hat{\psi}_j} \|\sigma_i \hat{x}_i - \hat{\mu}_j\|^2. \tag{7}$$

Another important issue related to the efficiency of k-means is the initialization of the clustering centers. Different choices of the initialization substantially affect the speed of divergence and also the performance. Instead of random sampling, we exploit the dominant local patterns, *i.e.*, the most frequently occurred patterns as initialization. In implementation, we sort local patterns by the descending order of occurrence, and then select the first K local patterns as the initialization of the clustering centers. We empirically observe that only several iterations are required for the objective function (7) to converge. Finally, in order to guarantee the fast mapping, look-up table (LUT) is offline built by mapping local patterns to the nearest clustering centers. An example of codebook learning is shown in Figure 4(a).

2.4.3 Implementation

For all face images, the sign-based, magnitude-based and orientation-based pattern are obtained. They are fed into the revised vector quantization to learn the three separate codebooks. The advantage is that we can explore the discriminative and most representative patterns for each one. For each verified image, we can get their histograms through mapping three patterns into the corresponding codebook, as shown in Figure 4(b). The histograms are finally concentrated into one histogram. In our approach, we make use of L^1-norm to normalize the final histogram. The final feature are denoted as CLQP$_{S,M,O}$.



Binary pattern	Index
00000000	17
00000001	46
...	...
11111111	6

(a)

(b)

Fig 4. An example of completed local quantized patterns. (a) *Codebook learning*: It extracts a local pattern of each pixel, and then these patterns and corresponding frequency vectors are fed into k-means clustering. (b) *Local pattern encoding*: It maps each local pattern into the look-up table.

2.4.4 Experiments

Experiments on texture classification and facial expression analysis are conducted to verify the performance of CLQP. Here, the results of facial expression analysis are described. More experiments about texture classification can be referred to the experiment section in Paper I.

Recent research on facial expression analysis has provided a protocol for diagnosing the pain of patient, especially for neonates who are incapable of articulating their pain experiences. We conducted experiments on the COPE database previously described in Section 1.2. For pain detection, face images of non-pain states (rest, cry, stimulus, friction) are combined to form a single class, and the ones of pain states are regarded as the other class. All experiments are conducted under leave-one-subject-out protocol and non-linear SVM with the Gaussian radial basis function kernel. The optimal values of kernel and cost parameters were determined using the grid search strategy, where the optimal values of kernel and cost parameters were searched exponentially in the ranges of $[2^{-15}, 2^{15}]$ and $[2^{-5}, 2^{10}]$. Considering the contour of the face, we divide face images into 2×2 regions. Our proposed descriptor with eight sampling points and the radius of 2 is used. The area under the ROC-curve (AUC) of different methods is presented in Table 4. From experimental results, our method achieves the highest AUC value among all methods under comparison.

2.5 Spatiotemporal monogenic binary patterns

Although LBP-TOP is an efficient descriptor for facial expression recognition, the operator on the pixel-level is not always stable for some noise caused by illumination variations or pose changes. In addition, the orientation in CLQP as previously mentioned can provide much information for facial expression recognition. Therefore, Paper II proposed a new two-layer method for enhancing the stability of LBP-TOP, as shown in Figure 5. In the first layer, an effective multi-scale monogenic signals analysis is employed as the first hidden layer. In the second latent layer, LBP is used to encode monogenic magnitude in three orthogonal planes (TOP), and the combination of Phase-quadrant Demodulation Coding (PQDC) and Local Exclusive OR operator (LXP) is used to encode the real and imaginary pictures of the orientation in TOP.

2.5.1 2-D multi-scale monogenic signal analysis

The monogenic signal generalizes the analytic signal to 2-D by the introduction of a Riesz filter $f(x,y)$, in which Fourier domain representation is $[\mathscr{F}_1(u,v), \mathscr{F}_2(u,v)]$. It can be used to interpret local phase, local orientation and local magnitude of an image in a rotation invariant way. Given an image $I(x,y)$, the monogenic signal is represented as the convolution of 2-D signal with two Riesz filtered components $f(x,y) = \hbar_s(\omega) * I(x,y) * [1, f_1(x,y), f_2(x,y)]$, where "$*$" is convolution operator, $f_i(x,y)$ is the spatial domain representation of $\mathscr{F}_i(u,v)$, and $\hbar_s(\omega)$ is a Log-Gabor function on the linear frequency scale s. Here, the Log-Gabor function has a transfer function (Field 1987) of the form

$$\hbar_s(\omega) = \mathscr{F}^{-1}(G(\omega)) = \mathscr{F}^{-1}(\exp^{(-\log(\omega/\omega_c)^2)/2(\log(k/\omega_c)^2)}), \tag{8}$$

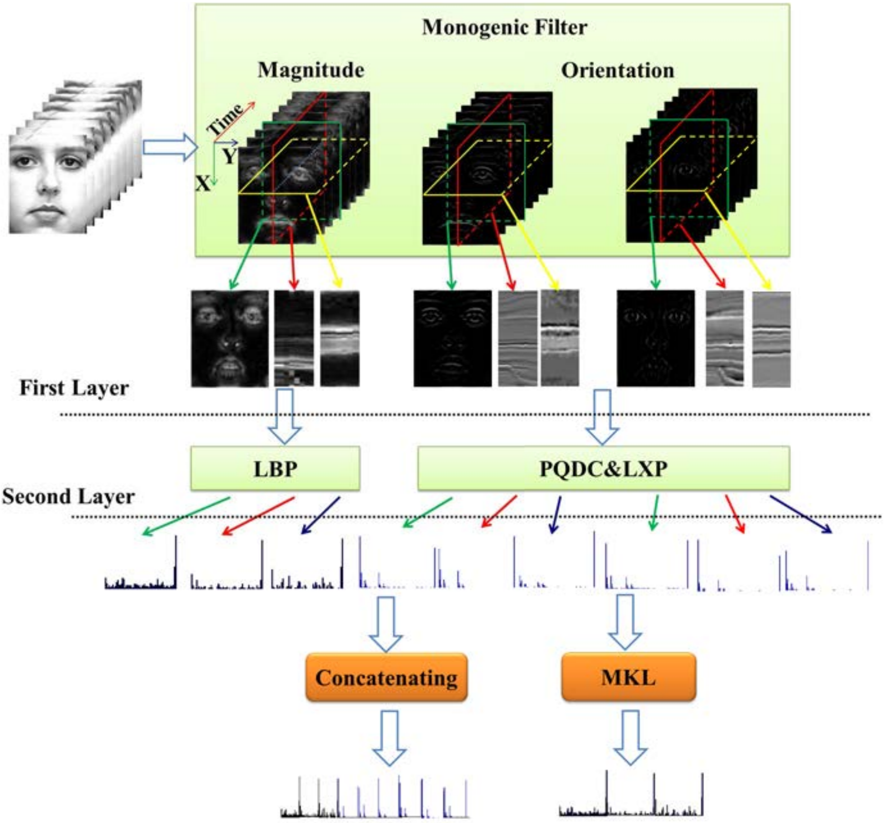


Fig 5. The proposed two-layer architecture for dynamic facial expressions.

where $\omega_c \propto s$ is the filter's center frequency and s is the number of wavelet scale of which a lower value will reveal more fine scale features, while a larger value will highlight coarse features.

Since the log-Gabor filters are band-pass filters, multi-scale monogenic representation is thus required to fully describe a signal. It is known that multi-scale representation for the sequence may cause high dimensionality and expensive computational time. For the trade-off between computational complexity and performance, the sum of multi-scale method is used to re-define the monogenic magnitude, phase and orientation as follows:

$$\widetilde{A} = \sqrt{\widetilde{f}_0^2 + \widetilde{f}_x^2 + \widetilde{f}_y^2}, \tag{9}$$

$$\widetilde{\phi} = \arctan(\sqrt{\widetilde{f}_x^2 + \widetilde{f}_y^2} / \widetilde{f}_0), \tag{10}$$

$$\widetilde{\theta} = \arctan(\widetilde{f}_x / \widetilde{f}_y), \tag{11}$$

where $\widetilde{f_0} = \sum_s \hbar_s(\omega) * I(\cdot)$, $\widetilde{f_x} = \sum_s \hbar_s(\omega) * I(\cdot) * f_1(\cdot)$, $\widetilde{f_y} = \sum_s \hbar_s(\omega) * I(\cdot) * f_2(\cdot)$.

2.5.2 Encoding procedure

For each facial image, it can be represented by two elements, *i.e.*, magnitude and orientation of its monogenic signal. For simplicity, the orientation information can be decomposed into the real and imaginary pictures. Magnitude and two sub-elements of orientation are further encoded by using the similar ways to LBP, respectively. These information can provide the robustness for LBP to noise and spatial translation, *etc*. Here, it begins with the investigation of the histogram for the still image. Three feature descriptors based on magnitude, real and imaginary pictures are denoted as local monogenic magnitude binary pattern (LMMBP), local monogenic real binary pattern (LMRBP), and local monogenic imaginary binary pattern (LMIBP), respectively.

LMMBP: In the monogenic representation of the image, the magnitude measures local structure energy, and the LBP operator can then be used to encode the variation of local energy. The histogram descriptor can thus be formulated as

$$h_{P,R} = \sum_{i=0}^{P-1} L((\widetilde{A}_{x_i,y_i} - \widetilde{A}_{x_c,y_c})) \geq 0)2^i, \tag{12}$$

where $\widetilde{A}_{x_c,y_c}$ is the monogenic magnitude value at the position (x_c, y_c), $\widetilde{A}_{x_i,y_i}$ is the magnitude value of P equally spaced pixels on a circle of radius R at this position.

LMR/IBP: Since the orientation is an important feature for indicating the dominant direction of local image variation, it is expected that it can then be used to provide complementary information. Different from the way in magnitude, PQDC is exploited to encode the orientation information into a quadrant bit. As previously described, the orientation is decomposed into the real and imaginary parts. Here, it takes the real part as an example for showing the encoding procedure. The same way is used regarding the imaginary part. The encoding measurement for two parts can be formulated as follows:

$$B(\widetilde{f_i}) = \begin{cases} 1, \widetilde{f_i} > 0 \\ 0, \widetilde{f_i} \leq 0 \end{cases}, \tag{13}$$

where $i = x$ for a real picture, and $i = y$ for an imaginary picture. For a brief description, $\widetilde{f_i}$ is denoted as $\widetilde{f}$.

Then, an alternative binary-compare function, LXP, is exploited to calculate the correlation of the center and its neighbors. The histogram of real/imaginary can be formulated as follows:

$$\hat{h}_{P,R} = \sum_{i=0}^{P-1} (B(\widetilde{f}_{x_i,y_i}) \bigoplus B(\widetilde{f}_{x_c,y_c}))2^i, \tag{14}$$

where $B(\widetilde{f}_{x_c,y_c})$ is the PQDC value of P equally spaced pixels on a circle of radius R at the position (x_c, y_c), and $\bigoplus$ denotes the bit exclusive or operator. Furthermore, the uniform pattern in LBP, which contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular, is used to preserve a simple rotation-invariant property and reduce the length of the feature vector.

Extension to temporal domain: It is known that a video sequence can provide much more information than the static image. This induces another problem how to extend these methods to video sequences. Regarding this problem, it borrows the nature of LBP-TOP. Firstly, the magnitude, the real picture, and the imaginary picture on facial expression images are orderly put into three cuboids, respectively. Secondly, the uniform LBP operator is used to calculate the histograms in TOP for sequential monogenic magnitude in a video. And then all histograms are concatenated into a new histogram. Here, it is denoted as spatiotemporal local monogenic magnitude binary pattern (STLMMBP). The procedure of LMRBP and LMIBP are utilized in TOP. It also obtains three histograms from three planes. Finally, they are also concatenated into one histogram. Here, they are denoted as spatiotemporal local monogenic real binary pattern (STLMRBP) and spatiotemporal local monogenic imaginary binary pattern (STLMIBP), respectively.

Feature fusion: To the end, feature fusion is an important issue in classification. Here, two methods are considered for fusing STLMMBP, STLMRBP, and STLMIBP histograms. Three histograms are (1) concatenated into one histogram (STLMBP-C) or (2) fused by MKL (STLMBP-MKL).

2.5.3 Experiments

To evaluate the performance of the proposed method, the CK+ (Lucey *et al.* 2010) and Oulu-CASIA NIR&VIS databases are used. To fairly compare with the previous works, multi-class SVM (Franc & Hlavác 2001) is adopted to learn the expression classifiers for all methods except STLMBP-MKL.

Table 5. Comparison among methods on CK+ database, where $*$ represents our methods (II, published by permission of IEEE).

Method	Average Recognition Rate (%)
Baseline (Lucey *et al.* 2010)	88.38
LBP-TOP (Zhao & Pietikäinen 2007)	90.8
LPQTOP (Jiang *et al.* 2011, Päivärinta *et al.* 2011)	89.5
VLPQ (Päivärinta *et al.* 2011)	91.8
STGabor (Petkov & Subramanian 2007)	81.54
STLMBP-C*	**92.31**
STLMBP-MKL*	**92.62**

CK+ database

CK+ database with 325 video sequences from 118 subjects was selected from the set of databases for evaluation. Here, the leave-one-subject-out protocol was employed for all methods. The comparative performance of all methods is shown in Table 5. It can be seen that STLMBP-C outperforms Spatiotemporal Gabor filters for motion processing (STGabor), LBPTOP, local phase quantization from three orthogonal planes (LPQTOP), and volume local phase quantization (VLPQ). Furthermore, the application of MKL (Sonnenburg *et al.* 2006) can achieve a higher degree performance than STLMBP-C. Here, we also compare the baseline result in Lucey *et al.* (2010) with our results, where they used the similarity-normalized shape and canonical appearance to obtain the recognition rate of 88.38%. Our two methods (STLMBP-C and STLMBP-MKL) outperformed it by about 3.93% and 4.24%, respectively.

Oulu-CASIC NIR&VIS database

The facial images are divided into 9×8 overlapping blocks, where the overlapping ratio is 43%. Here, a ten-fold-cross-validation scheme was used for the evaluation. LBP-TOP as a comparative method was implemented to obtain the results for three illuminations. The comparison is shown in Table 6. As seen, combining all information can boost the performance of a separate feature in most of illuminations.

Table 6. Comparison among methods on Oulu-CASIC NIR&VIS database, where $*$ represents our methods (II, published by permission of IEEE).

Method	Illumination	Recognition Rate (%)
LBP-TOP (Zhao & Pietikäinen 2007)	VIS_N	74.45
LBP-TOP (Zhao & Pietikäinen 2007)	VIS_W	60.84
LBPTOP (Zhao & Pietikäinen 2007)	VIS_D	57.49
STLMBP-MKL*	VIS_N	79.95
STLMBP-MKL*	VIS_W	64.55
STLMBP-MKL*	VIS_D	61.96
STLMBP-MKL*	NIR_N	78.61
STLMBP-MKL*	NIR_W	70.34
STLMBP-MKL*	NIR_D	72.23

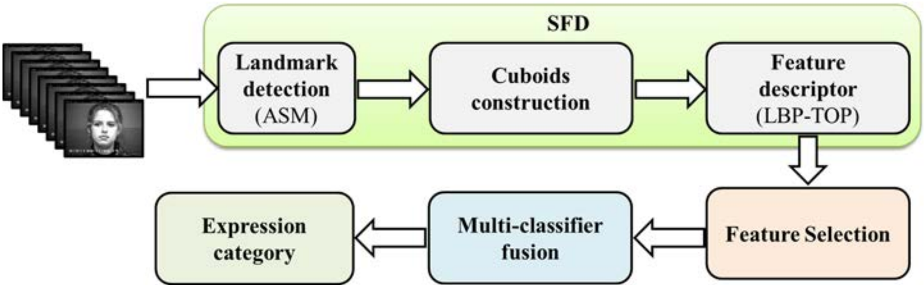


Fig 6. The proposed new facial expression system, which consists of feature descriptor and multi-classifier fusion.

2.6 Sparse spatiotemporal features and multi-classifier fusion

Paper III proposed a new facial expression recognition system, as shown in Figure 6. It consists of a novel spatiotemporal feature descriptor based on active shape model (ASM) and multiple classifier fusion. In this method, 38 important facial interest regions based on prior information are first determined, subsequently, the spatiotemporal feature descriptor is used to describe facial expressions from these areas. Furthermore, AdaBoost is used to select the most discriminative features for 114 slices (there are three slices or planes in one block volume for LBP-TOP). In this case, one slice represents a 59-bin histogram. In the classification step, a framework for combining voting results from several classifiers is presented.

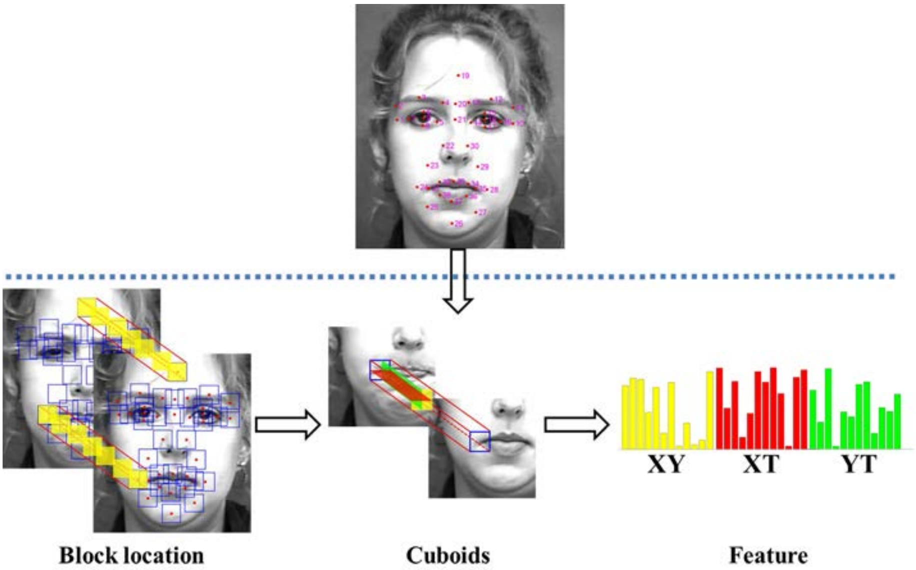


Fig 7. The proposed sparse spatiotemporal feature descriptor, where the number around the each point is the index of facial landmark (III, published by permission of Springer).

2.6.1 Sparse feature descriptor

An appearance-based feature can work well in illumination variation and spatial shifting, while a geometry-based feature can mostly focus on the interesting points in several important regions for facial expression recognition, like eyes and mouth, and it can also reduce the redundancy information that may be caused by appearance-based feature descriptors. Following Zhang *et al.* (1998), the interest point structure which includes 38 points is designed, as shown in Figure 7. The facial points can be obtained by ASM (Milborrow & Nicolls 2008). It is seen that these facial landmarks can accumulate in the specific regions, such as mouth, cheek, nose and eyes. It can help the system remove the face boundary and does not suffer from the influence of face identify. Then, a spatiotemporal feature descriptor from facial points of interest is developed. It has lower computational cost and more discriminative property. Details of implementation of our method in facial expression recognition are described as follows:

(1) Block location: Based on 38 facial interest points, the areas centered at these points have more discriminative information. From these areas, they could cover the

majority of features of eyes, mouth, cheek and forehead. The size of areas plays an important role in feature extraction. The best sizes of areas are chosen experimentally.

(2) Constructing cuboids: Given the size of area w and the time length T of an image sequence, ASM is used to detect the first frame. And then one cuboid goes through all images from the first frame for each facial point, where the size of each cuboid is $w \times w \times T$.

(3) Feature description: LBP-TOP is implemented to describe the motion feature and spatial feature of 38 cuboids. For each cuboid, we can obtain three uniform LBP histograms $\vec{x}_{\mathrm{XY}}$, $\vec{x}_{\mathrm{XT}}$ and $\vec{x}_{\mathrm{YT}}$ from XY, XT and YT planes, respectively. In the implementation, all radii and neighbor number of three planes are set as 3 and 8, respectively. Three histograms in a cuboid are concatenated into one feature vector of 177 dimensions.

(4) Implementation: Above mentioned, the histograms of all cuboids are concatenated into a single feature, and then it can be fed into the classifier, *e.g.*, SVM. Here, this feature descriptor is abbreviated as SFD.

2.6.2 Discriminative information enhancement

Though our system considers 38 facial regions preserving sufficient information for recognizing facial expressions, it is observed that not all facial regions play equally important roles in each facial expression (Kotsia *et al.* 2008a). For example, the information from mouth region can provide discriminative information to “happy” expression. It is necessary to select the most representative feature for intra-expression and inter-expression. Additionally, this procedure can further compress the dimension of the original feature. The concept of intra-expression similarity as well as extra-expression dissimilarity and feature selection are employed, where feature selection is based on AdaBoost theorem (Freund & Schapire 1997). The detail of feature selection can be referred to Zhao & Pietikäinen (2009).

Firstly, the inter-class dissimilarity and extra-class similarity are defined as follows: Given two slices of 59-bin histogram $\vec{x}_i$ and $\vec{x}_j$, their distance metric can be given as

$$\chi^2(\vec{x}_i,\vec{x}_j) = \sum_{m=1}^{D} \frac{(\vec{x}_i^m - \vec{x}_j^m)^2}{\vec{x}_i^m + \vec{x}_j^m}, \tag{15}$$

where D is the number of bins of histogram.

Secondly, for two facial video sequences, the dissimilarity of LBP-TOP histogram

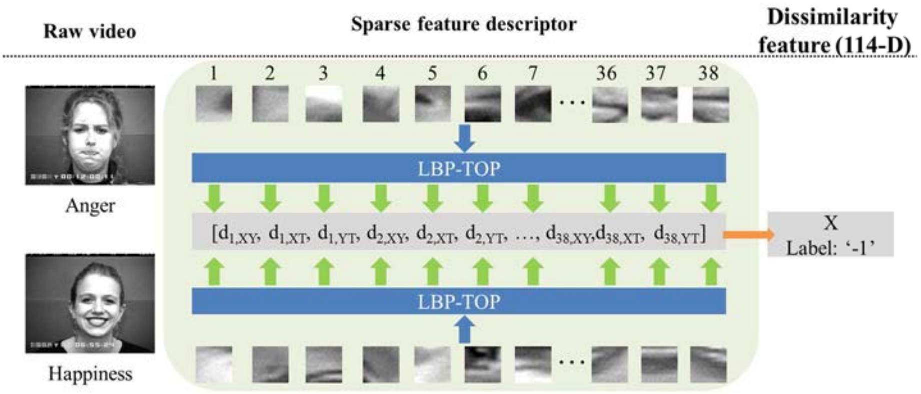

Raw video
Sparse feature descriptor
Dissimilarity feature (114-D)
1
2
3
4
5
6
7
36
37
38
LBP-TOP
[d1,XY, d1,XT, d1,YT, d2,XY, d2,XT, d2,YT, …, d38,XY,d38,XT, d38,YT]
X
Label: '-1'
LBP-TOP
Anger
Happiness

Fig 8. The procedure of generating a 114-D dissimilarity feature for two facial video sequences, where X represents the dissimilarity feature.

on each plane can be calculated by using Equation (15). In this case, for each region, there are 3-D distance vector $[d_{i,XY}, d_{i,XT}, d_{i,YT}]$. As for the feature concerned, the dissimilarity feature can be formulated by a 114-D vector. This dissimilarity features are further labeled '1' if two facial video sequences have the same class, otherwise '-1'. The procedure is shown in Figure 8. For any expression-pair, there are large-scale dissimilarity features concerned the spatial and motion information. Next, these dissimilarity features and class information are fed into a weak learner. The weights for 114 slices can be obtained after several rounds. Thus, the AdaBoost algorithm sorts and selects slices for discriminating two classes. After feature selection, the most discriminative slices can be selected. The raw features from 38 facial regions can be further filtered, and also have more discriminative ability. In brief, the discriminative feature descriptor is named as DisSFD.

2.6.3 Multi-classifier fusion

Given the feature of a facial expression video clip $\vec{x}$ and N classifiers, which are denoted as $\mathscr{C}_i, i = 1, \ldots, N$, the output of the ith single classifier can be approximated by a posteriori probabilities as

$$Pr(y = c|\mathscr{C}_i) = Pr(y = c|\vec{x}) + \varepsilon(\vec{x}), \tag{16}$$

where $Pr(y = c|\vec{x})$ represents the probability of the sample belonging to the class c, $\hat{c}$ is the prediction label, and $\varepsilon(\vec{x})$ represents the error that a single classifier introduces.

According to Bayesian theory, the sample should be assigned to the class $\hat{c}$ provided that the posteriori probability of that interpretation is maximum:

$$Pr(y = \hat{c}|\vec{x}, \mathscr{C}_1, \ldots, \mathscr{C}_N) = max_{c \in 1,\ldots,C} Pr(y = c|\vec{x}, \mathscr{C}_1, \ldots, \mathscr{C}_N). \tag{17}$$

For exploiting the complementary information among all classifiers, we investigated three decision rules (mean rule, product rule, and median rule). The comparative results of these three rules will be explained in the experiments. Assume that all classifiers used are generally statistically independent, and the prior probabilities of occurrence for the class k are under assumption of equal priors, the rule of multi-classifier fusion is simplified to

$$Pr(y = \hat{c}|\vec{x}) = max_{c \in 1,\ldots,C} [\otimes_{\vec{x} \in 1,\ldots,m} Pr(y = c|\vec{x}, \mathscr{C}_1, \ldots, \mathscr{C}_N)], \tag{18}$$

where C is the number of facial expressions, $\otimes$ represents one decision-level fusion rule, and $Pr(y = c|\vec{x}, \mathscr{C}_i)$ can be obtained by the soft-max function. Here, multi-classifier fusion is shorted as MCF. In the implementation, five classifiers, including SVMs based on linear, Gaussian and Polynomial kernels, boosting classifier and Fisher linear discriminant classifier are chosen.

2.6.4 Experiments

For our study, 374 sequences are selected from the Cohn-Kanade facial expression database for basic facial expression recognition. The selection criterion was that any sequence to be labeled was one of the six basic emotions (anger, disgust, fear, joy, sadness, and surprise). The sequences came from 97 subjects, with one to six emotions per subject. Ten-fold cross validation method was used in the whole scenario.

Firstly, we evaluate the performance of SFD with different block sizes. We give the mean accuracy of 8×8, 16×16, 32×32 and 64×64 block sizes. The mean accuracies are 87.97%, 94.92%, 94.12% and 93.85% for 8×8, 16×16, 32×32 and 64×64, respectively. SFD obtains the compromising result at 16×16 block size. Furthermore, based on MCF with median rule, SFD+MCF achieves the performance of 95.19% at the accuracy.

Next, the performance of DisSFD is evaluated. We give the mean accuracy of the different numbers of slices from 15 to 90 at the interval of 15. The mean accuracies are 90.37%, 91.98%, 94.12%, 93.32%, 93.05% and 92.25% for 15, 30, 45, 60, 75 and 90 slices, respectively. Unfortunately, DisSFD cannot be better than the SFD (94.92%), while the dimensionality of features is effectively decreased from 6,726 to 4,425.

Table 7. Comparative performance of the proposed methods and the state-of-the-art methods, where $*$ represents our methods (III, published by permission of Springer).

Method	Evaluation Protocol	Average Recognition Rate (%)
Yesin *et al.* (2006)	Five-fold	90.9
Shan *et al.* (2009)	Ten-fold	88.4
Aleksic & Katsaggelos (2006)	-	93.66
Littlewort *et al.* (2006)	Leave-one-subject-out	93.8
Zhao & Pietikäinen (2007)	Ten-fold	91.44
SFD*	Ten-fold	**94.92**
DisSFD*	Ten-fold	**94.12**
SFD+MCF*	Ten-fold	**95.19**
DisSFD+MCF*	Ten-fold	**96.32**

Finally, we combine all methods into the evaluation, where 16×16 block size and 45 slices are chosen for DisSFD, and MCF uses the median rule. It is good to see that DisSFD+MCF* can achieve the performance of 96.32%, which is better than disSFD and SFD+MCF. Finally, the comparative results of the proposed method with the state-of-the-art methods (Yesin *et al.* 2006, Shan *et al.* 2009, Aleksic & Katsaggelos 2006, Littlewort *et al.* 2006, Zhao & Pietikäinen 2007) are given in Table 7, where they provided the overall results obtained with the same database. From this table, we can see that SFD obtained better result than block-based LBP-TOP that divided face image into 8×8 overlapping blocks, with an increase of 3.48%. Additionally, SFD+MCF is better compared to SFD and DisSFD. DisSFD+MCF outperformed all the other methods.

2.7 Discussion

The method presented in Section 2.4, also in Paper I, is designed for facial expression recognition in still images. In order to enhance the representative capability of LBP, the completed local pattern method is designed for image description based on the difference of sign, magnitude and orientation. Additionally, to learn the discriminative and representative patterns for appearance and motion features, a simple yet powerful measurement is formulated. There are two parts of this measurement: first, predefined domain patterns are calculated in order to get the efficient computation, and second, the weight k-means clustering trickily learns the number of clusters of all patterns.

The contributions of this methods can be summarized as: (1) the statistical patterns are flexible and domain to various applications and (2) the features are compact. The solid experimental results in Paper I have shown the advantage of this method: (1) domain-adaptive and flexible statistical and (2) compact features. A limitation is that the feature extraction is on still images. Even so, it may be possible to extend it to the temporal domain by following the principle of LBP-TOP.

In the multi-layer spatiotemporal descriptor presented in Section 2.5 (Paper II), the useful monogenic filters as the first hidden layer are designed for the purpose of enhancing the representation of facial features, and then the LBP operator as the second latent layer was used to describe the facial features by local structure information and histogram of sub-regions. The contributions of this method lie in the following aspects: (1) the shape of facial images has been preserved in magnitude and orientation information; (2) the facial structures with respect to noise and illumination changes are suppressed by analyzing the local phase, which is a qualitative measure of a local structure; (3) the final features encode both local structure information and shape information; and (4) the representation capability of features is straightforwardly strengthened by learning the optimal weights for three histogram features. One of the derived descriptors combining with magnitude and orientations consistently achieves superior classification performance over all methods under comparison for facial expression recognition application. For this multi-layer spatiotemporal descriptor, we found that one of its limitations is that the computational time is slightly longer than that of the conventional LBP-TOP. However, as observed from experimental results, the improvement on recognition accuracies is more significant. Additionally, the activities of the monogenic filters may lead to different classification accuracies, since the scale among the controlled parameters of monogenic filters would give the fine or coarse structure of facial images. But it is experimentally observed that it gets the compromising achievement when the scale is 3. It also gets the tradeoff between the computational complexity and performance for facial expression recognition.

The method presented in Section 2.6 (Paper III) can be seen as the combination of geometric and appearance features. A slight difference to other hybrid feature methods is that the geometric information provides the regions of interest to the appearance descriptor. To enhance the discriminative capability of features, the feature selection based on AdaBoost is designed for 114 slices, including the appearance and motion features. However these discriminative features observed from the experiments cannot achieve better performance while they reduce the dimensionality. Finally, the multi-classifier fusion method is exploited to utilize the power of different classifier to enhance the final features. From the experimental evaluations on facial expression database it is found that: (1) the sparse features contain robust and reliable discriminative information which can improve the classification performance; (2) the feature selection significantly preserves the effectiveness of computation; and (3) the multi-classifier fusion may provide the potential implementation to facial expression recognition.

3 Facial expression recognition in uncontrolled conditions

This chapter presents the studies on facial expression recognition in uncontrolled conditions, including illumination variations, partial occlusion and head pose variation, which are originally presented in Papers IV-VII.

3.1 Background and motivation

In recent years, the experiments or applications among numerous approaches for facial expression recognition have been conducted in restricted experimental environment. For example, participants not wearing sunglasses sit in frontal of the camera under good illumination. In such a satisfying condition, a facial expression recognition system by Yesin *et al.* (2006), Aleksic & Katsaggelos (2006), Littlewort *et al.* (2006), Zhao & Pietikäinen (2007), Shan *et al.* (2009), for examples, including our previously proposed methods, can obtain promising performance. In general, these experimental conditions can be easily pre-acquired, *e.g.*, by requiring participants not to wear sunglass, providing a good lighting or using a high-resolution camera. An experimental setup such like that can be referred to Lucey *et al.* (2010), Gross *et al.* (2010).

However, facial expression recognition may suffer from some practical problems when it is combined with surveillance systems of airport, station, for instance. The rigorous conditions cannot be simply controlled in real-time environment. For example, people deliberately cover his/her face by hand, or people do not face to the camera. Even tiny changes of environment may expose numerous serious problems to the application of facial expression recognition. Three critical factors, including illumination, occlusion and pose, are well known to make serious challenges while applying facial expression recognition. Therefore, this thesis focuses on the classification problem of facial expressions on three previously described conditions.

Previous studies have suggested that the component-based approach can provide promising performance in some cases against pose motion or partial occlusion (Heisele & Koshizen 2004, Ivanov *et al.* 2004). Following our previous work in Section 2.6 (Paper III), this section begins with presenting two variants of the component-based method, which will be used to increase the robustness of features for illumination variation and occlusion. And then it continues to deal with the illumination variation in facial expression recognition, and handles the occlusion, finally presents our solution to pose changes.

Illumination variation

Most of facial expression data sets currently in use are captured in a visible light spectrum. However, different environment and time can cause significant variations between images. The variation of lighting conditions, and light angles in particular, change the appearance of the face in a significant way (Adini *et al.* 1997). Therefore, a facial expression recognition system should adapt to the environment, not vice versa. However, uncontrolled visible light (VIS) (380-750 mn) in ambient conditions can cause significant variations in image appearance and texture. The facial expression recognition method developed thus far by us and others performs well under controlled circumstances, but changes in illumination or light angle cause problems for the recognition systems. To meet the requirements of real world applications, it should be possible to work in varying illumination conditions and even in near darkness.

To date, there are many image processing methods (Li *et al.* 2008, Tan & Triggs 2010, Nabatchian *et al.* 2011) to handle the illumination changes. Unfortunately, algorithms are complicated and not very reliable, *e.g.*, for different lighting directions, and using the same preprocessing could not get satisfying results. Instead, there are certain convenient imaging sources to capture the face image in different light spectrum such as the thermal spectrum (8-12 μm) and near-infrared spectrum (780-1100 nm). The thermal spectrum imaging becomes a useful technique for recognizing facial expression under uncontrolled illumination condition (Yoshitomi *et al.* 1997, He *et al.* 2013), but thermal spectrum imaging has limitations in the following situations: (1) wearing glasses blocks a large portion of thermal energy of eye region, (2) variations of ambient temperature also significantly change the thermal characteristics of the face, and (3) some facial regions are not receptive to the emotion changes under thermal spectrum. (Kong *et al.* 2005, Ioannou *et al.* 2014, Nguyen *et al.* 2014). Additionally, Chen *et al.* (2005) showed that a thermal infrared system does not work in practice as well as a system based on VIS images, since elapsed time causes significant changes in the thermal patterns of the same subject.

In contrast, active near-infrared (NIR) has become a candidate for dealing with the illumination variations for facial expression recognition. As far as we know, active

NIR imaging is robust to illumination variations, and it has been used successfully for illumination invariant face recognition (Li *et al.* 2007). In their study, they firstly stated that the face images of the same subject obtained from a visual camera are negatively correlated, while ones obtained from NIR imaging in diverse visual illumination conditions are closely correlated. Because of the changes in the lighting intensity, NIR images are inclined to monotonic transform. They further compensated the monotonic transform by applying the LBP operator to NIR images because LBP is invariant with respect to monotonic gray scale changes. The combination of NIR imaging and LBP features obtained promising results for illumination invariant face recognition. Therefore, the robustness properties of NIR imaging may provide a good basis for facial expression recognition regardless of variations in VIS lighting. In Section 3.4, we present a new method that combines the NIR method and discriminative component-based LBP-TOP features beyond visual spectrum for illumination robust facial expression recognition.

Facial occlusion

Scientists conducted experiments in order to recognize the facial expressions from un-occluded facial images taken under controlled laboratory conditions. The facial expression recognition system developed by Zhao & Pietikäinen (2007), Shan *et al.* (2009), for examples, thus has a good ability to recognize the facial expression without presence of occlusion. Unfortunately, at times, the human may be wearing sunglasses or scarf; thus, the face may be partially occluded. In practice, the traditional facial expression recognition system suffers from the problems by the presence of occlusion, *e.g.*, “happiness” expression and “surprise” expressions are confused in case of mouth occlusion. Therefore, a practical facial expression recognition system should handle some common types of presence of occlusion.

However, most of recent work on recognizing facial expression in this circumstance has focused on still images. It is known that one can see emotion from an image sequence easier than from a still image. One reason for that is that the information from the image sequence contains not only the appearance feature, but also the motion feature. Due to this reason, the analysis of dynamic image sequences has become very attractive in computer vision. It motivates us to analyze facial occlusion of a dynamic facial expression sequence. In our observation, we find that un-occluded face regions can subsequently provide sufficient information to the classifier. Therefore, we try to develop a novel algorithm to combine component-based method and sparse representation to handle facial occlusion. The component-based approach in Paper V is exploited to represent facial expressions via two kinds of feature descriptors. Based on the component-based method, occlusion detection is discussed in Section 3.5. In the occlusion detection, we design a novel occlusion detection based on the theorem of sparse representation to find out the position of occluded regions. In the recognition stage, the weight learning and multiple feature fusion are used to learn the optimal weights for each feature and then fuse them to classify each image sequence to the expression class.

Pose change

Besides illumination and occlusion, the head pose makes a challenging problem in facial expression recognition. As previously mentioned, the facial expression recognition designed in recent years performs well under controlled environments, in which human subjects need to face the camera. However, in the real-world application, it is difficult to acquire the face images in frontal of the camera, *e.g.*, in the meeting or people talking. In this case, the existing system may fail to recognize the facial expression with the head pose change more than 30 degrees. It is important to make the system able to do the recognition at arbitrary views.

With the creation of several multi-view databases, such as Multi-PIE, the view-invariant approaches have been under investigation by the research community in facial expression recognition. The works on profile view expression recognition can be classified based upon the types of features employed: geometric features and various low-level features on pre-labeled landmarks points. It is noted that these approaches require the facial key-points location information, which needs to be pre-labeled. However, in real applications, key-points need to be automatically detected, which is a big challenge in itself in case of non-frontal faces, although multi-view 2D tracking methods (Zhu & Ramanan 2012, Anvar *et al.* 2013) can be used to register facial images. To address this issue, there have been some attempts which do not require key-point locations (Zheng *et al.* 2010, Tang *et al.* 2010, Moore & Bowden 2011).

Motivated by the existing research works, there are two questions that should be considered in multi-view facial expression recognition. The first issue is that whether there exist discriminative features for facial expressions or not. Recent research shows that appearance-based features achieve good performance, while they may carry irrelevant information of face identity. This irrelevant information can confuse the classifier for recognizing facial expressions. The other is addressed for utilizing the correlation between facial expressions and views. The classical framework regards the view and facial expression as two separate signals. It usually requires to accurately estimating pose. It therefore leads to a challenge to this framework. To address this, the exploration of their correlation may be a good way to make one system simple and avoid the cumulative error caused by pose estimation. The solution to these two issues is discussed in Section 3.6. Discriminative neighbor preserving embedding (DNPE) was proposed by using neighbor graph and maximizing margin criterion. Additionally, the multi-view framework was used to embed DNPE for coupling facial expressions with views.

In summary, we aim to design stable and reliable image descriptors for dealing with illumination variation, partial occlusion and pose changes, which are discussed in Sections 3.3, 3.4, 3.5 and 3.6.

3.2 Related works

For facial expression analysis, the existing challenging situations in the wild can be divided into three uncontrolled circumstances: illumination variations, the presence of occlusion, and head pose changes. So far, few complete research surveys have summarized these three issues at the same place. In order to clarify the purpose of this chapter, the literature in this section of facial expression recognition is orderly discussed under three respective conditions. A brief overview of representative methods for studying illumination variation, the presence of occlusion and head pose changes are summarized in Tables 8, 9 and 10, respectively.

Illumination variation

In the applications, uncontrolled visible light in ambient conditions creates serious difficulties in extracting efficiently facial representation from image appearance and texture. According to Adini *et al.* (1997), they pointed out that the lighting condition significantly changes the appearance of a face, and also stated that the local filters are themselves inadequate to overcome variations in the environmental lighting due to changes in the illumination direction. Therefore, recognition in uncontrolled indoor illumination conditions is one of the most important problem for practical facial expression recognition systems. It is possible to solve this problem by using three

Table 8. Representative methods for resolving the problem by illumination variations.

Reference	Type	Methods
Li *et al.* (2007)	Infrared imaging	LBP features on NIR and VIS imaging
Li *et al.* (2008)	Illumination correction	Face model construction and model fitting
Tan & Triggs (2010)	Illumination normalization	Gamma correction, DoG filtering, mask and equalization of variation
Hu (2011)	Illumination normalization	Dual-tree complex wavelet transform in logarithm domain
Maeng *et al.* (2011)	Infrared imaging	NIR imaging
Nabatchian *et al.* (2011)	Illumination invariation extraction	Illumination-reflection model by applying a high-pass filter on the logarithm of an image
Cao *et al.* (2012)	Feature extraction	Neighboring wavelet coefficients
Tzimiropoulos *et al.* (2012)	Feature extraction	Image gradient orientation
He *et al.* (2013)	Infrared imaging	Thermal infrared imaging and deep Boltzmann machine

different approaches. The first tries to normalize illumination variations, the second attempts to exploit the features, and the third takes the advantage of different imaging systems.

(1) Much work has been attempted to model and correct illumination changes on face in VIS images. The idea in illumination normalization is to remove unwanted illumination effects from the image, such as non-uniform illumination, highlights, shadowing, aliasing, noise and blurring. (Li *et al.* 2008, Tan & Triggs 2010, Hu 2011, Nabatchian *et al.* 2011). One of the disadvantages of the illumination normalization is that when effects caused by varying illumination are removed, also some useful information like wrinkles and skin detail will also vanish. Illumination normalization methods are not very reliable, because also some important information for recognition will be lost when removing effects caused by illumination variations. The use of illumination normalization methods is shown to improve recognition performance when there are illumination variations in the faces, but have not led to illumination invariant face representation due to significant difficulties, especially uncontrolled illumination directions.

(2) Alternative ways to suppress the illumination changes on face in visual spectrum are done through exploiting robust features. The essential in illumination suppression is to extract the useful features from intensity, *e.g.*, the orientation, which has been shown to be invariant to illumination, to eliminate the effect of illumination change using certain feature extraction methods. (Cao *et al.* 2012, Tzimiropoulos *et al.* 2012). Beyond illumination normalization, the feature extraction methods have better edge preserving ability. However, feature extraction is not very stable to some illuminations, especially in the dark evening. More importantly, it is observed that the previously described methods, which got satisfying results in face recognition, would vanish the skin detail and shape information when they are applied to facial expression recognition.

(3) Some research has been carried out recently on face imaging beyond the visible spectrum in face recognition. The first one is thermal infrared imaging, which reflects heat radiation. The use of thermal infrared imaging is a useful technique for identifying faces under uncontrolled illumination conditions (Yoshitomi *et al.* 1997, Hermosilla *et al.* 2012, He *et al.* 2013, Gade & Moeslund 2014). However, instability due to environmental temperatures and not-real-time due to elapsed time are two main disadvantages of thermal infrared imaging. Alternative of imaging acquisition is NIR imaging, which behaves more like normal VIS imaging. In Li *et al.* (2007), Maeng *et al.* (2011), they showed that an NIR system is more suitable for face recognition than a VIS imaging system , when there are changes in environmental illumination conditions, and especially, in the light angle. They also demonstrated that the influence of environmental lighting to the face was reduced considerably due to the present NIR imaging system. More specially, different from illumination normalization and feature extraction, using an NIR imaging system does not remove the skin texture and wrinkles, which are very important to facial expression recognition. Therefore, an NIR imaging brings a new dimension to illumination invariant face representation, specially for facial expression recognition.

Presence of occlusion

Partial occlusion sometimes occurs when people are wearing masks. More specifically, the occlusion can be regarded as the noise to facial expression recognition. Therefore, the ability to handle an occluded face is important for achieving robust recognition. Recently, the effect of occlusion on facial expression recognition has received much attention from the research community. Existing facial expression recognition studies have previously attempted to solve the occlusion through two ways: facial representation

Table 9. Representative methods for handling facial occlusion.

Reference	Type	Methods	Dynamic
Bourel *et al.* (2001)	Holistic feature extraction	Non-negative matrix factorization (NMF)	No
Buciu *et al.* (2005)	Holistic feature extraction	Gabor feature and classifier fusion	No
Kotsia & Pitas (2007)	Holistic feature extraction	Discriminant NMF	No
Mercier *et al.* (2007)	Occlusion detection	Model construction and residual image statistics	No
Towner & Slater (2007)	Geometric reconstruction	PCA-based methods to reconstruct the positions of occluded part	No
Hammal *et al.* (2009)	Face simulation	Facial point deformations and modified transferable belief model	No
Mao *et al.* (2009)	Model construction	Robust principal component analysis	No
Fanelli *et al.* (2010)	Feature extraction and classification	Log-Gabor response feature, dense optical flow, random forests and Hough voting	Yes
Cotter (2011)	Occlusion detection	Sparse representation classifier and a weight voting	No
Jiang & Jia (2011)	Holistic feature extraction	Eigen-faces and Fisher-faces	No
Zhang *et al.* (2011)	Holistic feature extraction	Gabor feature and template matching	No
Zhi *et al.* (2011)	Holistic feature extraction	Graph-preserving sparse NMF	No
Zhang *et al.* (2012)	Occlusion detection	Sparse representation and feature extraction	No

and occlusion detection.

(1) Facial representation methods have attempted to remove the impact of occlusion by recovering missing geometric or texture features based on the configuration and visual properties of the face. The facial representation methods for partial occlusion are broadly categorized into holistic and local feature sets.

For the holistic feature methods, one deals with the whole face by using Gabor or Eigen-faces. For example, the method in Buciu *et al.* (2005) presented the usage of Gabor feature of facial image and classifier fusion to solve the occlusion in static images. Zhang *et al.* (2011) further used the Gabor features to build the template, and then applied the template matching to recognize the facial expressions in partial occlusion. Other method in Towner & Slater (2007) utilized three PCA-based approaches to reconstruct the positions of missing points at the top and bottom of the face. Their results showed that occlusion of the top of the face can be reconstructed with little loss, while occlusion of the bottom is reconstructed less accurately but still gives comparable expression recognition accuracy. An analysis of Eigen-faces and Fisher-faces with near neighbor method and SVM was proposed by Jiang & Jia (2011) to be robust to facial occlusions. Some works have attempted to recognize facial expressions by using facial points. Hammal *et al.* (2009) recognized facial expression from partially occluded images based on facial point deformations and a modified transferable belief model. The different types of occlusion were simulated by adding bubble masks into the face and handled by the transferable belief model, which can integrate information from different local facial regions, and deal with uncertain and imprecise data.

By the nature of face structures, facial models using local facial information extraction have an advantage over the holistic feature methods (Bourel *et al.* 2001). The non-negative matrix factorization (NMF) is a classical local method to extract the features from partial occluded images. The recent study in Kotsia & Pitas (2007) proposed discriminant NMF for recognizing facial expressions under partial occlusion. Moreover, Zhi *et al.* (2011) attempted a graph-preserving sparse NMF to resolve the problem of facial occlusion. They transformed the high-dimensional facial images into a locality-preserving subspace.

(2) Besides facial representation, an alternative method, named as occlusion detection, has been proposed to automatically detect facial occlusion (Mercier *et al.* 2007, Mao *et al.* 2009, Xia *et al.* 2012). It recognizes the facial expression in occlusion through three steps: firstly building a statistical model from un-occluded faces, and then detecting and then removing the face region in occlusion, finally processing the un-occluded face regions. In Mercier *et al.* (2007), it was proposed to construct the model of un-occluded face by a fitting algorithm, and then detect occlusion by means of residual image statistics. In Mao *et al.* (2009), robust principal component analysis and saliency occlusion were used to detect facial occlusions. Alternatively, sparse representation is widely utilized in face recognition for robustness to face occlusion. A new algorithm for

Table 10. Representative methods for multi-view facial expression recognition on 2D images.

Reference	Methods
Hu *et al.* (2008b,a)	Two view-dependent frameworks based on pose estimation and facial expression recognition
Tang *et al.* (2010)	Dense scale-invariant feature transform (SIFT) feature vectors and ergodic Hidden Markov models
Zheng *et al.* (2009)	SIFT features on 83 facial landmarks
Zheng *et al.* (2010)	Regional covariance matrix representation and Bayes discriminant analysis via Gaussian mixture model
Rudovic *et al.* (2013)	Regression model to map non-frontal facial points into frontal one

facial expression recognition under occlusion was proposed by Cotter (2010b,a, 2011) to apply sparse representation for classifying the facial expression in each region and use a weight voting to combine the score of each class. Additionally, Zhang *et al.* (2012) applied sparse representation based on various features to recognize facial expressions in randomly pixel corruption and the random block occlusion.

In recent years, facial expression recognition based on the analysis of dynamic image sequences has become an active research topic. It conveys that dynamic features from the image sequences can provide much more information than the static images (Zhang & Ji 2005). However, only a few works are investigating dynamic facial expression recognition under occlusion by now (Bourel *et al.* 2000, Zhang & Ji 2005, Fanelli *et al.* 2010). In Bourel *et al.* (2000), they used six local facial regions and a rank-weighted k-nearest-neighbor classifier against partial occlusion in videos. Zhang & Ji (2005) inferred facial expressions in occluded frames via temporal reasoning. In Fanelli *et al.* (2010), a Hough forest-based method was proposed for facial expression sequences. These works mean that even though the facial sequence is partially occluded, the un-occluded regions in the video can still provide the motion information.

Pose change

The methods on facial expression recognition in arbitrary view can be divided into two groups according to the utilization of class labels: (1) one considers facial expressions separated from view labels (Hu *et al.* 2008b,a), and (2) the other couples facial expression labels with view information (Zheng *et al.* 2009, 2010).

The first group can be also called as the view-dependent approach (Moore & Bowden 2011, Rudovic *et al.* 2010, 2013). The essence of the view-dependent approach is an algorithm developed based on pose estimation and emotion classification. Hu *et al.* (2008b,a) established the general framework for multi-view emotion recognition. Deriving from this framework, certain research work in Moore & Bowden (2011) investigated the performance and effectiveness of the various feature descriptor and different dimension reduction schemes for estimating the emotion truth label. Additionally, Rudovic *et al.* (2013) applied the regression-based scheme to map the facial geometric features in the profile view to the optimal ones in the frontal view. Even though these works based on the view-dependent framework achieved promising results on BU-3DFE and Multi-PIE databases, they still have to build the subspace or regression model of each view. High accurate pose estimation should be required in, *e.g.*, Rudovic *et al.* (2013). However, in practice, this is inevitable to exactly estimate the head pose view of facial images.

Instead of advanced pose estimation, the other group considers view labels as complementary information to emotion class labels (Zheng *et al.* 2009, 2010). In other words, the view labels may be embedded in the probability theoretical framework. Zheng *et al.* (2009) pointed out that facial images from the same view and facial expression label can be considered as an independent subclass. The advantage is that the view labels are vanished in the emotion label. But they used a complicated feature with large numerous dimension of 83×128 from landmarks of 3D face model for representing facial images. This may severely limit its practical application due to a 3D face model not available. Therefore, Zheng *et al.* (2010) further presented to exploit the regional covariance matrix representation for appearance features. In addition, the Bayes discriminant analysis via Gaussian mixture model was proposed to reduce the dimensionality of feature vector while preserving the most discriminative information.

Different from the previously described categorization, the method of multi-view facial expression recognition can be also categorized into two classes according to image source: (1) regression model or classification model on 2D image (Zheng *et al.* 2010, Rudovic *et al.* 2013) and (2) 3D facial expression recognition (Tang & Huang 2008, Sun & Yin 2009, Sandbach *et al.* 2012).

It is known that the above mentioned works are mostly derived from 2D facial images, although Zheng *et al.* (2009) used landmarks from 3D face model. In recent years, 3D facial expression recognition has become an active research (Tang & Huang 2008, Sandbach *et al.* 2012). It is noticed that 3D facial expression techniques are mostly depending on the 3D image acquisition and tracking. With recent development,

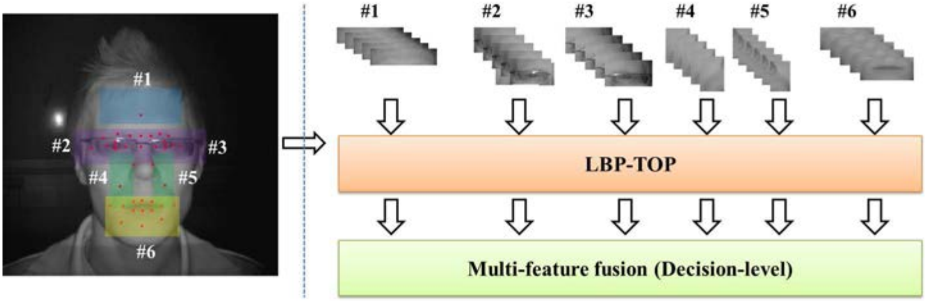


Fig 9. The illustration of component-based method single feature descriptor.

a 3D scanner becomes cheaper and affordable, which may be soon integrated in consumer devices. However, so far, 2D facial images are the most convenient ones to be acquired for a real-time facial expression recognition system, especially in outdoor environment. Recently, the combination of 2D and 3D facial expression recognition has been extensively discussed (Tsalakanidou & Malassiotis 2010).

3.3 Component-based feature descriptor

In recent years, the works of Heisele & Koshizen (2004), Ivanov *et al.* (2004), Li *et al.* (2009) cropped the facial image into some sub-regions or components, and then extracted the appearance features from those components. However, the features in the earlier component-based methods were only extracted from static images, even though Zhang & Ji (2005), Sun & Yin (2009) used HMMs or DBNs to integrate the static information with time development.

Researches (Heisele & Koshizen 2004, Ivanov *et al.* 2004) have also shown that the component-based approach is robust in some cases against pose motion or partial occlusion. In Section 2.6, though the sparse feature descriptor can perform well, each region is not always accurately located. Considering the advantages of the component-based approach, the merit of Section 2.6 can be further extended to solve the limitation of the component-based approach. Therefore, we propose two variants of component-based methods: (1) component-based single feature descriptor (CSFD) in Section 3.3.1, and (2) component-based multiple feature descriptor (CMFD) in Section 3.3.2. More details of CSFD and CMFD can be referred to Papers IV and V, respectively. More importantly, considering the unequal important role of each facial component, the weight learning procedure (WL) is further presented in Section 3.3.3. More details of WL can be referred

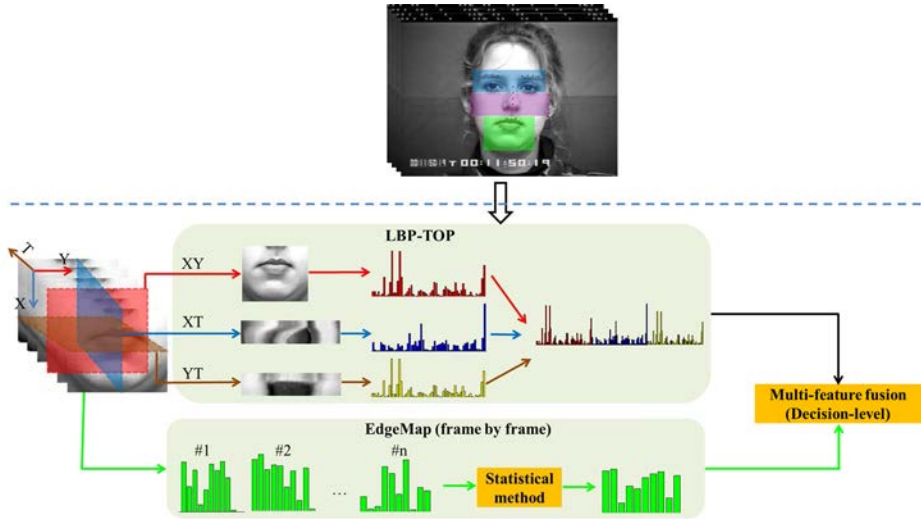


Fig 10. The illustration of using LBP and EdgeMap to construct component-based multiple feature descriptor, where the number is the index of frame (VI, published by permission of Elsevier).

to Paper V.

3.3.1 Component-based single feature descriptor

CSFD is presented based on six facial components and LBP-TOP. The procedure is shown in Figure 9. Firstly,considering time-consuming of manual labeling, the model in Section 2.6 is used to detect the facial landmarks of the first frame. According to 38 facial landmarks, six facial components, *i.e.*, forehead, two eyes, two cheeks and mouth, can be located. Secondly, the LBP-TOP feature descriptor is used to extract the spatiotemporal features of each component, where they are denoted as $\{\Omega_m\}$.

In general, the features from all face components are simply concatenated into a single feature. However, it may cause the curse of dimensionality to the classification. Instead, the decision-level fusion can allow each feature to play its role in facial expression and reduce the mutual effect of each region. The principle of the decision-level fusion has been previously described in Section 2.6.3. Considering the role of each face component, we rewrite the mathematical formulation of this fusion as follows:

Given all features vectors $\{\Omega_m\}_{m=1}^{\overline{M}}$, it can assign $\vec{x}_{test} \to \hat{c}$ if

$$\begin{aligned} & Pr(c = \hat{c}|\vec{x}_{test}, \Omega_1, \ldots, \Omega_{\overline{M}}) \\ = & \max_{c \in \{1,\ldots,C\}} \otimes \{Pr(c|\Omega_1)\beta_{c,1}, \ldots, P(c|\Omega_{\overline{M}})\beta_{c,\overline{M}}\}, \end{aligned} \tag{19}$$

where $\vec{x}_{test}$ and $\hat{c}$ represent the testing sample and its class, respectively, $\otimes$ is the decision-level rule, $\beta_{c,m}$ represents the weights of the mth feature vector to the cth expression. By default, the weights for all features vector are equal. In the solution, $P(c|\Omega_m)$ can be obtained by binary SVM and soft-max function.

3.3.2 Component-based multiple feature descriptor

CMFD aims to extent a multiple feature descriptor into the component-based approach. The method is shown in Figure 10. It contains the appearance and shape representations for each facial component. We firstly consider three components, eyes (including eyebrows), nose, and mouth derived from facial images according to facial configuration and facial landmarks. For each component, the appearance and shape features are extracted.

(1) For appearance representation, the LBP operator is implemented, since it is a monotonic gray-scale invariant texture primitive static and it has also been widely applied in various fields such as texture classification (Ojala *et al.* 2002). However, the appearance representations are not sufficient to give good features to classification. Here, we borrow the idea of LBP-TOP to further describe the motion features from the horizontal and vertical planes.

(2) For shape representation, Edge map (EdgeMap) (Gizatdinova & Surakka 2006), which describe structural features, is extended to describe the shape representation for three components in the image sequence. In the implementation, we directly extend EdgeMap into the temporal domain, where the mean histogram of all frames in one video sequence are statistically calculated for representing the shape variations along the temporal domain.

It is experimentally observed that features of each facial component cannot achieve promising results if feature descriptors were straightly used. Alternatively, for each component, it is divided into $m \times n$ grids. The feature descriptor, *e.g.*, LBP, is then applied in each grid. Finally, each component is formulated by one feature which is obtained by concatenating the features of grids. These component-based features are finally combined by using the decision-level method, as described in Equation (19).

3.3.3 Weight learning

It is observed that $\beta_{c,m}$ is an important element for Equation (19). In order to get optimal weights for different facial components, weight learning is proposed by using multiple kernel learning (MKL). Given multiple feature vectors $\{\Omega_m\}_{m=1}^{\overline{M}}$, each of them has N samples $\{\vec{x}_{m,i}\}_{i=1}^{N}$, and the corresponding class label of $\vec{x}_{m,i}$ is c_i, where $c_i \in \{+1,-1\}$, one can calculate multiple basis kernels for each feature vector. Hence, the kernel of the MKL is computed as a convex combination of the basis kernels

$$k_{i,j} = \sum_{m=1}^{\overline{M}} \beta_m k(\vec{x}_{m,i}, \vec{x}_{m,j}), \tag{20}$$

and restricted by $\beta_m \geqslant 0, \sum_{m=1}^{\overline{M}} \beta_m = 1$, where β_m is the weight of the mth feature vector, $k(\vec{x}_{m,i}, \vec{x}_{m,j})$ is the kernel of $\vec{x}_{m,i}$ and $\vec{x}_{m,j}$. The Equation (20) is to show the purpose of the MKL that combines the multiple feature vectors into a single feature vector by assigning different weights. It thus is defined based on all components. Here, we use a linear kernel function to compute the basis kernel. And then we normalize all kernel matrices of feature sets to unit trace. To make sure each kernel matrix is positive-definite, we have added the absolute of the smallest eigenvalue of the kernel matrix to the diagonal of this kernel matrix, if this smallest eigenvalue is negative. This formulation enables the kernel combination weights to be learnt within the SVM framework. For kernel algorithms, the solution of the learning problem of each component is of the form $f_m(\vec{x}_{m,test}) = \sum_i \alpha_i K_m(\vec{x}_{m,test}, \vec{x}_{m,i}) + b$, where α_i and b are some coefficients to be learned from samples, $\vec{x}_{m,test}$ is the mth feature vector of $\vec{x}_{test}$.

The classifier model parameter and the weights are given by the optimization problem of the MKL based on SVM as

$$\min \sum_m \frac{1}{\beta_m} ||f_m||^2_{\mathscr{H}_m} + \hat{C} \sum_i \xi_i, \tag{21}$$

and it is restricted by (1) $y_i \sum_m f_m(x_{m,i}) + y_i b \geq 1 - \xi_i$, (2) $\xi_i \geq 0$, (3) $\beta_m \geq 0$, and (4) $\sum_m \beta_m = 1$, where ξ_i is the slack afforded to each sample, $\hat{C}$ is the regularization parameter, and f_m belongs to an reproducing kernel Hilbert space $\mathscr{H}_m$ associated with a kernel function. In our implementation, the stop criterion of the optimization to Equation (21) is based on the variation of coefficient β_m between two consecutive steps.

3.4 Discriminative component-based spatiotemporal features with NIR imaging for illumination variations

Images via NIR imaging are subject to a monotonic transform, and LBP-TOP features are robust to monotonic gray-scale changes. A facial expression recognition system invariant to illumination changes can be generated, when an NIR imaging system and LBP-TOP features are combined together. Although LBP-TOP and NIR images can outperform some illumination normalization methods, there is still space to further improve them. The framework is presented in Paper IV, deriving it from some works presented in Paper III.

3.4.1 NIR imaging

The NIR imaging system used in Paper IV consists of an NIR camera, a color camera, a camera box and 18 NIR light-emitting diodes mounted on the camera box. The NIR imaging system was used to collect a new facial expression database for both NIR and VIS images on dark, weak and normal illuminations. For a brief description, NIR and VIS images on dark, weak and normal illuminations are abbreviated as NIR_D, NIR_W, NIR_N, VIS_D, VIS_W and VIS_N, respectively.

In the NIR imaging system, two methods are used to control the light direction: (1) active lights are mounted on the camera to provide frontal lighting and (2) environmental lighting is minimized. NIR LEDs are used as active lights. A reasonable wavelength for active lights is 850 nm, which is in the NIR spectrum (780-1,100 nm).

3.4.2 Discriminative component-based features

Based on images obtained from the NIR imaging system, the CSFD, presented in Section 3.3.1, is used to extract features of facial expressions. In the implementation of ASM (Milborrow & Nicolls 2008), it is found that it mostly failed to detect facial points from NIR images when using an ASM model trained from VIS images, since the model based on VIS images cannot be adapted to NIR spectrum. Instead, some NIR images are chosen for training a new ASM model. In the experiments, it is found that the ASM model based on NIR images can work promisingly for NIR and VIS images in all illuminations. The detection rates were: for NIR images, 98.93% in normal and weak illumination, 98.37% in dark illumination; for VIS images, 97.07% in normal, 98.95% in weak, 95.79% in dark illumination. Some misalignment occurred in a few images, but this error was acceptable in our experiments, and we did not make any further manual processing to remove it.

The size of each component is so large that more than one block is needed to describe its local spatiotemporal information. However, the areas of facial components are different. For example, both cheek areas are much smaller than forehead and mouth areas. This means that using the same number of blocks for all components is not reasonable. Thus, a different number of blocks was used for different components in our experiments. Some tiny areas in one component usually contain more discriminative information than others. It is not necessary to use all the information available in the image, but only the most important areas in terms of distinguishing between subjects or events. The method in Section 2.6.2, also presented in Paper III, is implemented for selecting the most discriminative features for each component. Therefore, the CSFD has discriminative ability, where it is named as DisCSFD.

3.4.3 Experiments

For evaluating the proposed method, we apply 10-fold cross-validation test scheme on the Oulu-CASIA NIR&VIS facial expression database, which is described in Paper IV. Each facial component has discriminative information for classification. It is found that the areas of different facial components are not equal. Since using the same block size cannot preserve the optimal capability of each component, it is necessary to select the optimal grid-size for each component. In the experiments, it is observed that the highest performance is achieved with 5×4 blocks for the right eye, 7×7 for the left eye, 5×4 for the right cheek, 4×3 for the left cheek, 8×7 for the mouth and 5×5 blocks for the forehead, respectively. For feature selection, we implement these parameters 45, 30, 30, 60, 90, 30 for right furrow, left furrow, right eye, left eye, mouth and forehead, respectively.

In the experiments, we evaluate the performance of the system consisting of NIR imaging system and CSFD, DisCSFD and their weighted version. The ‘product’ rule is employed in Equation 19. Table 11 gives the accuracies of different expressions and average performance of the proposed method in VIS and NIR imaging system. We also implement the LBP-TOP (Zhao & Pietikäinen 2007) into VIS and NIR imaging systems. As seen from Table 11, some interesting things can be seen:

(1) It is found that the NIR imaging system provides more robustness for facial

Table 11. Comparative performance (%) of LBP-TOP and the proposed methods on VIS and NIR images, where $*$ represents our proposed methods (IV, published by permission of Elsevier).

Illumination	VIS_N	VIS_W	VIS_D	NIR_N	NIR_W	NIR_D
LBP-TOP	73.54	60.44	56.55	72.09	66.99	69.24
CSFD*	73.06	65.78	54.37	73.79	70.63	69.66
CSFD+WL*	73.54	65.78	53.35	73.54	71.15	69.66
DisCSFD*	75.97	67.96	60.92	75.00	73.79	72.33
DisCSFD+WL*	76.45	67.75	60.92	75.66	73.79	73.74

expression recognition than VIS imaging. It demonstrates that NIR imaging can be more robust to illumination variations, even for poor illumination.

(2) Based on VIS imaging, it is seen that the proposed DisCSFD method performs better than LBP-TOP. More specifically, DisCSFD can still preserve the accuracy of 60.92% under VIS_D, while LBP-TOP only achieves the accuracy of 56.55%.

(3) WL method gives a little improvement for CSFD and DisCSFD at most of illuminations except VIS_W. One main reason is that it may be caused by mis-alignment error.

(4) Comparing NIR imaging and LBP-TOP, it is seen that the proposed system combining DisCSFD and NIR imaging is least influenced by different illuminations. Even in dark illumination, the performance is close to the one in normal illumination. It demonstrates that the proposed system in Paper IV works well in different illuminations.

3.5 Occlusion detection for dynamic occlusions

How to handle facial occlusion is an active research topic for facial expression recognition (Bourel *et al.* 2001, Buciu *et al.* 2005, Kotsia *et al.* 2008a, Jiang & Jia 2011), and has also become a challenging problem for achieving promising results. It is found that CMFD, previously described in Section 3.3.2 (Paper V), may be influenced by facial occlusion. More specifically, the weight learning in Section 3.3.3 may assign the false weights for facial occlusion. It is necessary to marginally filter the noise caused by occlusion. In this section, we present an occlusion detection (OD) system for handling facial occlusion,which is presented in Paper VI. It compresses the influence of noise to weight learning. The completed system (CMFD+OD+WL) which can be used in un-occluded and occluded video sequences is shown in Figure 11. More details can be

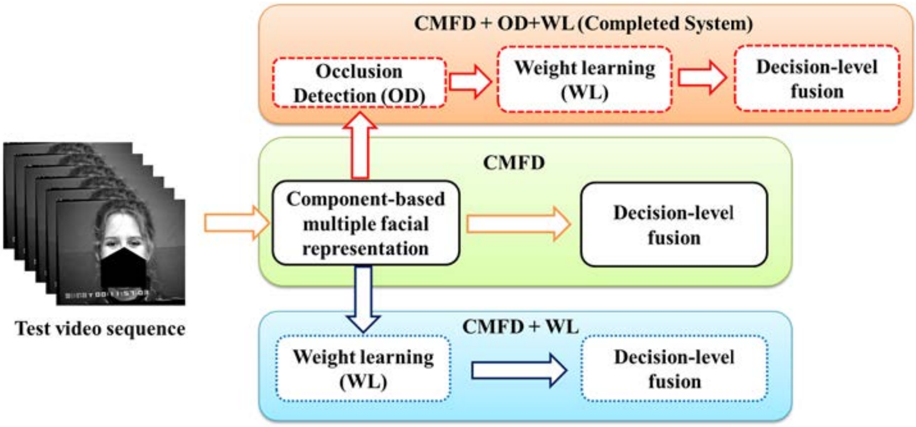


Fig 11. The flowchart of the completed system for handling occlusion in video sequences.

referred to Paper VI.

3.5.1 Construction of occlusion detection

The aim of occlusion detection is to determine whether the region is occluded or not. We exploit sparse representation to develop a real-time system to detect the occlusion for facial expressions. Firstly, we introduce the theorem of sparse representation. Given training samples of the cth object class, each of the training samples is denoted as $\vec{x}_c^j \in \Re^d$, where d is the dimensionality, j represents the index of a training sample of the ith object class. Let us define $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_C]$, where $\mathbf{X}_c$ is the set of samples of the cth object class and C represents the number of the object class. Given any test sample $\vec{x}_{test} \in \Re^d$ from the same class, it can be linearly spanned by the training samples of the ith object class as follows:

$$\vec{x}_{test} = \omega_c^1 \vec{x}_c^1 + \omega_c^2 \vec{x}_c^2 + \cdots + \omega_c^{N_c} \vec{x}_c^{N_c} = \mathbf{X}_c \vec{\omega}_c, \tag{22}$$

where ω_c^j is a reconstruction coefficient of the jth training sample of the cth object class, $\vec{\omega}_c = [\omega_c^1, \ldots, \omega_c^{N_c}]$ and N_c is the number of training samples of cth object class.

The sparse representation is implemented to reconstruct a facial sequence. It is achieved by optimizing the following l_1-optimization that consists of the reconstruction error and the sparsity of the representation:

$$\hat{\omega} = \arg\min \|\vec{\omega}\|_1, \text{subject to } \|\mathbf{X}\vec{\omega} - \vec{x}_{test}\|_2^2 \leq \varepsilon, \tag{23}$$

where ε is the threshold, and vector $\vec{\omega} = [\omega_1, \ldots, \omega_N]$ depicts the contribution of the training images, $N = \sum_{i=1}^{C} N_i$.

In the implementation, for a given facial video sequence, three facial components are easily obtained. Here, we take an example of analyzing eye component. Considering one component may not be absolutely occluded, the block-partition-based method with sparse representation is employed. In this case, eye component is divided into M blocks. For given $\vec{F}_{m,test}$ of the mth block, where $\vec{F}_{m,test}$ are concatenated by appearance and shape features, the coefficient vectors $[\widetilde{\omega}_1, \ldots, \widetilde{\omega}_C]$ are obtained and ordered by the index of training samples. Using only the coefficients associated with the cth expression class, we approximate $\vec{F}_{m,test}$ as $\vec{F}_{m,c}\widetilde{\omega}_c$. In our paper, the residual between $\vec{F}_{m,test}$ and its approximation in the cth expression class is defined as $r_c(\vec{F}_{m,test}) = \|\vec{F}_{m,test} - \vec{F}_{m,c}\widetilde{\omega}_c\|_2$, and the residuals associated with all expression classes of the test sample are represented by $\{r_c\}_{c=1}^{C}$. Furthermore, the minimum residual r_{min} is chosen, *i.e.*, $\min\{r_c\}_{c=1}^{C}$. For classifying occlusion, the threshold machine is defined as

$$\triangle(\vec{x}_{m,test}) = \begin{cases} 0, r_{min} \geq \gamma \\ 1, r_{min} < \gamma \end{cases}, \tag{24}$$

where γ is the threshold of residual, ‘0’ and ‘1’ represent the occluded and un-occluded status of the mth block, respectively. The procedure can be seen in Algorithm 1.

3.5.2 Experimental analysis

For evaluating the proposed method, we apply the same protocol described in Section 2.5 on CK+ database (Lucey *et al.* 2010) and its simulated database. We compare our method with dynamic EdgeMap (Gizatdinova & Surakka 2006), FSE (*i.e.*, Mean-voting fusion of LBP-TOP and EdgeMap) and (Fanelli *et al.* 2010). The simulated database of facial occlusion is in detail presented in Paper VI. Firstly, we choose the optimal block size for eyes, nose, and mouth. Experimental results in Paper V show that eyes, nose and mouth reach the best recognition when using 9×8, 11×10, and 8×8 grids, respectively. These grid parameters are chosen in the following discussion. More experiments about facial expression recognition based a normal video sequence can be referred to the experiment sections in Papers V and VI.

Algorithm 1: Real-time occlusion detection based on block-based method with sparse representation.

```
input  : Appearance features $\vec{A}$ and shape features $\vec{B}$ from $N$ training samples,
         $\vec{A}_{test}$ and $\vec{B}_{test}$ from testing sample $\vec{x}$, the grid size $M$, threshold of residual
         $\gamma$, $C$ classes
output : $\{\triangle(\vec{x}_{m,test})\}_{m=1}^{M}$
for $m \leftarrow 1$ to $M$ do
    Extract the appearance and shape features $\vec{A}_m$ and $\vec{B}_m$ of the $m$th block;
    $\vec{F}_m = [\vec{A}_m \ \vec{B}_m]$;
    Extract the appearance and shape features $\vec{A}_{m,test}$ and $\vec{B}_{m,test}$ of the $m$th block
    of testing sample;
    $\vec{F}_{m,test} = [\vec{A}_{m,test} \ \vec{B}_{m,test}]$;
    Optimize $\hat{\omega} = \arg\min \|\vec{\omega}\|_1$, subject to $\|\vec{F}_m\vec{\omega} - \vec{F}_{m,test}\|_2^2 \leq \varepsilon$;
    Obtain $\hat{\omega}$;
    for $c \leftarrow 1$ to $C$ do
        Select $\widetilde{\omega}_c$ with the $c$th class from $\hat{\omega}$;
        Calculate $r_c(\vec{F}_{m,test}) = \|\vec{F}_{m,test} - \vec{F}_{m,c}\widetilde{\omega}_c\|_2$;
    Find $r_{min} = \min\{r_c\}_{c=1}^{C}$;
    Obtain $\triangle(\vec{x}_{m,test})$ according to Equation (24);
```

Non-occlusion, eye occlusion, mouth occlusion, lower-face occlusion and random occlusion are considered in the experiments. The results of the proposed method and comparison methods are listed in Table 12, where the 'sum' rule is employed in decision-level fusion of CMFD. Some observations are seen as follows:

(1) CMFD is more robust to the partial occlusion and has better performance than LBP-TOP and dynamic EdgeMap. With the help of OD, CMFD can be further improved by the recognition rate of 6.77%, 7.39%, 32.92%, and 16.47% for eyes, mouth, lower-face occlusion and random occlusion, respectively. It demonstrates that OD has played an important role in reducing the effect of occlusion for CMFD.

(2) As we know, the WL is a critical stage in our framework, because it can learn the optimal weights for face components. But it would be seriously influenced by occlusion cases except the eye occlusion if OD is not used.

(3) It is found that the completed system (CMFD+OD+WL) can achieve promising results on four occlusion cases. More importantly, this system can also work well in

Table 12. The average recognition rate (%) using different approaches under non-occlusion, eye occlusion, mouth occlusion, lower-face occlusion and random occlusion (Area size of occlusion is 50%), where $*$ represents our methods (VI, published by permission of Elsevier).

Case	Non Occlusion	Eye Occlusion	Mouth Occlusion	Lower-face Occlusion	Random Occlusion
LBP-TOP	87.08	63.38	31.38	10.46	31.38
EdgeMap	82.77	60.62	24.31	15.69	33.23
FSE	92	69.23	36.62	10.46	38.46
Fanelli *et al.* (2010)	87.1	-	-	-	37
CMFD*	**89.85**	**78.77**	**66.15**	**34.46**	**60.15**
CMFD+WL*	**93.23**	**92.31**	**63.38**	**31.38**	**61.85**
CMFD+OD*	**89.23**	**85.54**	**73.54**	**67.38**	**76.62**
Completed system*	**92.32**	**93**	**79.08**	**73.54**	**79.69**

non-occlusion case, although it is lower than CMFD+WL. Comparing with Fanelli *et al.* (2010), our system can perform better.

The recognition rates under eyes occlusion using Gabor filters, the DNMF, and shape-based SVMs in Kotsia *et al.* (2008a) is 86.8%, 84.2%, and 82.9%, respectively. And for lower-face occlusion, they are 84.4%, 82.9%, 86.7%. Though the results are not directly comparable due to different experimental setups, processing methods, the number of sequences used, *etc.*, they still give an indication of the discriminative power of each approach. It is found that our proposed system under eye occlusion outperforms their algorithms. Unfortunately, the performance under lower-face occlusion is worse than from these methods. It is 13.16% lower than the best performance in Kotsia *et al.* (2008a).

3.6 Multi-view neighborhood preserving embedding

In practice, nearly frontal-view facial images may not be available. Therefore, it is important to investigate a method to recognize the facial expression at arbitrary views. This section attempts to address two questions: (1) whether there exist discriminative features for facial expressions or not and (2) how can we utilize the correlation between facial expressions and views. According to the study of Moore & Bowden (2011), it is seen that local Gabor binary pattern feature operators (LGBP) (Zhang *et al.* 2005) outperforms other variants of LBP in multi-view facial expression recognition. It motivates us to use this feature as an appearance feature based on still images. And then a multi-view discriminative framework based on neighborhood preserving embedding is presented to address the previously mentioned questions. More details can be referred to Paper VII.

3.6.1 Multi-view discriminative framework

Emotion recognition from arbitrary view can be regarded as one issue how to construct a multi-view model to recognize facial expression. In this section, we first propose discriminative neighborhood preserving embedding (DNPE) by introducing Fisher criterion and an intrinsic graph, and then propose to use the theorem of multi-view model to DNPE.

Given n training images with C classes, they are denoted as $\mathbf{X} = [\vec{x}_1, \ldots, \vec{x}_N] \in \Re^d$. With the class label and Euclidean distance, we can obtain the intrinsic graph $\mathbf{G}^{\mathrm{wi}} = \{\mathbf{X}, \mathbf{W}^{\mathrm{wi}}\}$, which preserves the intrinsic structure of intra-class samples, where $\mathbf{W}^{\mathrm{wi}}$ is the similarity matrix of $\mathbf{G}^{\mathrm{wi}}$. We can also obtain the penalty graph $\mathbf{G}^{\mathrm{bw}} = \{\mathbf{X}, \mathbf{W}^{\mathrm{bw}}\}$ that describes the margin across inter-class boundaries in the same way to build an intrinsic graph, which the penalty similarity matrix $\mathbf{W}^{\mathrm{bw}}$ describes the similarity among $\vec{x}_i$ and inter-class ones $\vec{x}_j$.

(1) It constructs the within-class set Ω_p^{wi} of the sample $\vec{x}_i$, where $i = 1, \ldots, N$. This set contains k^{wi} nearest neighbors $\vec{x}_j$ $(j = 1, \ldots, \mathrm{k}^{\mathrm{wi}})$ of $\vec{x}_i$ with the same label of $\vec{x}_i$. In the intrinsic graph, there exists the reconstruction weight matrix $\mathbf{W}^{\mathrm{wi}}$ for $\mathbf{X}$ that can be obtained by minimizing the following formulation:

$$\varepsilon^{\mathrm{wi}}(\mathbf{X}) = \sum_i \parallel \vec{x}_i - \sum_{\vec{x}_j \in \Omega_i^{\mathrm{wi}}} w_{i,j}^{\mathrm{wi}} \vec{x}_j \parallel^2, \tag{25}$$

with the constraint $\sum_i w_{i,j}^{\mathrm{wi}} = 1$, where $w_{i,j}^{\mathrm{wi}}$ in $\mathbf{W}^{\mathrm{wi}}$ is the weight of the edge from $\vec{x}_i$ to $\vec{x}_j$.

Given the lower dimensional feature space $\mathbf{U} \in \Re^{\hat{\mathrm{d}}}$, where $\hat{\mathrm{d}}$ $(\hat{\mathrm{d}} \ll \mathrm{d})$ is the dimensionality of this space, the sample $\vec{x}_i$ is transformed into this space via $\vec{y}_i = \mathbf{U}'\vec{x}_i$. Therefore, the sample $\vec{y}_i$ can be represented as a linear combination of its neighbors with the corresponding coefficients fw_{ij}^{wi}. The corresponding cost function is defined as follows:

$$\begin{aligned}\varepsilon^{\mathrm{wi}}(\mathbf{Y}) &= \sum_i \| \vec{y}_i - \sum_{\vec{y}_j \in \Omega_i^{\mathrm{wi}}} w_{i,j}^{\mathrm{wi}} \vec{y}_j \|^2 \\ &= \mathbf{U}'\mathbf{S}^{\mathrm{wi}}\mathbf{U}, \end{aligned} \tag{26}$$

where $\mathbf{S}^{\mathrm{wi}} = \mathbf{X}(\mathbf{I} - \mathbf{W}^{\mathrm{wi}})(\mathbf{I} - \mathbf{W}^{\mathrm{wi}})'\mathbf{X}'$ represents the local geometric structure of intra-class samples.

(2) It also builds the between-class set Ω_i^{bw} ('bw' means between-class). It contains k^{bw} samples $\vec{x}_j$ $(j = 1, \ldots, \mathrm{k}^{\mathrm{bw}})$ nearest neighboring to $\vec{x}_i$ that have different class labels to $\vec{x}_i$. The $\mathbf{W}^{\mathrm{bw}}$ in the penalty graph $\mathbf{G}^{\mathrm{bw}}$ that can be obtained by minimizing the following formulation:

$$\varepsilon^{bw}(\mathbf{X}) = \sum_i \| \vec{x}_i - \sum_{\vec{x}_r \in \Omega_i^{\mathrm{bw}}} v_{i,j}^{\mathrm{bw}} \vec{x}_r \|^2, \tag{27}$$

with the constraint $\sum_j v_{i,j}^{\mathrm{bw}} = 1$, where $v_{i,j}^{\mathrm{bw}}$ in $\mathbf{W}^{\mathrm{bw}}$ represents the weight of the edge from $\vec{x}_i$ to $\vec{x}_j$ with different class labels.

In the lower dimensional feature space, one aims to maximize the boundary of samples with different class labels. Therefore, $\mathbf{U}$ makes the following objective function maximizes while $\mathbf{W}^{\mathrm{bw}}$ is fixed:

$$\begin{aligned}\varepsilon^{bw}(\mathbf{Y}) &= \sum_i \| \vec{y}_i - \sum_{\vec{y}_j \in \Omega_i^{\mathrm{bw}}} v_{i,j}^{\mathrm{bw}} \vec{y}_j \|^2 \\ &= \mathbf{U}'\mathbf{S}^{\mathrm{bw}}\mathbf{U}, \end{aligned} \tag{28}$$

where $\mathbf{S}^{\mathrm{bw}} = \mathbf{X}(\mathbf{I} - \mathbf{V}^{\mathrm{bw}})(\mathbf{I} - \mathbf{V}^{\mathrm{bw}})'\mathbf{X}'$ represents the local geometric structure of inter-class samples.

According to maximum margin criterion, the objective function, that maximizes the margin between classes and avoids the small sample size problem, is defined as

$$\mathbf{U} = \arg\max\{\mathbf{U}'\mathbf{S}^{\mathrm{bw}}\mathbf{U} - \eta \mathbf{U}'\mathbf{S}^{\mathrm{wi}}\mathbf{U}\}, \tag{29}$$

where η is the balancing factor which adjusts the second term to ensure a positive objective function. We empirically observe that η has not much impact on the performance. For avoiding the degeneration, Equation (29) is restricted by the constraint $\mathbf{U}'\mathbf{X}\mathbf{X}'\mathbf{U} = 1$. It is known that Equation (29) is a constrained quadratic programming problem.

In practice, there can be an arbitrary view angle for facial expression recognition. Here, we simply suppose that there exist V views for each sample. Given the samples with the ith view ($i \in \{1,\dots,V\}$), they are denoted as $\mathbf{X}_i$. Concerning all views, Equation (29) can be formulated as follows:

$$[\mathbf{U}_1,\dots,\mathbf{U}_V] = \arg\max \sum_{i=1}^{V} \mu_i \mathbf{U}_i'(\mathbf{S}_i^{\mathrm{bw}} - \eta_i \mathbf{S}_i^{\mathrm{wi}})\mathbf{U}_i, \tag{30}$$

with constraints $\mathbf{U}_i'\mathbf{X}_i\mathbf{X}_i'\mathbf{U}_i = 1, \forall i$, where the positive term μ_i is included to bring the balance between multiple objectives. Because of non-linear constraints, it leads to no closed form solution in the current form. Instead, the relaxed version of the problem can be obtained by coupling all constraints as $\sum_i \mathbf{U}_i'\mathbf{X}_i\mathbf{X}_i'\mathbf{U}_i = 1$.

In order to make the samples with the same expression label yet in various views accumulating very close, the new objective function is defined as follows:

$$[\mathbf{U}_1,\dots,\mathbf{U}_V] = \arg\max \sum_{i=1}^{V} \sum_{j=1,j\neq i}^{V} \mathbf{U}_i'\mathbf{M}_i\mathbf{M}_j'\mathbf{U}_j, \tag{31}$$

with the constraint $\sum_{i=1}^{V} \mathbf{U}_i'\mathbf{X}_i\mathbf{X}_i'\mathbf{U}_i = 1$, where $\mathbf{M}_i$ is the class mean matrix of samples on the ith view.

Based on Equations (30) and (31), we can finally obtain the completed formulation as follows:

$$[\mathbf{U}_1,\dots,\mathbf{U}_V] = \arg\max \sum_{i=1}^{V} \{\mu_i \mathbf{U}_i'(\mathbf{S}_i^{\mathrm{bw}} - \eta_i \mathbf{S}_i^{\mathrm{wi}})\mathbf{U}_i + \sum_{j=1,j\neq i}^{V} \alpha_{i,j} \mathbf{U}_i'\mathbf{M}_i\mathbf{M}_j'\mathbf{U}_j\}, \tag{32}$$

with the constraint $\sum_{i=1}^{V} \beta_i \mathbf{U}_i'\mathbf{X}_i\mathbf{X}_i'\mathbf{U}_i = 1$, where μ_i, $\alpha_{i,j}$ and β_i are balancing parameters which adjust the importance of terms in the objective function and constraint item.

It is observed that Equation (32) is a standard generalized eigenvalue problem that can be solved using any Eigen-solver. Through this formula, we can obtain discriminative feature space of facial expression $\mathbf{U}_i$ in each view. Here, we name our method as multi-view discriminative neighbor preserving embedding (MVDNPE).

3.6.2 Implementation

Our aim is to match two face images with the same or different facial expression label in different views. The mean-correlation maximization classifier is designed to classify this sample as follows:

$$\xi(\vec{x}) = \max_c(\operatorname*{mean}_c(\max_i\{\mathrm{corr}(\mathbf{U}_i'\mathbf{X}_{i,c}, \mathbf{U}_i'\vec{x})\}_{i=1}^{V})), \tag{33}$$

Table 13. Comparative performance (%) of the methods on BU-3DFE and Multi-PIE databases, where $*$ represents our methods (VII, published by permission of BMVA).

Method	BU-3DFE	Multi-PIE
PCA (Turk & Pentland 1991)	67.70	69.12
LDA (Belhumeur *et al.* 1997)	65.87	76.09
LPP (He & Niyogi 2003)	69.33	74.57
NPE (He *et al.* 2005)	69.63	74.12
SIFT+Multi-view Bayes (Zheng *et al.* 2009)	68.35	-
BDA/GMM (Zheng *et al.* 2010)	68.28	-
DNPE*	**69.72**	**74.35**
MVDNPE*	**72.47**	**76.83**

where $\mathbf{X}_{i,c}$ represents training samples of the cth facial expression label with the ith view, *corr* represents Pearson's linear correlation coefficient operator, V is the number of views, *mean* and *max* are the mean and maximum value operators, respectively.

3.6.3 Experiments

The presented method is extensively evaluated on the BU-3DFE (Yin *et al.* 2006) and Multi-PIE (Gross *et al.* 2010) databases and is compared with several state-of-the-art facial expression recognition approaches. Experimental results in the original Paper VII demonstrated that this method consistently achieved the highest recognition accuracies among other methods under comparison.

BU-3DFE database

By projecting 3D facial expression models in various directions, we can generate a set of 2D facial images with various facial views. In our experiment, we choose 3D models with the highest level of intensity to generate five yaw views (0°, $+30°$, $+45°$, $+60°$, $+90°$) with six facial expressions (anger, disgust, fear, happiness, sadness, and surprise).

In experiments, we randomly divided 100 subjects into ten groups, each one having ten subjects. In each trial of the experiment, we choose one group as the test set and the other ones as the training set. We conduct ten trials of the experiment in total such that each subject is used as test data once. The experimental parameter setup can be

Table 14. Accuracy (%) of facial expressions at five views on BU-3DFE database (VII, published by permission of BMVA).

View	0	30	45	60	90
Anger	73	72	71	72	67
Disgust	77	72	68	60	60
Fear	65	58	57	50	49
Happiness	91	89	89	86	88
Sadness	69	65	60	65	57
Surprise	89	90	91	90	84
Average	**77.33**	**74.33**	**72.67**	**70.50**	**67.50**

referred to Paper VII. The confusion matrix of this method is computed. It is shown that the proposed method achieves good performance on each facial expression (*i.e.*, 71% for "Anger", 67.40% for "Disgust", 55.80% for "Fear", 88.60% for "Happiness", 63.20% for "Sadness" and 88.80% for "Surprise"). The comparison with two recent methods, SIFT+Multi-view Bayes (Zheng *et al.* 2009) and Bayes discriminant analysis via Gaussian mixture model (BDA/GMM) (Zheng *et al.* 2010) is shown in Table 13. We also give the accuracy of PCA (Turk & Pentland 1991), LDA (Belhumeur *et al.* 1997), LPP (He & Niyogi 2003), NPE (He *et al.* 2005) based on the framework of Hu *et al.* (2008b). The very satisfactory performance achieved among all the methods under comparison reflects the effectiveness and robustness of the proposed method.

Table 14 shows overall recognition rates as well as the recognition rates of facial expressions of the proposed method across various views. The increasing view angles can affect much the performance of facial expression recognition. One reason for that may be that the information that can be obtained is less with the increasing of view angles. While the angle reaches 90°, there is a 9.83% in the performance. The optimal results for average, anger, disgust, fear, happiness, sadness are achieved when face images are in the frontal view. From Table 14, it can be seen that "surprise" expression was recognized more easily at angles of 30°, 45° and 60°. These are most likely because the lip movement provides respective evidence to surprise expression in the non-frontal view. It is surprising to see that the performance of sadness at 60° is better than the one at 45°. This is mostly likely due to that the lip movement has more evidence at 60°. We can also see a similar situation for anger at 45° and 60° because of the eye movement.

Multi-PIE database

In this experiment, we have chosen 100 subjects with five facial expressions (disgust, scream, smile, squint and surprise), normal illumination and thirteen poses, ranging from the left profile (-90°) to the right profile (+90°) at an interval of 15°.

The confusion matrix is also given in Paper VII. It is shown that the proposed method achieves good performance on each facial expression (*i.e.*, 65.77% for "Disgust", 83.54% for "Scream", 78.69% for "Smile", 73.54% for "Squint", and 82.62% for "Surprise"). Therefore, the facial expressions with the recognition performance ranked from the best to the worst are the following: scream, surprise, smile, squint, disgust. The recognition rate for scream is 83.54%, while for disgust it is only 65.77%.

The comparison with PCA, LDA, LPP and NPE is shown in Table 13. It can be seen that our approach on this database outperforms the PCA, LPP and NPE, in which their rates are 7.71%, 2.27% and 2.71% lower, respectively. And our method also performs better than LDA by 0.74%.The very satisfactory performance achieved among all the methods under comparison reflects the effectiveness and robustness of the proposed method.

3.7 Discussion

This chapter presented two variants of the component-based method and three methods to handle and compress the influence of three factors, *i.e.*, illumination variations, face occlusion and face poses.

The methods proposed in Section 3.3 are motivated by the recent component-based method and the previous work in Section 2.6. They aim to develop the new variants of the component-based approach, including component-based single feature descriptor (CSFD) and component-based multiple feature descriptor (CMFD). The CSFD presented in Paper IV contains two stages: the face component localization and facial representation. In the first stage, the ASM is used for detecting facial points on facial images. According to the geometric information, six face components are located. In the second stage, LBP-TOP was utilized to each face component. In contrast, the CMFD presented in Paper V only focuses on three facial components, where LBP-TOP and EdgeMap are used for each facial component. For CSFD and CMFD, they employ decision-level fusion for fusing features of all components. Additionally, they also use the weight learning approach to learn the optimal weights for all components. It is found that CSFD and CMFD only apply the LBP-TOP feature descriptor for describing the feature of facial regions. Other spatiotemporal features are not evaluated for both methods. This is the limitation of CSFD and CMFD. Actually, other spatiotemporal features, *e.g.*, STLMBP presented in Section 2.5 and the temporal extension of CLQP in Section 2.4, can be used as well.

The approach proposed in Section 3.4 (Paper IV) aims to solve the illumination variation problem by using the NIR imaging system and CSFD. The NIR imaging system provides the robustness to monotonic gray-scale changes caused by illumination changes for facial expression recognition. Especially for increasing discriminative power, the feature selection based on AdaBoost was applied to each face component on the CSFD. The advantages of this method include: (1) It combines the advantage of NIR images and component-based LBP-TOP approach, which enhances the ability of robustness to illumination variation; (2) it considers abundant temporal discriminative information, which makes it flexibly adaptive to various illuminations. This method has been evaluated qualitatively and quantitatively, and compared with well-known feature descriptor and illumination normalization in Paper IV.

For the second factor, a solution to the dynamic facial occluded expression recognition problem was presented. Based on CMFD, the occlusion detection was further proposed towards facial occlusion in video sequences. It is flexible to build the adaptive occlusion detection for throwing away the occluded blocks. A completed facial expression system was presented in Paper VI. It was the first complete way of addressing dynamic facial expression recognition under occlusion by combining feature descriptor and occlusion detector. It has been thoroughly evaluated in various conditions of occlusion, and quantitatively compared with the state of the arts. Experimental results show that it consistently achieves good performance under some classical types of occlusions. Though the proposed method obtains promising results, it still hard resolves other challenging type such as the face is totally occluded.

For the last factor, a manifold-learning-based method was used to solve the face pose. This method was described in Section 3.6 (Paper VII). It consists of two stages: multi-view model learning and the classification of emotions. In the first stage, each discriminative subspace was created by using the presented discriminative neighborhood preserving embedding method. Additionally, each subspace further preserved much discriminant information according to intra-class boundary. Furthermore, the correlation of each subspace was explored to obtain a new feature space, which in addition contains information on view. In the recognition stage, the mean-correlation maximization classifier was designed to determine its type of facial expression of various views. This method is the derivation of cross-view face recognition of tackling the face pose in facial expression recognition. Numerous experiments demonstrate that it achieves better performance than the methods under comparison. Unfortunately, Section 3.6 (Paper VII) suffers from the following limitations: (1) the influence of other feature descriptors, *e.g.*, CLQP in Section 2.4, is not evaluated, and (2) it only tests the method on still images.

4 Summary

Facial expression recognition provides a good protocol to evaluate the emotional state of human beings. This thesis first involves spatial and spatiotemporal descriptors for facial expression recognition, and then further discusses potential solutions to the problems caused by uncontrolled environments, including illumination variations, facial occlusion and pose changes.

4.1 Methods and contributions

All methodologies proposed in the thesis inherit the advantages and abilities of LBP and LBP-TOP and subsequently enhance their performance.

In Chapter 2, three new feature extraction descriptors are presented, which are originally described in Paper I-III, respectively.

(1) A more generalized LBP method (Paper I) has focused on the advantage of CLBP and LQP. Firstly, the new descriptor re-visits the decision function of LBP following the principle of CLBP. Additionally, the use of orientation information is exploited for facial expression. The completed information can make the feature set robust. Moreover, it attempts to resolve the problems of LBP restricted into specific patterns. The revised vector quantization based on k-means is utilized to learn the flexible pattern, and it also allows using the deep spatial structures and having efficient computational cost.

(2) A new two layers spatiotemporal descriptor (Paper II) is studied. The multi-layer structure of LBP in still images motivates us to develop a similar framework for describing dynamic facial expressions, since this structure is robust to gray-scale changes. Therefore, a new descriptor exploits a compact representation of features and the principle of LBP-TOP for facial expressions. It aims to have little information loss and be not restricted to one octave. More specifically, the magnitude and orientation information of the monogenic filter are extracted, and then they are encoded into a statistical histogram using the framework of LBP-TOP.

(3) A sparse spatiotemporal feature method combines the geometric information and LBP-TOP. It aims to resolve the problems caused by the facial outliers and unnecessary facial parts. It also attempts to compress the redundant information from the whole face. In the implementation, the feature descriptor is achieved by using ASM and

LBP-TOP, where ASM can focus on regions of interest and LBP-TOP efficiently obtains the appearance and motion features. Furthermore, the discriminative information is obtained by using feature selection. Finally, multi-classifier fusion is further utilized in the output of features for boosting the performance.

The thesis also considers three common environment conditions.

(1) It firstly studies two variants of the component-based approach, including component-based single feature descriptor (CSFD) and component-based multiple feature descriptor (CMFD). In CSFD, six interesting face regions are chosen, and then LBP-TOP is utilized to extract the features. In CMFD, multiple feature vectors are extracted from three facial components. Instead of concentrating features, decision-level fusion and weight learning are developed to combine all features and assign the optimal weights for all features, respectively.

(2) The thesis further employs an NIR imaging system and CSFD for handling with the illumination variation in facial expression recognition. The NIR imaging system is robust to illumination variation. Furthermore, CSFD is shown to outperform LBP-TOP. Additionally, an attempt is made to increase the discriminative ability of CSFD by using a feature selection method. The NIR imaging system can enhance the resistance of CSFD features and disCSFD to the poor illuminations. A combination of CSFD or DisCSFD and NIR imaging system works well in different illuminations.

(3) The thesis investigates the major problem of facial expression, *i.e.*, if the hand or other stuff occludes some face regions, it would influence the system of facial expression recognition. Here, CMFD is presented to show its good performance in normal videos and occluded ones. More importantly, an occlusion detector based on sparse representation is presented to throw the occluded region away. As well, weight learning is used to assign the weights to three facial components. Different from the traditional methods in occlusion of facial expression, the work successfully combines the advantage of feature descriptor and occlusion detection. To our knowledge, the proposed method is the first one to use occlusion detection in video sequences.

(4) The thesis finally discussed the critical issue of facial expression, namely, how to tackle the pose changes. A multi-view discriminative neighbor preserving embedding approach is presented to recognize the facial expression in arbitrary views. It is found that the intra-class intrinsic structure is still ignored by manifold learning. It first exploits the intra-class intrinsic structure and the inter-class penalty graph to strengthen the discriminative power of neighbor preserving embedding. In addition, the maximum margin criterion revisits the objective function. Furthermore, the latest multi-view model

of face recognition and multi-set canonical correlation analysis are exploited to make the correlation of intra-class samples with distinct angles maximization. These schemes lead to lower dimensional features with discriminative capability of facial expressions.

The contributions of this thesis can therefore be briefly summarized as:

(1) Investigation of facial expression representation goes through spatial and temporal analysis. In the progress, three new feature descriptors are proposed to provide robust and stable methods for future facial expression recognition. In addition, they have the efficient computational cost. More specifically, firstly, a completed generalized texture description is presented by encoding the deeper sampling structure and producing a flexible pattern for facial expression recognition. The method compensates the lost dominant property. Secondly, a dynamic texture description is designed based on LBP-TOP for the dynamic facial expression recognition problem. It enhances the robust ability of LBP-TOP to illumination changes. Finally, a combination of geometric features and LBP-TOP produces a new texture descriptor that compresses the redundancy of facial images. These feature descriptors can be viewed as the new variants of LBP.

(2) Comprehensive analysis for three critical conditions, including illumination variation, partial occlusion and pose changes, is studied. Three new ways are presented to resolve these problems. They also make the facial expression recognition system suitable for the real-world conditions. More specifically, an NIR imaging system with a component-based LBP-TOP method provides a new perspective to easily solve the illumination variation. Then, a completed system based on component-based method and occlusion detection provides a new application to handle occlusion. Finally, a multi-view framework avoids pose estimation and makes it reliable to recognize facial expressions at arbitrary views.

4.2 Limitation and future work

In the view of feature extraction in facial expression recognition, all feature descriptors in this thesis are mainly designed based on LBP and LBP-TOP. The proposed three facial representation methods in Section 2 provide new aspects and tools for the problem of facial expression recognition. From a practical viewpoint, they can be applicable to many areas of interest as there is a wide range of topics that involve the particular problem of face analysis, such as micro-expression analysis, pain detection and video-based face recognition, amongst many others.

Despite the promising results, the presented approaches in this thesis are limited to acted facial expressions. In practice, spontaneous and subtle facial expressions can more reveal the real emotional state of human beings. The proposed methods in Section 2 may suffer from the subtle changes and irregular motion variation of facial expression in spontaneous behavior. In many applications of human-computer interaction (HCI), it is important to be able to detect the emotional states of the person in a natural situation. Measuring the intensity of spontaneous facial expressions is, of course, more difficult than measuring acted facial expressions due to the complexity, subtlety and variability of natural expressions. Acted facial expressions may differ in appearance and timing from spontaneously occurring expressions. Hence, there is still room for improvement and extension to spontaneous facial expressions in order to make a dynamic facial descriptor sufficiently generalized, stable, efficient and accurate.

Another important part of this thesis is that we study the effects of illumination changes, partial occlusion and variations for the applications of facial expression recognition. Furthermore, the methods are proposed in Section 3 to resolve the problem caused by these real-world conditions. These approaches exploit the features and machine learning methods in subtle ways. The novel solutions in Section 3 can provide new aspects and tools for resolving the problem caused by real-world conditions for other applications, *e.g.*, face recognition and gender classification.

Though the proposed approaches seem practical and robust to these effects, most of the experiments are based on artificial databases. As far as we know, the light, facial occlusion and view changes are still difficult problems, not only for facial expression recognition, but also for face recognition and micro-expression analysis. Also, how to automatically recognize facial expression when three cases occur at the same time remains a challenging issue. Up to now, there are no complete databases, including all three conditions of illumination, occlusion and pose, especially in videos. Thus, it would be interesting to study whether a combination of the proposed methods would be working at more difficult environments.

In addition, it is found that the features used in Section 3 are limited into LBP-TOP. In fact, other spatiotemporal features presented in Section 2 can be developed instead of LBP-TOP. They would obtain more promising results. On the other hand, the proposed approach in Section 3.6, only considers the view changes in still images. It would also suggest a possible solution for the dynamic facial expression recognition with arbitrary views through a multi-view model.

References

Adini Y, Moses Y & Ullman S (1997) Face recognition: The problem of compensating for changes in illumination direction. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7): 721–732.

Ahonen T, Hadid A & Pietikäinen M (2006) Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12): 2037–2041.

Aleksic S & Katsaggelos K (2006) Automatic facial expression recognition using facial animation parameters and multi-stream HMMS. IEEE Transactions on Information Forensics and Security 1(1): 3–11.

Almaev T & Valstar M (2013) Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction, 356–361.

Anvar S, Yau W & Teoh E (2013) Multiview face detection and registration requiring minimal manual intervention. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(10): 2484–2497.

Bayramoglu N, Zhao G & Pietikäinen M (2013) CS-3DLBP and geometry based person independent 3d facial action unit detection. Proc. International Conference on Biometrics, 1–6.

Belhumeur P, Hespanha J & Kriegman D (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7): 711–720.

Bourel F, Chibelushi C & Low A (2000) Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. Proc. International Conference on Automatic Face and Gesture Recognition, 106–111.

Bourel F, Chibelushi C & Low A (2001) Recognition of facial expressions in the presence of occlusion. Proc. British Machine Vision Conference, 213–222.

Brahnam S, Nanni L & Sexton R (2007) Introduction to neonatal facial pain detection using common and advanced face classification techniques. Proc. Advanced Computational Intelligence Paradigms in Healthcare-1, 225–253.

Buciu I, Kotropoulos C & Pitas I (2003) ICA and Gabor representation for facial expression recognition. Proc. International Conference on Image Processing, 855–858.

Buciu I, Kotsia I & Pitas I (2005) Facial expression analysis under partial occlusion. Proc. International Conference on Acoustics, Speech, and Signal Processing, 453–456.

Cao X, Shen W, Yu L, Wang Y, Yang J & Zhang Z (2012) Illumination invariant extraction for face recognition using neighboring wavelet coefficients. Pattern Recognition 45(4): 1299–1305.

Chan C, Goswami B, Kittler J & Christmas W (2012) Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication. IEEE Transactions on Information Forensics and Security 7(2): 602–612.

Chang K, Chen C & Hung Y (2013) Intensity rank estimation of facial expressions based on a single image. Proc. International Conference on System, Man and Cybernetics, 3157–3162.

Chen S, Tian Y, Liu Q & Metaxas D (2013) Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. Image and Vision Computing 31(2): 175–185.

Chen X, Flynn P & Bowyer K (2005) IR and visible light face recognition. Computer Vision and Image Understanding 99: 332–358.

Cotter S (2010a) Sparse representation for accurate classification of corrupted and occluded facial expressions. Proc. International Conference on Acoustics Speech and Signal Processing, 838–814.

Cotter S (2010b) Weighted voting of sparse representation classifiers for facial expression recognition. Proc. European Signal Processing Conference, 1164–1168.

Cotter S (2011) Recognition of occluded facial expressions using a fusion of localized sparse representation classifier. Proc. Digital Signal Processing Workshop and Signal Processing Education Workshop, 437–442.

Darwin C (1872) The Expression of the Emotions in Man and Animals. London: John Murray, anniversary edition.

Ekman P (1993) Facial expression and emotion. American Psychologist 48(4): 384–392.

Ekman P & Davidson R (1994) The nature of emotion: fundamental questions. Oxford University Press.

Ekman P & Friesen W (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press.

Fanelli G, Yao A, Noel P, Gall J & Gool L (2010) Hough forest-based facial expression recognition from video sequences. Proc. International Workshop on Sign, Gesture and Activity, 195–206.

Fasel B & Luettin J (2003) Automatic facial expression analysis: a survey. Pattern Recognition 33(1): 259–275.

Feng X, Cui J, Pietikäinen M & Hadid A (2005a) Real time facial expression recognition using local binary patterns and linear programming. Proc. Mexican International Conference on Artificial Intelligence, 328–336.

Feng X, Lai Y, Mao X, Peng J, Jiang X & Hadid A (2013) Extracting local binary patterns from image key points: Application to automatic facial expression recognition. Proc. Scandinavian Conference on Image Analysis, 339–348.

Feng X, Pietikäinen M & Hadid A (2005b) Facial expression recognition with local binary patterns and linear programming. Pattern Recognition and Image Analysis 15(2): 546–548.

Field D (1987) Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A 4(12): 2379–2394.

Franc V & Hlavác V (2001) Multi-class support vector machines. Proc. International Conference on Pattern Recognition, 236–239.

Freund Y & Schapire R (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1): 119–139.

Gade R & Moeslund T (2014) Thermal cameras and applications: a survey. Machine Vision and Applications 25(1): 245–262.

Gajsek R, Struc V & Mihelic F (2010) Multi-modal emotion recognition using canonical correlations and acoustic features. Proc. International Conference on Pattern Recognition, 4133–4136.

Gizatdinova Y & Surakka V (2006) Feature-based detection of facial landmarks from neutral and expressive facial images. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(1): 135–139.

Gross R, Matthews I, Cohn J, Kanade T & Baker S (2010) Multi-PIE. Image and Vision Computing 28(5): 807–813.

Guo Y, Zhao G & Pietikäinen M (2012) Discriminative features for texture description. Pattern

Recognition 45(10): 3834–3843.

Guo Z, Zhang L & Zhang D (2010) A completed modeling of local binary pattern operator for texture classification. IEEE Transactions on Image Processing 19(6): 1657–1663.

Hammal Z, Arguin M & Gosselin F (2009) Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions. Journal of Vision 9(2): 1–19.

He S, Wang S, Lan W, Fu H & Ji Q (2013) Facial expression recognition using deep Boltzmann machine from thermal infrared images. Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction, 239–244.

He X, Cai D, Yan S & Zhang H (2005) Neighborhood preserving embedding. Proc. International Conference on Computer Vision, 1208–1213.

He X & Niyogi P (2003) Locality preserving projections. Proc. Advances in Neural Information Processing System, 153–160.

Heisele B & Koshizen B (2004) Components for face recognition. Proc. International Conference on Automatic Face and Gesture Recognition, 153–158.

Hermosilla G, del Solar JR, Verschae R & Correa M (2012) A comparative study of thermal face recognition methods in unconstrained environments. Pattern Recognition 45(7): 2445–2459.

Hu H (2011) Multiscale illumination normalization for face recognition using dual-tree complex wavelet transform in logarithm domain. Computer Vision and Image Understanding 115(10): 1384–1394.

Hu Y, Zeng Z, Yin L, Wei X, Tu J & Huang T (2008a) Multi-view facial expression recognition. Proc. International Conference on Automatic Face and Gesture Recognition, 1–6.

Hu Y, Zeng Z, Yin L, Wei X, Tu J & Huang T (2008b) A study of non-frontal-view facial expression recognition. Proc. International Conference on Pattern Recognition, 1–4.

Huang D, Shan C, Ardabilian M, Wang Y & Chen L (2011) Local binary patterns and its applications to facial image analysis: A survey. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews 41(6): 765–781.

Hussain S, Napoleon T & Jurie F (2012) Face recognition using local quantized patterns. Proc. British Machine Vision Conference, 1–11.

Hussain S & Triggs B (2012) Visual recognition using local quantized patterns. Proc. European Conference on Computer Vision, 716–729.

Ioannou S, Gallese V & Merla A (2014) Thermal infrared imaging in psychophysiology: Potentialities and limits. Psychophysiology 51(10): 951–963.

Ivanov Y, Heisele B & Serre T (2004) Using component features for face recognition. Proc. International Conference on Automatic Face and Gesture Recognition, 421–426.

Jabid T, Kabir M & Chae O (2010) Facial expression recognition using local directional pattern. Proc. International Conference on Image Processing, 1605–1608.

Jain S, Hu C & Aggarwal J (2011) Facial expression recognition with temporal modeling of shapes. Proc. International Conference on Computer Vision, 1642–1649.

Jiang B & Jia K (2011) Research of robust facial expression recognition under facial occlusion condition. Proc. International Conference on Active Media Technology, 92–100.

Jiang B, Valstar M, Martinez B & Pantic M (2014) A dynamic appearance descriptor approach to facial actions temporal modeling. IEEE Transactions on Cybernetics 44(2): 161–174.

Jiang B, Valstar M & Pantic M (2011) Action unit detection using sparse appearance descriptors in space-time video volumes. Proc. International Conference on Automatic Face and Gesture Recognition, 1–11.

Jun B, Kim T & Kim D (2011) A compact local binary pattern using maximization of mutual information for face analysis. Pattern Recognition 44(3): 532–543.

Kaltwang S, Rudovic O & Pantic M (2012) Continuous pain intensity estimation from facial expressions. Proc. International Symposium on Visual Computing, 368–377.

Kanade T, Cohn J & Tian Y (2000) Comprehensive database for facial expression analysis. Proc. International Conference onAutomatic Face and Gesture Recognition, 484–490.

Kanaujia A & Metaxas D (2006) Recognizing facial expressions by tracking feature shapes. Proc. International Conference on Pattern Recognition, 33–38.

Kellokumpu V, Zhao G & Pietikäinen M (2011) Recognition of human actions using texture descriptors. Machine Vision and Applications 22(5): 767–780.

Kong S, Heo J, Abidi B, Paik J & Abidi M (2005) Recent advances in visual and infrared face recognition - a review. Computer Vision and Image Understanding 97(1): 103–135.

Kotsia I, Buciu I & Pitas I (2008a) An analysis of facial expression recognition under partial facial image occlusion. Image and Vision Computing 26(7): 1052–1067.

Kotsia I & Pitas I (2007) Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Transactions on Image Processing 16(1): 172–187.

Kotsia I, Zafeiriou S & Pitas I (2008b) Texture and shape information fusion for facial expression and facial action unit recognition. Pattern Recognition 41(3): 833–851.

Krumhuber E, Kappas A & Manstead A (2013) Effects of dynamic aspects of facial expressions: a review. Emotion Review 5(1): 41–46.

Lei Z, Liao S, He R, Pietikäinen M & Li S (2008) Gabor volume based local binary pattern for face representation and recognition. Proc. International Conference on Automatic Face and Gesture Recognition, 1–6.

Li H, Buenaposada J & Baumela L (2008) Real-time facial expression recognition with illumination-corrected image sequences. Proc. International Conference on Automatic Face and Gesture Recognition, 1–6.

Li S, Chu R, Liao S & Zhang L (2007) Illumination invariant face recognition using near-infrared images. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4): 627–639.

Li X, Pfister T, Huang X, Zhao G & Pietikäinen (2013) A spontaneous micro-expression database: Inducement, collection and baseline. Proc. International Conference on Automatic Face and Gesture Recognition, 1–6.

Li Z, Imai J & Kaneko M (2009) Facial-component-based bag of words and PHOG descriptor for facial expression recognition. Proc. International Conference on Systems, Man, and Cybernetics, 1353–1358.

Lien J, Kanade T, Cohn J & Li C (1998) Automated facial expression recognition based on FACS action units. Proc. International Conference on Automatic Face and Gesture Recognition, 390–395.

Littlewort G, Bartlett M, Fasel I, Susskind J & Movellan J (2006) Dynamics of facial expression extracted automatically from video. Image and Vision Computing 24(6): 615–625.

Long F, Wu T, Movellan J, Bartlett M & Littlewort G (2012) Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. Neurocomputing 93: 126–132.

Lucey P, Cohn J, Kanade T, Saragih J & Ambadar Z (2010) The extended Conn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. Proc. International Conference on Computer Vision and Pattern Recognition, 94–101.

Lyons M, Akemastu S, Kamachi M & Gyoba J (1998) Coding facial expressions with gabor wavelets. Proc. International Conference on Automatic Face and Gesture Recognition, 200–205.

Maeng H, Choi H, Park U, Lee S & Jain A (2011) Nfrad: Near-infrared face recognition at a distance. Proc. International Joint Conference on Biometrics, 1–7.

Majumder A, Behera L & Subramanian V (2013) Facial expression recognition with regional features using local binary patterns. Proc. International Conference on Computer Analysis of Images and Patterns, 556–563.

Mao X, Li Y, Li Z, Huang K & Lv S (2009) Robust facial expression recognition based on RPCA and AdaBoost. Proc. Workshop on Image Analysis for Multimedia Interactive Services, 113–116.

Mehrabian A (1968) Communication without words. Psychology Today 2(4): 53–56.

Meng H, Romera-Paredes B & Bianchi-Berthouze N (2011) Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. Proc. International Conference on Automatic Face and Gesture Recognition, 854–859.

Mercier H, Peyras J & Dalle P (2007) Occluded facial expression tracking. Proc. Scandinavian Conference on Image Analysis, 72–81.

Milborrow S & Nicolls F (2008) Locating facial features with an extended active shape model. Proc. European Conference on Computer Vision, 504–513.

Moore S & Bowden S (2011) Local binary patterns for multi-view facial expression recognition. Computer Vision and Image Understanding 115(4): 541–558.

Nabatchian A, Abdel-Raheem E & Ahmadi M (2011) Illumination invariant feature extraction and mutual-information-based local matching for face recognition under illumination variation and occlusion. Pattern Recognition 44(10-11): 2576–2587.

Nanni L, Brahnam S & Lumini A (2011) Local ternary patterns from three orthogonal planes for human action classification. Expert Systems with Applications 38(5): 5125–5128.

Nanni L, Luminni A & Brahnam S (2010) Local binary patterns variants as texture descriptors for medical image analysis. Artificial Intelligence in Medicine 49(2): 117–125.

Nguyen H, Chen F, Kotani K & Le B (2014) Human emotion estimation using wavelet transform and t-ROIs for fusion of visible images and thermal image sequences. Proc. Computational Science and Its Application, 224–235.

Ojala T, Pietikäinen M & Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7): 917–987.

Ouyang Y & Sang N (2013) A facial expression recognition method by fusing multiple sparse representation based classifiers. In: Advances in Neural Networks, 479–488.

Päivärinta J, Rahtu E & Heikkilä J (2011) Volume local phase quantization for blur insensitive dynamic texture classification. Proc. Scandinavian Conference on Image Analysis, 360–369.

Petkov N & Subramanian E (2007) Motion detection, noise reduction, texture suppression and contour enhancement by spatiotemporal gabor filters with surround inhibition. Biological Cybernetics 97(5): 423–439.

Pfister T, Li X, Zhao G & Pietikäinen M (2011) Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. Proc. International Conference on Computer Vision Workshops, 868–875.

Pietikäinen M, Hadid A, Zhao G & Ahonen T (2011) Computer Vision using Local Binary Patterns. Springer.

Raducanu B & Dornaika F (2008) Dynamic vs. static recognition of facial expressions. Proc.

European Conference on Ambient Intelligence, 13–25.

Roy A & Marcel S (2009) Haar local binary pattern feature for fast illumination invariant face detection. Proc. British Machine Vision Conference, 1–12.

Rudovic O, Pantic M & Patras I (2013) Coupled Gassign processes for pose-invariant facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(6): 1357–1369.

Rudovic O, Patras I & Pantic M (2010) Regression-based multi-view facial expression recognition. Proc. International Conference on Pattern Recognition, 4121–4124.

Ruiz-Hernandez J & Pietikäinen M (2013) Encoding local binary patterns using the re-parametrization of the second order gaussian jet. Proc. International Conference and Workshop on Automatic Face and Gesture Recognition, 1–6.

Sánchez A, Ruiz J, Moreno A, Montemayor A, Hernández J & Pantrigo J (2011) Differential optical flow applied to automatic facial expression recognition. Neurocomputing 74(8): 1272–1282.

Sandbach G, Zafeiriou S, Pantic M & Yin L (2012) Static and dynamic 3D facial expression recognition: A comprehensive survey. Image and Vision Computing 30(10): 683–697.

Shan C & Braspenning R (2010) Recognizing facial expressions automatically from video. In: Handbook of Ambient Intelligence and Smart Environment, 479–509.

Shan C, Gong S & McOwan P (2005) Robust facial expression recognition using local binary patterns. Proc. International Conference on Image Processing, 370–373.

Shan C, Gong S & McOwan P (2009) Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27(6): 803–816.

Shin G & Chun J (2008) Spatio-temporal facial expression recognition using optical flow and HMM. In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 27–38.

Sikka K, Dykstra K, Sathyanarayana S, Littlewort G & Bartlett M (2013) Multiple kernel learning for emotion recognition in the wild. Proc. ACM on International conference on multimodal interaction, 517–524.

Smith R & Windeatt T (2010) Facial expression detection using filtered local binary pattern features with ECOC classifiers and platt scaling. Proc. JMRL Workshop on Applications of Pattern Analysis, 111–118.

Sonnenburg S, Rätsch G, Schäfer C & Schölkopf B (2006) Large scale multiple kernel learning. Journal of Machine Learning Research 7: 1531–1565.

Sun Y & Yin L (2009) Evaluation of spatio-temporal regional features for 3D face analysis. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 13–19.

Suwa M, Sugie N & Fujimora K (1978) A preliminary note on pattern recognition of human emotion expression. Proc. International Joint Conference on Pattern Recognition, 408–410.

Tan X & Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting condition. IEEE Transactions on Image Processing 19(6): 1635–1650.

Tang H, Hasegawa-Johnson M & Huang T (2010) Non-frontal view facial expression recognition based on ergodic hidden markov model super vectors. Proc. International Conference on Multimedia and Expo, 1202–1207.

Tang H & Huang T (2008) 3D facial expression recognition based on automatically selected features. Proc. International Conference on Computer Vision and Pattern Recognition, 1–8.

Tian Y, Kanade T & Cohn J (2002) Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. Proc. International Conference on Automatic

Face and Gesture Recognition, 229–234.

Tian Y, Kanade T & Cohn J (2005) Facial expression analysis. In: Li S & Jain AK (eds) Handbook of face recognition, Springer, 247–275.

Tian Y, Kanade T & Cohn J (2011) Facial expression recognition. In: Li S & Jain AK (eds) Handbook of face recognition, Springer, 487–519.

Towner H & Slater M (2007) Reconstruction and recognition of occluded facial expressions using PCA. Proc. International Conference on ACII, 36–47.

Tsalakanidou F & Malassiotis S (2010) Real-time 2D+ 3D facial action and expression recognition. Pattern Recognition 43(5): 1763–1775.

Turk M & Pentland A (1991) Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(11): 71–86.

Tzimiropoulos G, Zafeirious S & Pantic M (2012) Subspace learning from image gradient orientations. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(12): 2454–2466.

Valstar M, Jiang B, Mehu M & Pantic M (2011) The first facial expression recognition and analysis challenge. Proc. International Conference on Automatic Face and Gesture Recognition, 921–926.

Whitehill J, Bartlett M & Movellan J (2013) Automatic facial expression recognition. In: Gratch J & Marsella S (eds) Social Emotions in Nature and Artifact, Oxford University Press, 88–109.

Wu T, Bartlett M & Movellan J (2010) Facial expression recognition using Gabor motion energy filters. Proc. International Conference on Computer Vision and Pattern Recognition, 42–47.

Xia H, Xu R & Song S (2012) Robust facial expression recognition via sparse representation over overcomplete dictionaries. Journal of Computational Information Systems 8(1): 425–433.

Yacoob Y & Davis L (1996) Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(6): 636–642.

Yang H, Su H, Zheng S, Wei S & Fan Y (2011) The large-scale crowd density estimation based on sparse spatiotemporal local binary pattern. Proc. International Conference on Multimedia and Expo, 1–6.

Yang P, Liu Q & Metaxas D (2007) Boosting coded dynamic features for facial action units and facial expression recognition. Proc. Conference on Computer Vision and Pattern Recognition, 1–6.

Yesin M, Bullot B & Sharma R (2006) Recognition of facial expressions and measurement of levels of interest from video. IEEE Transactions on Multimedia 8(3): 500–508.

Yin L, Wei X, Sun Y, Wang J & Rostao M (2006) A 3D facial expression database for facial behavior research. Proc. International Conference on Automatic Face and Gesture Recognition, 211–216.

Yoshitomi Y, Miyawaki N, Tomita S & Kimura S (1997) Facial expression recognition using thermal image processing and neural network. Proc. International Workshop on Robot and Human Communication, 380–385.

Yu K, Wang Z, Zhuo L, Wang J, Chi Z & Feng D (2013) Learning realistic facial expressions from web images. Pattern Recognition 46(8): 2144–2155.

Yuce A, Sorci M & Thiran J (2013) Improved local binary pattern based action unit detection using morphological and bilateral filters. Proc. International Conference on Automatic Face and Gesture Recognition, 1–7.

Zafeirious S & Petrou M (2010) Sparse representation for facial expressions recognition via l1 optimization. Proc. International Conference on Computer Vision and Pattern Recognition, 32–39.

Zavaschi T, Britto A, Oliveira L & Koerich A (2013) Fusion of feature sets and classifiers for facial expression recognition. Expert Systems with Applications 40(2): 646–655.

Zeng Z, Pantic M, Roisman G & Huang T (2009) A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1): 39–58.

Zhang L, Tjondronegoro D & Chandran V (2011) Toward a more robust facial expression recognition in occluded images using randomly sampled Gabor based templates. Proc. International Conference on Multimedia and Expo, 1–6.

Zhang S, Zhao X & Lei B (2012) Robust facial expression recognition via compressive sensing. Sensors 12(3): 3737–3761.

Zhang W, Shan S, Gao W, Chen X & Zhang H (2005) Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. Proc. International Conference on Computer Vision, 786–791.

Zhang X, Mahoor M & Voyles R (2013) Facial expression recognition using hessianMKL based multiclass-SVM. Proc. International Conference on Automatic Face and Gesture Recognition, 1–6.

Zhang Y & Ji Q (2005) Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5): 699–714.

Zhang Z, Lyons M, Schuster M & Akamatsu S (1998) Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. Proc. International Conference on Automatic Face and Gesture Recognition, 454–459.

Zhao G, Ahonen T, Matas J & Pietikäinen M (2012) Rotation-invariant image and video description with local binary pattern features. IEEE Transactions on Image Processing 21(4): 1465–1477.

Zhao G & Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6): 915–928.

Zhao G & Pietikäinen M (2009) Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recognition Letters 30(12): 1117–1127.

Zheng W, Tang H, Lin Z & Huang T (2009) A novel approach to expression recognition from non-frontal face images. Proc. International Conference on Computer Vision, 1901–1908.

Zheng W, Tang H, Lin Z & Huang T (2010) Emotion recognition from arbitrary view facial images. Proc. European Conference on Computer Vision, 490–503.

Zhi R, Flierl M, Ruan Q & Kleijin W (2011) Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(1): 38–52.

Zhu X & Ramanan D (2012) Face detection, pose estimation and landmark localization in the wild. Proc. International Conference on Computer Vision and Pattern Recognition, 2879–2886.

Original articles

I Huang X & Zhao G & Hong X & Pietikäinen M & Zheng W (2013) Texture description with completed local quantized patterns. In: Image Analysis, SCIA 2013 Proceedings, Lecture Notes in Computer Science, 7944:1-10.

II Huang X & Zhao G & Zheng W & Pietikäinen M(2012) Spatiotemporal local monogenic binary patterns for facial expression recognition. IEEE Signal Processing Letters, 19(5):243-246.

III Huang X & Zhao G & Pietikäinen M & Zheng W (2010) Dynamic facial expression recognition using boosted component-based spatiotemporal features and multi-classifier fusion. In: Advanced Concepts for Intelligent Vision Systems, ACIVS 2010 Proceedings, Lecture Notes in Computer Science, 6475:312-322.

IV Zhao G & Huang X & Taini M & Li SZ & Pietikäinen M (2011) Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619.

V Huang X & Zhao G & Pietikäinen M & Zheng W (2011) Expression recognition in videos using a weighted component-based feature descriptor. In: Image Analysis, SCIA 2011 Proceedings, Lecture Notes in Computer Science, 6688:569-578.

VI Huang X & Zhao G & Zheng W & Pietikäinen M (2012) Towards a dynamic expression recognition system under facial occlusion. Pattern Recognition Letters, 33(16): 2181-2191.

VII Huang X & Zhao G & Pietikäinen M (2013) Emotion recognition from facial images with arbitrary views. Proc. the British Machine Vision Conference (BMVC 2013): 76.1-76.11.

Reprinted with permission from Springer (I, III, V), IEEE(II), Elsevier (IV, VI), and BMVA (VII).

Original publications are not included in the electronic version of the dissertation.

ACTA UNIVERSITATIS OULUENSIS

SERIES C TECHNICA

492. Sliz, Rafal (2014) Analysis of wetting and optical properties of materials developed for novel printed solar cells

493. Juntunen, Jouni (2014) Enhancing organizational ambidexterity of the Finnish Defence Forces' supply chain management

494. Hänninen, Kai (2014) Rapid productisation process : managing an unexpected product increment

495. Mehtonen, Saara (2014) The behavior of stabilized high-chromium ferritic stainless steels in hot deformation

496. Majava, Jukka (2014) Product development : drivers, stakeholders, and customer representation during early development

497. Myllylä, Teemu (2014) Multimodal biomedical measurement methods to study brain functions simultaneously with functional magnetic resonance imaging

498. Tamminen, Satu (2014) Modelling the rejection probability of a quality test consisting of multiple measurements

499. Tuovinen, Lauri (2014) From machine learning to learning with machines : remodeling the knowledge discovery process

500. Hosio, Simo (2014) Leveraging Social Networking Services on Multipurpose Public Displays

501. Ohenoja, Katja (2014) Particle size distribution and suspension stability in aqueous submicron grinding of $CaCO_3$ and TiO_2

502. Puustinen, Jarkko (2014) Phase structure and surface morphology effects on the optical properties of nanocrystalline PZT thin films

503. Tuhkala, Marko (2014) Dielectric characterization of powdery substances using an indirectly coupled open-ended coaxial cavity resonator

504. Rezazadegan Tavakoli, Hamed (2014) Visual saliency and eye movement : modeling and applications

505. Tuovinen, Tommi (2014) Operation of IR-UWB WBAN antennas close to human tissues

506. Vasikainen, Soili (2014) Performance management of the university education process

507. Jurmu, Marko (2014) Towards engaging multipurpose public displays : design space and case studies

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-952-62-0655-4 (Paperback)
ISBN 978-952-62-0656-1 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)

UNIVERSITY of OULU
OULUN YLIOPISTO