

# Face Frontalization for Facial Expression Recognition in the Wild



Yiming Wang

School of Creative Technologies

University of Portsmouth

A thesis submitted for the degree of

*Doctor of Philosophy*

September 2018

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Word Count: 29,474



## **Acknowledgements**

Foremost, I would like to express my sincere gratitude to my supervisor, Prof. Hui Yu, for his constant support during these four years of PhD study. He taught me how to think like a scientist. Further thanks to my second and third supervisor Dr. Brett Stevens and Dr. Neil Dansey for their valuable help and guidance. They gave me many inspirations during my research. Equivalent thanks to Prof. Honghai Liu and Dr. Zhaojie Ju, who provide me helps after my arrival in the United Kingdom. I would like to thank other colleagues of the Visual Computing Group. A big thanks to Mr Jianwen Lou, Dr. Shu Zhang, Miss Qiongdan Cao, Mr. Martin Kearn, Ms. Xiaoxu Cai, Ms. Jianyuan Sun, Mr. Hao Fan, Mr. Yifan Xia for the great moments that we have shared. I would also like to thank Mr. Dalin Zhou, Mr. Haibin Cai, Dr. Dongxu Gao, Ms. Ting Wang, Mr. Peter Boyd, Mr. Uche Ogenyi, Mr. Charles Phiri, Ms. Bangli Liu, Miss. Kairu Li for sharing their knowledge with me. I am grateful to my parents for their unconditional love throughout all my studies. Special thanks to Weihong Gao, who stood by me all the moments.

## **Abstract**

Automatic machine analysis of facial expressions has attracted increasing attentions and has been widely applied to various domains such as animation, multimedia and security. Sensing and understanding facial behaviours are the fundamental requirement of human-machine interaction systems. As computing has become more powerful and ubiquitous, there is an urgent demand for the facial expression recognition system which is designed under real-world conditions and enables generalization across population. The proposed work in this thesis addresses four main challenges of facial expression recognition in the wild: 1) identity bias which refers to the fact that facial features are always discriminative in terms of identity but difficult to distinguish in terms of expressions, 2) head pose variations, 3) occlusions, and 4) irregularity of spontaneous expressions and presents approaches to tackle these challenges. Inspired by the success of existing research and benchmarks of facial expression recognition under controlled conditions where identity bias is the only challenge and there are very small or even no variations in terms of the other three problems, we propose to normalize the facial images under unconstrained situations into lab conditions by presenting spatial face normalization and texture reconstruction based on face frontalization. The goal is to design a powerful and flexible system which can improve the facial expression recognition performance under unconstrained real-world conditions.

We introduce a novel Facial Expression-Aware face Frontalization (FEAF) method based on spatial normalization strategy. Compared with most existing methods that only addressed one or several challenges, a joint consideration of all the four challenges is taken into account. To effectively solve the problem of identity bias and irregularity of expressions,

we present a multi-template model to normalize shape variations deliberately designing multiple frontal shape templates that contain meaningful expressions to fit in with various shapes of facial expressions. Hence, every face is aligned to one of a group of shared template no matter how ambiguous or who the input face is. Subsequently, shape normalization that map the facial shape to a normalized frontal emotional template in order to solve head-pose variations. Finally, we have employed face frontalization techniques to reconstruct facial appearances, where occlusions are removed by maintaining an additional error matrix to restore sparse errors caused by occlusions. The reconstructed faces will be strictly in frontal view. Given the reconstructed faces, some commonly used feature extraction methods and machine learning techniques can be employed for facial emotional states recognition. The state-of-the-art performance is achieved in the task of static facial expression recognition in the wild. We also have demonstrated the superior performance of our proposed models on the task of interpersonal relation prediction.

To capture more subtle facial expression cues and further improve recognition rate, we propose an extended FEAF approach for dynamic facial expression analysis based on accurate shape processing. Contrary to the majority of existing dynamic methods that focus on time alignment, the consideration of head pose variations and occlusions are addressed in this method by using regression-based spatial alignment that estimates an accurate frontal view of facial shape given the a non-frontal face. We develop a cascade regression model to learn the pair-wise relationship between non-frontal facial shape and its frontal counterpart. Different from static FEAF, this method can capture subtle facial muscle changes in an image sequence and, therefore, it can be used for dynamic facial expression recognition. Superior performance has been achieved on several public datasets under both lab and unconstrained conditions.

# Contents

<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problems . . . . .	2
1.1.1 Facial Expression Recognition . . . . .	2
1.1.2 FER: Controlled vs. Unconstrained Conditions . . . . .	3
1.2 Challenges . . . . .	5
1.2.1 Multi-View and View-Invariant FER . . . . .	6
1.2.2 Person-Independent FER . . . . .	8
1.2.3 Face Frontalization . . . . .	9
1.2.4 Research Gap of FER in the Wild . . . . .	10
1.3 Contributions . . . . .	12
1.4 Outline of Thesis . . . . .	13
<b>2 Literature Review</b>	<b>15</b>
2.1 Facial Expression Analysis . . . . .	15
2.1.1 Facial Expression Models and Benchmarks . . . . .	15
2.1.2 Facial Expression Analysis . . . . .	19
2.1.3 FER Applications . . . . .	23
2.2 Facial Frontalization . . . . .	25
2.2.1 Face Normalization . . . . .	26
2.2.2 Face Generation . . . . .	28

2.3	Relation to Our Work . . . . .	30
2.4	Summary . . . . .	31
<b>3</b>	<b>Dynamic FER Under Controlled Conditions Using Key Features</b>	<b>32</b>
3.1	Introduction . . . . .	32
3.2	Related Works . . . . .	34
3.3	Method . . . . .	37
3.3.1	Facial Landmark Detection . . . . .	38
3.3.2	Extract Localized Patches . . . . .	38
3.3.3	Temporal Feature Extraction . . . . .	39
3.3.4	Emotion Recognition . . . . .	41
3.4	Experiment . . . . .	41
3.4.1	Dataset . . . . .	41
3.4.2	Evaluation . . . . .	42
3.4.3	Discussion . . . . .	44
3.5	Conclusion . . . . .	44
<b>4</b>	<b>Facial Expression-Aware Face Frontalization for Static FER in the Wild</b>	<b>46</b>
4.1	Introduction . . . . .	47
4.2	Related Works . . . . .	50
4.3	FEAF Model . . . . .	52
4.3.1	Initial Design of Multi-template Model . . . . .	52
4.3.2	Template Design and Template Matching . . . . .	59
4.3.3	Texture Reconstruction . . . . .	64
4.4	FEAF for Interpersonal Relation Estimation . . . . .	70
4.5	Experiment . . . . .	72
4.5.1	Database and experimental design . . . . .	72
4.5.2	Face frontalization . . . . .	73
4.5.3	Facial expression recognition . . . . .	76
4.5.4	Interpersonal Relation Prediction . . . . .	77
4.6	Summary . . . . .	78



<b>5</b>	<b>Cascade Regression-Based Face Frontalization for Dynamic FER</b>	<b>85</b>
5.1	Introduction . . . . .	85
5.2	Related Works . . . . .	88
5.3	Methodology . . . . .	91
5.3.1	Regression Approaches . . . . .	91
5.3.2	Cascade Regression Model . . . . .	93
5.4	Experiment . . . . .	95
5.4.1	Training data collection . . . . .	95
5.4.2	Spatial Alignment . . . . .	97
5.4.3	Static FER on SFEW . . . . .	98
5.4.4	Dynamic FER on AFEW . . . . .	98
5.5	Summary . . . . .	102
 <b>6</b>	 <b>Conclusions</b>	 <b>103</b>
6.1	Overview . . . . .	103
6.2	Contributions . . . . .	104
6.2.1	Patch-based Person-independent FER . . . . .	104
6.2.2	Static Model of Facial Expression-Aware Face Frontalization .	104
6.2.3	Dynamic Model of Facial Expression-Aware Face Frontalization	105
6.3	Future Work . . . . .	105
6.3.1	Facial Expression Applications . . . . .	106
6.3.2	Facial Expression synthesis . . . . .	106
 <b>References</b>		 <b>107</b>
 <b>A</b>	 <b>Publications</b>	 <b>124</b>
 <b>B</b>	 <b>Research Ethics</b>	 <b>125</b>

# List of Figures

1.1	Challenges of FER in unconstrained conditions . . . . .	6
1.2	Spontaneous vs Posed expressions . . . . .	7
2.1	Six basic emotions . . . . .	17
2.2	Valence and arousal dimensional model . . . . .	18
2.3	Process of traditional FER architecture . . . . .	19
2.4	Face Frontalization = Frontal shape estimation + texture fitting . . . . .	26
3.1	Black squares indicate weight 0.0, dark gray 1.0, light gray 2.0, and white 4.0. (a) Subregions were weighted for face recognition. (b) The blocks were weighted for facial expression recognition. . . . .	33
3.2	The outline of proposed system . . . . .	35
3.3	Local patch extraction: the first patch centers around point number 11, the second centers around point number 14, point number 32 and 38 are used to design the bounding box of the third patch. . . . .	39
3.4	Examples of some uniform patterns that show meaningful local structures	40
4.1	Relations between universal emotions and AUs . . . . .	52
4.2	FACS AUs on eyebrow and eye regions . . . . .	53
4.3	FACS AUs on mouth regions . . . . .	54
4.4	Principal shape template of eyebrow behaviours . . . . .	56
4.5	Initial shape template (part 1) . . . . .	57
4.6	Initial shape template (part 2) . . . . .	58
4.7	Template Matching . . . . .	62
4.8	Eight templates of face shape . . . . .	63
4.9	Eigen faces . . . . .	66

## LIST OF FIGURES

---

4.10 Texture Fitting . . . . .	66
4.11 Interpersonal relation traits . . . . .	70
4.12 Approach for interpersonal relation prediction . . . . .	71
4.13 Face frontalization on unconstrained images . . . . .	73
4.14 Face frontalization . . . . .	75
5.1 Overview of the proposed method . . . . .	87
5.2 Cascade regression for frontal shape estimation . . . . .	95
5.3 Landmark positions: 2D vs 3D . . . . .	96
5.4 3D rendering . . . . .	96

# List of Tables

1.1	Controlled vs. Unconstrained . . . . .	4
3.1	Confusion matrix (%) of 6-class facial expression recognition for this work . . . . .	43
3.2	Comparison of different methods for “leave one subject out” cross-validation . . . . .	43
3.3	Comparison of different methods for “leave one subject out” cross-validation and person-independent validation . . . . .	43
4.1	Measurement of initial template matching on SFEW databse . . . . .	60
4.2	The configuration of CNN . . . . .	79
4.3	Comparison of generic Frontalization methods . . . . .	80
4.4	Confusion matrix (%) on SFEW databse . . . . .	81
4.5	Comparison of recognition rate (%) of generic frontalization methods on SFEW databse . . . . .	82
4.6	Comparison of recognition rate (%) of the state-of-the-art methods on SFEW databse . . . . .	83
4.7	Comparison of relation traits prediction performance . . . . .	84
5.1	Alignment Error (%) of different regression methods . . . . .	97
5.2	Comparison of recognition rate (%) of the state-of-the-art methods on SFEW databse . . . . .	99
5.3	Comparison of recognition rate (%) of the state-of-the-art methods on AFEW databse . . . . .	100

# List of Abbreviations

- 3DMM** 3D Morphable Model
- AAM** Active Appearance Model
- ADMM** Alternating Directions Method of Multipliers
- AFEW** Acted Facial Expressions In The Wild
- AI** Artificial Intelligence
- ALM** Augmented Lagrangian Method
- AR** Augmented Reality
- ASM** Active Shape Model
- AU** Action Unit
- DCNN** Deep Convolutional Neural Network
- DNN** Deep Neural Network
- EmotiW** Emotion Recognition in the Wild Challenge
- FACS** Facial Action Code System
- FEAF** Facial Expression-Aware face Frontalization
- FER** Facial Expression Recognition
- FFD** Free Form Deformation
- GAN** Generative Adversarial Network
- GP** Gaussian Process
- GPR** Gaussian Process Regression

**HCI** Human Computer Interaction

**HOG** Histogram of Gradients

**ITS** Intelligent Tutorial System

**LBP** Local Binary Patterns

**LBP-TOP** Local Binary Pattern on Three Orthogonal Planes

**LPQ** Local Phase Quantization

**LPQ-TOP** Local Phase Quantization on Three Orthogonal Planes

**MVP** Multi-View Perceptron

**PCA** Principal Component Analysis

**PIFR** Pose-Invariant Face Recognition

**RBF** Radial Basis Function

**ROI** Regions of Interest

**RSF** Robust Statistical face Frontalization

**SDM** Supervised Decent Method

**SFEW** Statistical Facial Expression in the Wild

**SIFT** Scale-Invariant Feature Transform

**SVM** Support Vector Machine

**SVR** Support Vector Regression

**TOP** Three Orthogonal Planes

**UMM** Universal Manifold Model

**VR** Virtual Reality

# Chapter 1

## Introduction

Facial expressions are human beings' naturally and cognitively affective responses to situations or contexts. They convey massive information of a person with regards to affects during communication. In 1967, Mehrabian and his colleagues (Mehrabian & Wiener, 1967) (Mehrabian & Ferris, 1967) presented two studies on how human decided whether they liked or disliked one another. The studies were based on human "positive versus negative" affective response, where Mehrabian was seeking for the relative impact of facial expressions, tone of voice and spoken words. By these studies, a well-known "7%-38%-55% Rule" was provided: words account for 7%, tone of voice accounts for 38%, and facial expression accounts for 55%. When communicators were talking about their feelings or attitudes, this rule suggested that human beings preferred to trust the information conveyed by facial expressions rather than the literal meaning of words. Therefore, facial expression is one of the most important visual features for human beings to convey various subtle signals of affects.

The origin of facial expression analysis can be traced back to 19th century when this normally conducted by Darwin. As computing has and will continue to be more powerful and ubiquitous, it is now much more common to capture and process facial affective features by machines. Facial Expression Recognition (FER) aims to automatically identify facial muscle behaviours or affective states of human beings from digital images or video sequences. FER has been gaining significant attention over past years due to various applications in security, psychology, automatic counselling, music for mood, animation etc.

This chapter describes problems and challenges of FER. Then the main contributions and the outline of this thesis are presented.

## 1.1 Problems

### 1.1.1 Facial Expression Recognition

Facial expressions are believed to be one of the most natural means for human beings to transmit information about intention and emotion. Facial expressions convey the most enriched non-verbal cues and, thus, they play an important role in interpersonal communication and human-environment interaction. There is a growing amount of evidence showing that emotional skills of facial expressions are part of human intelligence. Automatic machine recognition of facial expressions has attracted increasing interests for over two decades. Considering the importance of facial expressions in human communication, facial behaviour understanding is becoming a fundamental requirement of Human Computer Interaction (HCI).

The ultimate goal of FER is to construct an intelligent system for automatic analysis of human affects. Although there are some existing works for FER based on biological sensors, vision-based approaches are still the mainstream. Traditional approaches for visual recognition of facial expressions mostly focus on two main tasks: six basic emotion recognition and Action Unit (AU) recognition. Six basic emotions refer to angry, disgust, fear, happy, sad and surprise and AUs are systematically encoded facial muscle movements. Recent works on vision-based FER have extended to various attractive research topics, such as continuous emotion intensity estimation (Zafeiriou *et al.*, 2016), compound facial expression recognition (Du *et al.*, 2014), 3D/4D facial expression recognition (Yin *et al.*, 2006, 2008), pain monitoring (Lucey *et al.*, 2011a,b), driver drowsy detection (Weng *et al.*, 2016), among others. Conventionally, the architecture of vision-based FER often involves three steps: 1) face registration, 2) feature extraction and representation, 3) recognition.

Face registration is the preprocessing of FER, which registers all the faces in the same coordinate system. Whole face registration is the most commonly used strategy where facial areas are detected and cropped from images in order to preserve the whole facial features. Many existing works use an open source face detector (such as



OpenCV) for this task. Alternatively, some systems use the position of facial fiducial points (such as the coordinates of eye corner / center and nose tip) to localize and align faces based on point detectors. Note that crucial cues of facial expressions are often conveyed by eyes, eyebrows and mouth but not the whole face. So, there are a few approaches based on part registration strategy which only registers facial Regions of Interest (ROI) instead of the whole facial region.

The objective of feature extraction and representation is to quantify a facial image into a meaningful and discriminative feature vectors. Facial extraction methods can be classified into appearance-based approaches and geometric-based approaches. Appearance-based approaches extract local low-level texture features and present statistical representation. Currently, the best known texture feature descriptors are Local Binary Patterns (LBP) and Local Phase Quantization (LPQ). Geometric-based approaches represent a face by straightforwardly concatenating the coordinates of all the fiducial landmark points.

Given a facial image and its derived feature representation, the final step is recognition which classifies this face into one of the expression categories. In this step, machine learning techniques, such as Support Vector Machine (SVM), are usually employed.

The above mentioned structure is a traditional learning paradigm where features are extracted according to human-designed rules (so-called handcrafted) and classification step is independent to feature extraction step. The recent booming of deep learning paradigm has shown a novel framework where feature extraction and recognition steps can be performed jointly. Correspondingly, deep learning is a completely data-driven system that is not associated to any handcrafted rules or human expert knowledge.

### 1.1.2 FER: Controlled vs. Unconstrained Conditions

Most existing works of FER were conducted on the facial expression images under lab conditions and could often achieve good recognition results. However, those FER models trained on controlled facial expression datasets are not practical since the performance will drop significantly when they are applied to real-world images captured under various unconstrained conditions. In recent years, research of FER has started to addresses “in-the-wild” problem and some progress has been made. Nowadays, FER

Table 1.1: Controlled vs. Unconstrained

Dataset		Typical results	
Controlled	CK+	96.26%(T)	
	MMI	93.33%(T)	
	JAFFE	94.76%(T)	
Unconstrained	SFEW (EmotiW2015)	39.13%(T)	53.80%(DL)
	AFEW (EmotiW2015)	39.33%(T)	56.16%(DL)
	FER2013	71.16%(DL)	

in the wild is still challenging because of complicated backgrounds, illumination variations, occlusions, and particularly, large variations in person-specific appearance and head pose.

The datasets of controlled lab conditions involves posed expressions, in which the subjects (often trained actors) were instructed to act a specific emotion (often without emotion stimuli). The expressions are performed in a regularized way and the expression intensity is often in high levels. Furthermore, the subject’s face was captured in frontal or near frontal view and there was uniform illumination condition and no occlusions. Since so many factors were artificially manipulated, the recognition performance of many approaches on these kind of datasets was always impressively good.

The in-the-wild facial expression images contain spontaneous expressions captured in real-world conditions. In this kind of unconstrained image dataset, there are various changes in head pose, illumination and occlusions (as is shown in Fig. 1.1), which are non-linear factors for machine learning approaches. Beside, another difference between controlled and unconstrained images is that posed expressions are often exhibited in high intensities, in which there are significantly visible facial muscle movements. Whilst, spontaneous expressions are naturally performed in various intensities, which results in many more complex and subtle changes in facial appearances. The visual comparison between posed and spontaneous expressions is shown in Fig. 1.2, in which the images were collected from public datasets (CK+ and SFEW). From this figure we can see two main differences between posed and spontaneous expressions. Firstly, the intensity degree of posed expressions are high. For example, a surprise emotion in posed expressions always involves remarkably wide opening of eyes and

mouth, while a spontaneous surprise is not always exhibited in terms of mouth widely open. Secondly, posed expressions often follow a uniformed pattern while spontaneous expressions are irregular. An obvious example is that a happy emotion commonly involves AU12 (lip corner puller) but Fig. 1.2 shows a spontaneous happy emotion with AU15 (lip corner depressor) which is often associated with sad. Therefore, spontaneous expressions contain much more subtle facial muscle changes, which makes it more difficult to analyze them.

Table 1.1 shows several typical results obtained from several facial expression datasets under controlled or unconstrained conditions. In this table, DL represents deep learning methods and T is short for traditional handcrafted feature-based methods. All the traditional methods are based on LBP + SVM. The results of deep learning methods are the winners EmotiW2015 and FER2013 competitions (Kim *et al.*, 2015; Tang, 2013; Yao *et al.*, 2015). From this table we can see that the recognition rate on controlled facial datasets are much higher than in-the-wild datasets no matter whether traditional method or deep learning method is used. Meanwhile, deep learning methods do not show significantly superior results compared to traditional methods. Considering that deep learning methods have achieved remarkably good performance in many applications (e.g. object detection, face recognition), the results of FER is not so impressive. Training a deep model requires large labelled facial expression images which are not always readily available. What is worth mentioning is that a good deep learning model is empirically obtained by training on millions of labelled examples. For example, ImageNet for object detection contains over 14 millions images and Google FaceNet model for face recognition is trained on over 200 million faces. However, there is no such large publicly available dataset for facial expression analysis. Therefore, deep learning facial expression approaches are currently not prominent and the mainstream of FER research is still based on the traditional paradigm.

## 1.2 Challenges

As discussed in Section 1.1.2, the problems of FER in the wild are that a) there are so many non-linear factors for traditional approaches and b) deep learning facial expression approaches suffer from insufficient training data. By comparing FER under controlled and unconstrained conditions, the main challenges of facial expression analysis

can be summarized as follows 1) identity bias, 2) head pose variations, 3) occlusions, and 4) irregularity of spontaneous expressions. Although illumination change is also one of the main non-linear factors, it is not so challenging as the other four issues since most commonly used feature extraction methods are inherently robust to illumination changes.



Figure 1.1: Challenges of FER in unconstrained conditions

Identity bias is reported as a common challenge, where facial visual features always contains more identity-related information but not expression-related cues. There are some existing works on how to extracted expression-enriched facial features and conducted person-independent FER methods. Another common non-linear factor is head-pose variation which refers to various out-of-plane rotations of faces. Several multi-view & view-invariant FER approaches have been presented to model the head-pose variations. Face frontalization is a newly rising technique for view-invariant facial analysis, which aims to reconstruct a frontal view of the face given a profile facing view. Face frontalization approaches were originally proposed for the face recognition task but not FER, but it is highly associated to the topic of this thesis. Therefore, we will elaborate on multi-view & view-invariant, person-independent FER and face frontalization in the rest of this section.

### 1.2.1 Multi-View and View-Invariant FER

Multi-view or view-invariant FER aims to perform expression recognition on both frontal and non-frontal faces. Moore et al. (Moore & Bowden, 2011) divided the






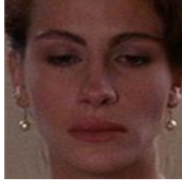



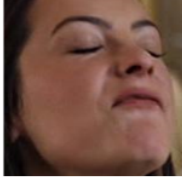
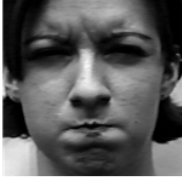
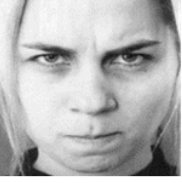




Emotions	Spontaneous		Posed	
Happy				
Sad				
Angry				
Surprise				

Figure 1.2: Spontaneous vs Posed expressions

face images into several discrete viewpoints and matched head pose to its closest view angle. Then they facilitate LBP (and its variants) feature extraction and expression recognition in each specific viewpoint. It is obvious that a pose estimation step must be taken first, so the whole system has to be trained per viewpoint/person/expression. The accuracy of this approach will drop dramatically when training samples are insufficient. Similar approaches can refer to (Eleftheriadis *et al.*, 2015; Hesse *et al.*, 2012; Kumano *et al.*, 2009).

Rudovic *et al.* (Rudovic *et al.*, 2013) present a Coupled Gaussian Processes regression model for pair-wise viewpoint normalization in which non-frontal facial geometric features can be normalized to their frontal viewpoint. The principal advantage

of this approach is that it can estimate facial expression categories in an untrained viewpoint. However, the accuracy is relatively low because this approach only use normalized geometric features, whilst texture features were withdrawn.

### 1.2.2 Person-Independent FER

The main motivation for person-independent FER is to remove the variations in person-specific appearance and establish an expression-enhanced facial representation in order to facilitate the recognition of new (untrained) faces.

Person-independent FER addresses the problem of identity bias. Identity bias means the extracted features preserve more identity-related cues rather than expression (Sariyanidi *et al.*, 2015). Geometric-based approaches have weak tolerance against individual variations. But for appearance-based approaches, this problem remains.

To date, only a few works have been proposed to deal with identity bias. The main approaches for person-independent FER can be classified into three categories: 1) introducing part registration, 2) extending to temporal model, 3) proposing a subsequent system.

In contrast to whole face registration, part registration only focuses on “interest regions”. This strategy selects the interest regions by incorporating domain experts knowledge. Shan *et al.* (Shan *et al.*, 2009) exploited a region weighted method to highlight the interest regions. In this method, the whole face is divided into several blocks and each block is weighted according to its contribution to expression recognition. Xue *et al.* (Xue *et al.*, 2013) used fusion features in which the MEb (Mouth and Eyebrow) features are addressed based on the fact that mouth and eyebrows possess the principal information related to expressions. Many part-registration-based researches did not address person-independent validation (Zhang & Tjondronegoro, 2011) (Nicolle *et al.*, 2012) (Zhu *et al.*, 2011) (Jeni *et al.*, 2013), but it is obvious that these methods focus on removing individual differences in texture.

Spatio-temporal models introduce temporal dependency in image sequences and will result in dynamic facial expression analysis. Spatio-temporal approaches can model facial muscle deformations, which enable them to capture more subtle expressions than spatial approaches. It is reported that Spatio-temporal approaches outperform spatial ones (Sariyanidi *et al.*, 2015). Zhao *et al.* (Zhao & Pietikainen, 2007)

extended LBP features to Orthogonal Planes (TOP) and conduct a spatio-temporal representation. LBP-TOP achieves good performance on person-independent validation. Other works on person-independent FER using dynamic texture analysis include (Jiang *et al.*, 2014; Koelstra *et al.*, 2010; Yeasin *et al.*, 2006).

Presenting a multi-layer recognition system is an effective strategy to reduce identity bias. Most previous works introduce a feature selection step before recognition. AdaBoost and GentleBoost are the most frequently used methods for feature selection. The combination of AdaBoost / GentleBoost and Support Vector Machine (SVM) has been used for both AU recognition (Jiang *et al.*, 2014; Valstar & Pantic, 2012) and emotion recognition (Shan *et al.*, 2009). Principal Component Analysis (PCA) is another useful technique for this task. Mohammadi *et al.* (Mohammadi *et al.*, 2014) propose a sparse dictionary representation based on (PCA). Cross database validation verifies that this method is less dependent on identity.

A few person-independent experiments were conducted by applying Gabor features (Gu *et al.*, 2012; Shojaeilangari *et al.*, 2011) or 3D models (Tekguc *et al.*, 2009). There are also a few approaches using geometric features to complement appearance representations (Chew *et al.*, 2011; Xue *et al.*, 2013) in order to reduce identity bias.

### 1.2.3 Face Frontalization

Recently, face frontalization has attracted wide attention due to its effectiveness in facial analysis. It is commonly accepted that more robust features can be captured from frontal faces rather than profile faces. Thus, the main objective of face frontalization is to recover the frontal faces from non-frontal viewpoints. Meanwhile, there are also several extended face frontalization approaches that generate facial images in not only frontal view, but also other views in order to capture more facial features. In general, face frontalization includes two key components: frontal facial shape estimation and frontal facial texture fitting.

Frontal shape estimation starts from facial landmark detection. Recently, many breakthroughs have been made on automatic facial landmark detection [1,2,3,4]. The objective of frontal shape estimation is to align the non-frontal facial landmarks to their frontal positions. Then, frontal texture-fitting recovers facial appearances by texture warping and rectification. When the non-frontal facial textures are overlaid on a frontal

face mesh, there will be one half face (left or right side of face) with rich pixels and some regions the other half without visible pixels. The task of texture fitting is to approximate the pixels on these regions and rectify the whole facial textures.

There are currently two main problems in the existing face frontalization approaches: 1) it is difficult to achieve real-time performance 2) most methods could only be used for face recognition but not FER.

Face frontalization is difficult to work in real-time because most approaches are unsupervised so that it takes a long time to perform the optimization. Meanwhile, a majority of approaches achieve only person-specific face frontalization in which novel subjects cannot be normalized to their frontal view. Therefore, they are not suitable for FER because a good FER system should work well on any unseen faces. In the existing generic (person-independent) face frontalization methods, a rough solution to frontal shape estimation is presented in which an unmodified shape template (in frontal view) is used as reference for all the query images and then texture-fitting is performed based on this single template (Hassner *et al.*, 2015) (Sagonas *et al.*, 2015). This strategy is called hard frontalization in which the reconstructed frontal faces will share a common 2D/3D face shape. The template is often made in a neutral shape as a compromise. However, the facial expression cues are ignored when reconstructing frontal face. It is a challenging task to remain or recover facial expressions during the process of face frontalization.

As far as we know, there is no attempt so far that performs face frontalization with full considerations of facial expressions. To this end, we present a novel approach that develops an supervised approach for real-time face frontalization and combine the expert knowledge of AUs to achieve a facial expression-aware approach for FER in the wild.

### 1.2.4 Research Gap of FER in the Wild

1. Identity bias is the most common problem of FER. There are only a few works on this topic. Salient facial region registration and temporal model have been reported to be most effective. But these two strategies are always investigated separately. No previous work has been done to combine them together. And



there are only a few reports or discussions for the advantages and disadvantages of the two strategies.

2. Facial expression recognition in the wild is a challenging task and most existing methods have achieved very low recognition performance. Deep learning facial expression approaches suffer from insufficient training data. Traditional paradigms will face the four challenges as discussed in Section 1.2 and the existing methods, like multi-view & view-invariant and person-independent approaches, only consider one or several of the four non-linear factors. Currently, no existing methods has proposed to jointly consider all the four non-linear factors. Face frontalization is highly related to the topic of this research due to its effectiveness in tackling head-pose variations and occlusion, and potential capability of reducing the influence of identity bias. But it was only applied for face recognition and there is no attempt so far performing face frontalization with full considerations of facial expressions.
3. Facial expressions are inherently dynamic behaviours and dynamic facial expression recognition methods capture facial features regarding movements. A video clip of facial expression often involves five sequential stages which start from neutral, then onset, till peak emotion, followed by offset and finally back to a neutral face. It is obvious that the durations a specific stage in different clips are usually quite different. Therefore, dynamic FER approaches often consider the registration of time changes, in which different video clips are aligned and registered to a common temporal template in order to normalize the non-linear factors caused by temporal variations. However, most existing approaches ignored spatial changes of head-pose and occlusions which are reported as the main challenges of FER. There is no works on dynamic FER based on spatial alignment. At this point, face frontalization is effective in spatial alignment that aligns faces into a common spatial template, which is well suited to static facial analysis but not dynamic facial expression analysis. Sequential face frontalization for dynamic FER is currently still an unexplored area of research.

### 1.3 Contributions

Accordingly, three main contributions of this thesis are listed below.

1. Firstly, A study of the identity bias problem is presented in which a) a novel method is proposed, which combines salient region registration and temporal feature extraction to derive a person-independent emotion-enhanced facial representation; b) this method is validated on a facial expression dataset under controlled conditions and the experimental result shows superior performance on person-independent validation; c) based on the experimental result, a detailed discussion is provided, which paves the way to the following research on face frontalization.
2. This thesis proposes a novel Facial Expression-Aware face Frontalization (FEAF) method which has three main contributions. 1) we make a comprehensive study of the difference between controlled facial expression images and in-the-wild images and illuminate the non-linear factors for FER in-the-wild. Accordingly, a clear idea is presented that better results can be obtained by normalizing facial images under unconstrained conditions into controlled conditions. FEAF is the first work that jointly considers all the non-linear factors of FER, which is theoretically superior than the existing FER approaches that only take one or several factors into account. 2) This work takes the advantages of face frontalization in which a realistic clean (without occlusions) frontal face can be reconstructed. At the same time, a novel multi-template model and template matching method are developed in which the vivid facial expression cues are well maintained, which fills the gap of face frontalization that often loses information of expressions in reconstructed faces. The proposed template-based face reconstruction strategy provides a better solution to identity bias than salient region registration-based approaches. The experimental results show FEAF outperforms any other small-sample learning methods. 3) a novel FEAF-based deep learning model is proposed which achieves state-of-the-art performance on a large-scale dataset.
3. An extended face frontalization method based on cascade regression is proposed for dynamic FER. Different from existing dynamic FER approaches which focus

on time alignment, this method considers spatial alignment that uses regression-based method to recover all the non-frontal faces to frontal views without using any templates. It takes advantage of face frontalization that is robust to head-pose and occlusions, and applies it to dynamic FER to reduce the affect of identity bias.

## 1.4 Outline of Thesis

The rest of the thesis is organised as follows.

**Chapter 2** surveys the state-of-the-art FER and face frontalization technologies. This chapter reviews physiologically emotional models and computer vision methods for automatic recognition of facial expressions. Then, face frontalization methods are reviewed. In the end, this chapter outlines some research challenges and future directions, as well as the potential solutions to these challenges. The aim of this chapter is to provide readers a systemic and comprehensive understanding of the background of this thesis.

**Chapter 3** studies the common problem of identity bias and present a framework for person-independent FER. This chapter proposes to combine “interest regions” detection and spatio-temporal features for person-independent facial expression analysis under controlled conditions.

**Chapter 4** presents a novel facial expression-aware face frontalization method which has been very well applied to facial expression recognition in the wild and interpersonal relation prediction. Firstly, we have deliberately designed multiple shape templates to fit in with various shapes of facial expressions. Each template describes certain facial activities which consist of a group of facial action units defined by the Facial Action Coding System. Secondly, a template matching strategy is applied to match the query image with an appropriate template. Finally, we employed robust statistical face frontalization and Active Appearance Model to reconstruct facial appearances. We have evaluated the method on both small-scale and large-scale public database and it outperforms state-of-the-art facial expression recognition approaches. We also illuminate the visual effects of frontalization result for comparison.

**Chapter 5** presents a novel dynamic facial expression recognition method based on facial expression-aware face frontalization which normalizes head-pose changes by

reconstructing frontal facial appearances of each non-frontal face. In this method, we firstly collect facial images in a pair of a non-frontal face and its corresponding frontal image, and the pair-wise relation between non-frontal face-shape and frontal counterpart will be learned through a regression model. Considering that such a relation is highly non-linear, a sequentially cascade manner is proposed to iteratively fulfill this task. The obtained cascade regression model will be used as frontal face-shape predictor which transforms the non-frontal face-shape to its frontal view. Finally, the estimated frontal face-shape can be seen as a base template and Active Appearance Model fitting is employed to reconstruct facial appearances. We have evaluated the method on a public database and it outperforms state-of-the-art facial expression recognition approaches. We also illuminate the visual effects of frontalization result for comparison.

**Chapter 6** summaries the thesis with a discussion of the contributions and future work.

# Chapter 2

## Literature Review

As is discussed in Chapter 1, the main challenges of FER can be summarized as identity bias, head pose variations, occlusions, different illumination conditions and various spontaneous expression changes. This chapter is going deep into the problems of FER challenges and approaches. Meanwhile, face frontalization is a newly rising technique that is highly associated with our proposed methods due to its professional design in facial head pose and occlusion normalization. In the remainder of this chapter, we start from the existing facial expression models and benchmarks for affective computing, and then review approaches for facial expression analysis. We then go through the state-of-the-art face frontalization approaches. Finally, we summarize those works and discuss the association with methods proposed in this thesis.

### 2.1 Facial Expression Analysis

#### 2.1.1 Facial Expression Models and Benchmarks

There has been a long history of research on facial expression analysis. Darwin conducted one of the first studies on human emotion analysis. In the book "The expression of the emotions in man and animals" (Darwin & Prodger, 1998) published in 1872, Darwin investigated human facial expressions and body gestures and argued that all human beings presented similar behavioural characteristics when communicating a particular emotion. He also attempted to point out humans and some other animals

(mammals) shared similar expressions so that emotion also had an evolutionary history, which is a support to his theory of evolution.

Darwin stated in the book that only a few specific emotions can be observed and studied through visible facial muscle movements, but he didn't tell which exact emotions were universal. Nearly a hundred years later in 1971, this blank was filled Ekman and his colleagues (Ekman & Friesen, 1971) who presented a cross-cultural study and found that there are six basic emotions: angry, disgust, fear, happy, sad and surprise. Although some researchers argued that human facial expressions of emotions were not culturally universal (Jack *et al.*, 2012), many psychologists still agree that the six emotions are universal to all humans. Another great contribution from Ekman was the Facial Action Code System (FACS) (Ekman, 2002) which systematically described and encoded physical expressions via visible facial Actions Units (AUs). These studies formed the essence of affective computing (Lisetti, 1998) which combined psychologists' expert knowledge with Artificial Intelligence (AI) research to enable computers sensing, understanding and generating facial affective features.

Emotions are human natural and cognitively affective responses to situations or contexts. Emotions can be measured and recognized by many different indicators, such as physiology, body gestures, and individual experience. Among them, facial expression is one of the most important visual signals for human beings to convey affect.

According to a cross-cultural study from Ekman and his colleagues (Ekman, 1992), there are six universal emotions: angry, disgust, fear, happy, sad and surprise. Each of them involves several particular characteristics attached to facial expressions. Most existing approaches for automatic machine recognition of facial expressions aimed to identify six basic emotions.

Apart from six basic emotions, human beings actually act much more broadly distributed affective states. Barrett (Barrett, 2006) proposed that emotions were mutually involved and the boundary of different states were not clear. In (Du *et al.*, 2014), compound emotional categories were defined, where each compound emotional state is a combination of two basic emotions (e.g. happily surprise, sadly disgust, etc.). Currently, there are two datasets for compound facial expression recognition: EmotioNet (Benitez-Quiroz *et al.*, 2017) and RAF-DB (Li *et al.*, 2017).

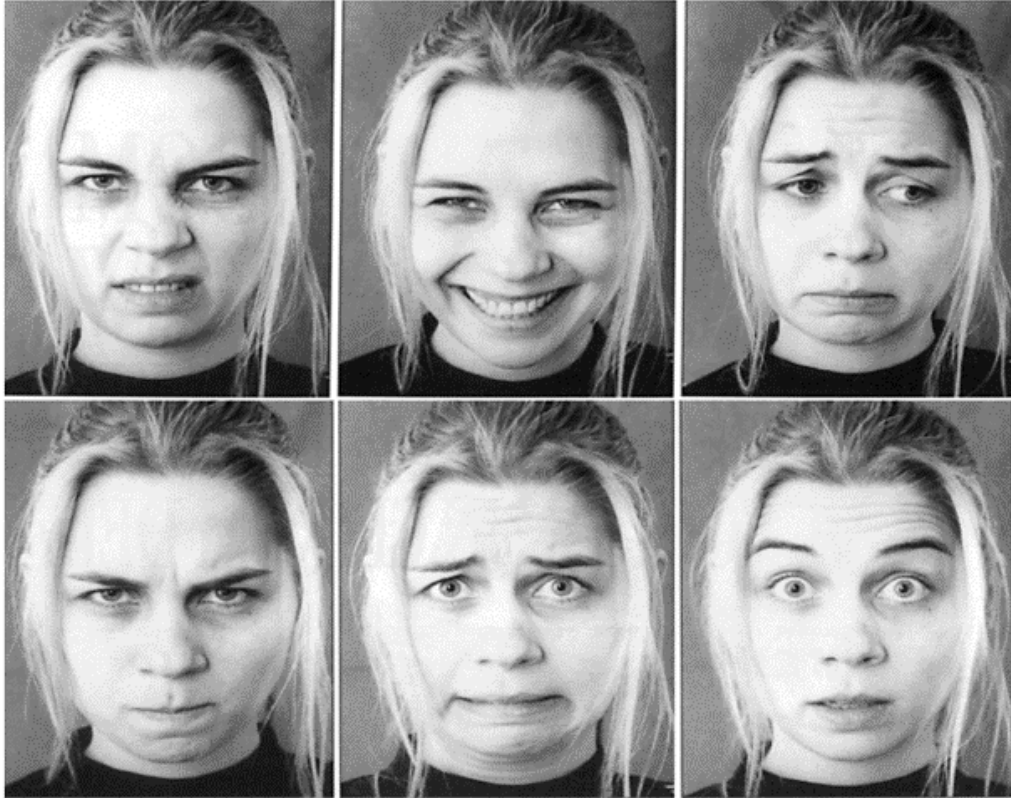


Figure 2.1: Six basic emotions

Another way to quantify affects is valence and arousal dimensional model which is used to describe continuous emotional changes (Russell, 1980). Emotions should be joint events or process, reflected on whatever facial expressions, psychophysiology or brain activities, rather than a specific discrete state. According to Plutchik's research of emotion wheel (Plutchik, 1984), a basic emotion expressed in varying intensity will result in several other states (e.g. high degree of fear is terror while its early stage is apprehension). In the dimensional model, continuous affective changes were divided into two dimensions: binary emotional categories and intensity. Valence considers changes from negative emotion (-1) to positive one (1). The changes of arousal value reflect emotional intensity from calm (-1) to exciting (1). Current benchmarks for valence and arousal detection include Aff-Wild (Zafeiriou *et al.*, 2016) and AffectNet (Mollahosseini *et al.*, 2017).

Recent FER research also addresses 3D/4D facial expression models (Yin *et al.*,

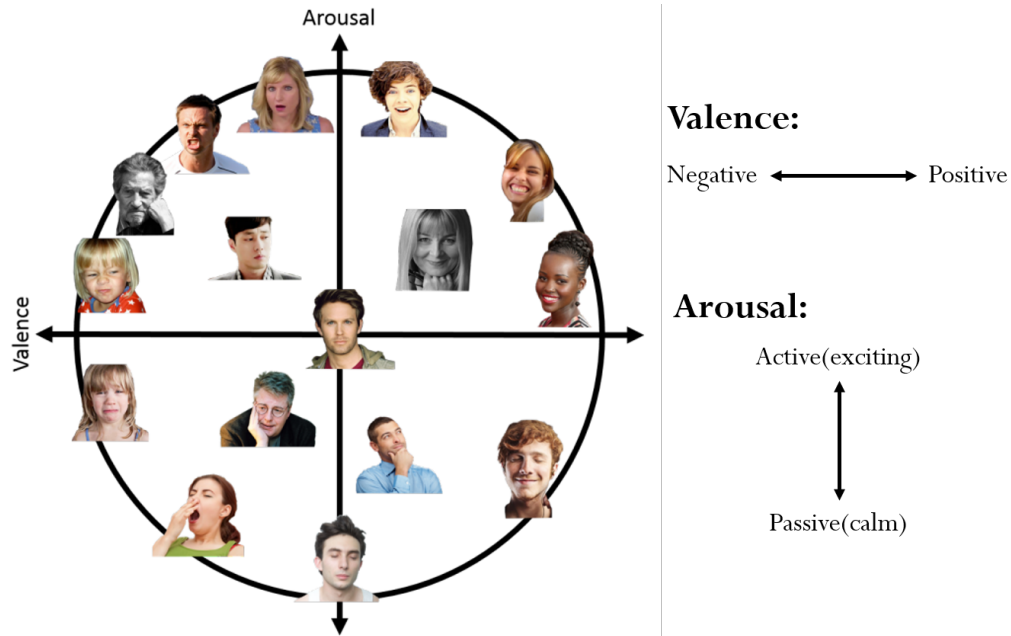


Figure 2.2: Valence and arousal dimensional model

2006, 2008). A 3D facial model includes both RGB pixel values and the corresponding three-dimensional geometric location of each pixel. A 4D facial model refers to a temporal sequence of 3D face models. It has been reported that 3D facial models contain more vivid cues of facial expressions and 3D-based FER methods perform better than 2D-based methods. However, capturing a 2D image is so convenient that every one can take a picture with their cell phones. Whilst, 3D depth sensor is expensive and inconvenient to take along. Therefore, 2D-based facial analysis is still the mainstream for FER research rather than 3D methods. This thesis will mainly focus on 2D-based methodologies and experiments.

Basic and compound emotional categories do not cover all the possible human affects. Valence and arousal are not intuitive for interpretation of emotional states. Although there has already been many benchmarks and studies on facial expression recognition, the research on automatic machine recognition of broadly distributed emotions is still at the infant stage.



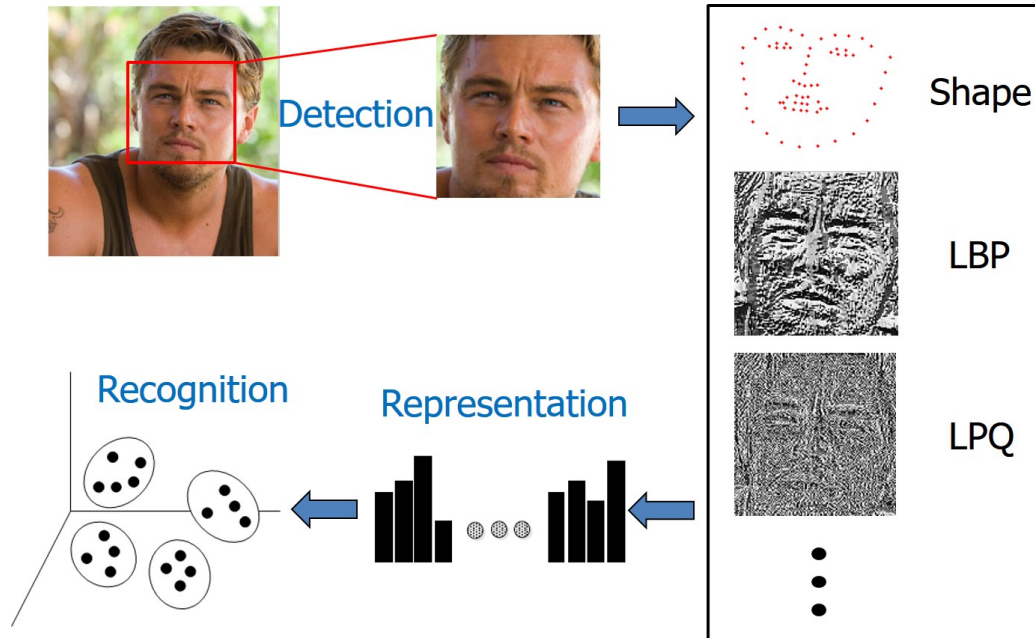


Figure 2.3: Process of traditional FER architecture

### 2.1.2 Facial Expression Analysis

As is mentioned in Section 1.1, six emotions recognition and AU recognition are the mainstream of FER though there have already been many other facial models and benchmarks. Facial Action Coding System (FACS) is the most commonly used approach for human observers to describe the surface of facial behaviours. It specifies a group of atomic facial muscle actions named Action Units (AU) and defines ordinal-level variance in AU intensity. AUs are the smallest visible discrete muscle actions that may occur alone or in combination to perform expressions. With FACS, nearly any anatomically possible facial expressions were coded. Automatic AU recognition is widely regarded as an active topic in facial expression recognition.

Emotion recognition considers six basic categories: happiness, sadness, fear, anger, disgust and surprise. There are some subsets of AUs that usually co-occur on the face to generate meaningful emotions. Although there are obvious connections between the combinations of AUs and universal facial emotions, most existing methods still do not benefit from these connections (Taheri *et al.*, 2014) due to the lack of exploiting domain experts knowledge on AU composition rules and ambiguous semantic nature

of AUs. Furthermore, lack of a commonly accepted evaluation protocol and, typically, lack of sufficient details needed to reproduce the reported individual results make it difficult to compare different systems. This hinders the process of this field.

Traditional methods for facial expression recognition include three main steps: face registration, facial representation and recognition. Empirically, these three steps were executed separately, but sometimes one can achieve a fusion that iteratively learns the facial feature representation and trains a classifier (Tariq *et al.*, 2014).

Face registration can be regarded as a preprocessing step for facial expression recognition. Point detection, especially eye point detection (Shan *et al.*, 2009; Valstar *et al.*, 2011) is usually applied for face registration. Eye point alignment can locate the same facial component in a prototypical frontal face, which is used to register for in-plane head rotation and scale. Further alignment of facial features, such as alignment of the nose and mouth (Jiang *et al.*, 2014), can be performed to enhance stability. But this method will face an identity-bias problem since it emphasizes individual differences in face shapes.

The second step is feature processing which derives an effective facial representation from an image. Facial feature extraction methods vary in terms of different backgrounds. For 2D static facial feature detection, two common approaches are: geometric feature-based methods and appearance-based methods. Geometric feature-based methods present the shape and location of facial components, whilst appearance features describe skin texture changes. For 3D static analysis, there are three kinds of models: distance-based feature, patch-based feature and Morphable models. For dynamic analysis, motion-based feature is widely used in both 2D and 3D images.

Approaches that only use geometric features mostly rely on facial fiducial points (Mattivi & Shao, 2009; Sariyanidi *et al.*, 2015; Zhao & Pietikainen, 2007). Geometric-based recognition can be viewed as facial points movement classification. Spatial relations and motions of these points are often used for expression analysis. In dynamic shape analysis, one can model the temporal characteristics into sequences of temporal segmentations: neutral, onset, apex and offset, which naturally solves the problem of unpredictability in temporal extent. Also, much research focuses on pose-wise facial expression recognition (Mattivi & Shao, 2009; Yang *et al.*, 2011) and have made great contributions. However, these methods totally ignore the information presented in skin

texture changes, and it is reported that they are often outperformed by appearance-based approaches. So in this work, we only focus on appearance-based methods.

Conventionally, the word "texture" is used to describe image regions that exhibit certain spatially stationary stochastic properties. low-rank textures correspond to regions in an image that have rather deterministic regular or periodic structures. Most appearance-based approaches used low-level histogram representations. Popular low-level features include LBP, Local Phase Quantisation (LPQ) (Peng *et al.*, 2005), Histogram of Gradients (HoG) (Sariyanidi *et al.*, 2015), SIFT etc. Among them LBP and LPQ features extraction methods are extended to Three Orthogonal Plane for dynamic application. Low-level appearance features have an identity bias problem in which identity-related cues dominate over expression-related cues (Ahonen *et al.*, 2006; Sariyanidi *et al.*, 2015). The static local appearance descriptors LBP and LPQ show a promising solution to facial biometric analysis, as well as many other visual recognition applications (Mattivi & Shao, 2009; Ren *et al.*, 2015). A histogram representation computed over the whole face encodes only co-occurrence statistics of local patterns without any of their spatial relations. But a holistic facial description sometimes seems not reasonable since it is important to retain information about spatial relations. To overcome this effect, one can consider the block-based representation. In this approach, the face image is divided into several regions (a regular grid, non-overlapping or overlapping nonadjacent blocks) from which feature vectors are extracted and concatenated to form a final representation (Ahonen *et al.*, 2006; Zhao & Pietikainen, 2007).

When using a holistic approach, it is difficult to take the pose invariance into account, because different local face parts change their appearances in different manners in terms of head pose. In order to model local face appearance, patch-based spatial approaches were proposed by dividing a face image into several patches and modelling each patch independently (Ashraf *et al.*, 2008; Prince *et al.*, 2008). For a huge database, most of the features may be irrelevant. Feature selection techniques aim to reduce redundancy and noise whilst preserving principal and discriminative information from the original feature vectors. They provide a powerful tool in various domains, including computer vision, pattern recognition, information retrieval, multimedia analysis (Peng *et al.*, 2005; Yang *et al.*, 2011, 2013; Zhao *et al.*, 2015), etc. Some previous

works have shown an effective feature selection technique can yield a better performance. Alternatively, although sometimes the accuracy may not be enhanced, the computational cost can be alleviated due to the significant reduction of feature dimensionality. Feature selection learns a subspace in which the original feature vectors are filtered or combined into an enhanced feature subset. According to the availability of class labels, feature selection techniques can be classified into two groups: supervised approaches and unsupervised approaches. It is reported that the supervised methods outperform unsupervised methods (Sariyanidi *et al.*, 2015; Yang *et al.*, 2013). Feature selection can be used to undertake several challenges in facial expression recognition, such as illumination variation, registration error and identity bias.

Most methods only capture the characteristic frames at the apex and use a mug shot of each expression. Spatio-temporal representation is used for dynamic facial expression recognition. Spatio-temporal relations capture the facial muscle motions in time intervals. It models facial events happening sequentially and simultaneously, considering a facial expression as a continuous activity over a time slice. Wang *et al.* considered an overlapping time-interval model instead of traditional time-sliced model. (Wang *et al.*, 2013) They used a graphical model to learn pairwise temporal dependency. Facial action is adequate to describe temporal evolution.

Considering the co-occurrences in subregions and their locations, recent research often uses block-based technique for texture analysis (Jiang *et al.*, 2014; Ren *et al.*, 2015; Zhao *et al.*, 2015). This approach was first applied by Ahonen *et al.* in the application to face recognition (Wang *et al.*, 2013). They divided a face image into several non-overlapping regions from which LBP features were extracted and concatenated into a spatially enhanced histogram. Considering that some regions contribute more than others regarding identity variance, the regions were correspondingly weighted based on their contributions. The weighted Chi square distribution was employed for this application. Zhao *et al.* extended this method to overlapping block-based approach in their experiments (Tariq *et al.*, 2014). In the dynamic application to facial expression, the best results were obtained with an overlapping ratio of 70% of original blocks. However, they did not apply any techniques to remove personal-related information. The deriving LBP descriptors were highly effected by identity bias. Shan *et al.* was also inspired by (Wang *et al.*, 2013), they exploited an expression-based region weighted method that only considered the importance of expressions instead of

identities (Valstar *et al.*, 2011). Nevertheless, this method does not take advantage of overlapping blocks. Also, the experiment that was used to test person-independency did not achieve an ideal result.

The final step is recognition that classifies a facial image into one or several parallel facial expression categories. Support Vector Machine (SVM) is one of the most robust and accurate methods among all the existing classification algorithms. In a two-class learning task, SVM optimizes its decision boundary via maximizing the margin, as defined by maximum margin hyperplane, between two classes. This approach offers the best generalization ability that not only guarantees a good partition on training data, but also leaves much room to classify future data with high predictive accuracy. SVM receives good trade-off between model complexity and training error. It shows peculiar advantages in recognition of limited samples, nonlinear and high-dimensional data patterns.

### 2.1.3 FER Applications

FER has been widely used in many application areas. This section will focus on the main applications and the corresponding techniques of FER. Animation: Facial expression is an important factor in animated movies and illuminations. The main task of animation is to transfer expressions from human beings to stylized characters. Facial animation with enriched facial expressions has been well applied to animated movies, Virtual Reality (VR) & Augmented Reality (AR), personalized character design, face reenactment, etc. Some recent advances are introduced in the following.

It is very difficult to synthesize a realistic facial expression image. In (Aneja *et al.*, 2016), an approach is proposed to map expressions from humans to characters by using deep learning method and transfer learning techniques, which results in a group of stylized characters with impressive expressions. In (Thies *et al.*, 2016), a face-to-face real-time facial reenactment is presented where a monocular target video sequence (e.g. from YouTube) is reenacted based on the expressions of a source actor who is recorded live with a commodity webcam. In (Ichim *et al.*, 2015), a cell-phone based software is developed for a dynamic 3D Avatar creation which recover the facial expression dynamics of users in real-time. This method used an adaptive blend-shape

model that integrates feature tracking, optical flow and shape from shading. The software enables human users to create a fully rigged, personalized 3D facial avatar using cell-phone camera. These methods are believed to have a significant impact on many exciting applications of VR / AR, gaming and film industries.

**Assisted diagnosis and nursing:** There are some diseases related to facial expressions, such as facial palsy (Pereira *et al.*, 2011), Parkinsons disease (Gray & Tickle-Degnen, 2010), Schizophrenia (Mandal *et al.*, 1998), autism (Celani *et al.*, 1999) etc. Reports have shown that patients with the above-mentioned diseases will have troubles in perceiving and expressing facial emotions. The diagnose and nursing of this kind of diseases is often tedious and costly.

Currently, FER has not yet applied to this field due to its low recognition accuracy. In this task, a facial system can provide a two-stage function: 1) FER for assisted diagnosis, 2) VR techniques based facial exercise therapy. For the first stage, it is necessary to achieve high-quality facial sensing and recognition. With the development of image-capturing equipment and 3D techniques, many breakthroughs have been made on 3D facial sensing and 3D face reconstruction (Dhall *et al.*, 2011; Jeni *et al.*, 2015; Roth *et al.*, 2015). The progress of 3D techniques will pave the way of the FER applications in automatic diagnosis. For the second stage, facial exercise therapy is the common way to help patients train their facial muscles and recover emotional skills of facial expressions. As is mentioned in animation section, VR techniques are expected to take the place of human labours.

Although there is no existing practical system for automatic counselling, diagnose and therapy of facial expression diseases, FER has already demonstrated its high values in this potential application.

**Multimedia:** With rapid increase of multimedia collections, image & video tagging and retrieval attract significantly increasing attentions in recent years. As emotion is one of the key factors during communication, retrieval by emotion is in high demands. The need of automatic emotion tagging technique is, therefore, increasing.

Although 3D FER has made great progress and proved to be more effective than 2D FER, the main online visual resources are still 2D. Meanwhile, multimedia data is often very varied which means that the type of the data can be very complex. Considering these two issues, emotion retrieval for multimedia application is very challenging. In (Wang *et al.*, 2015), a probabilistic framework is presented for multiple emotion

media tagging by modelling the dependency between facial emotions. This study is validated on multimedia data, such as music emotion database, film data, video data etc. In (Dureha, 2014), an automatic system of generating a music playlist, based on the emotion of facial expressions of users, is investigated. Music is an important source of multimedia and plays an important role in our lives. Human users often select the playlist of music according to their current mood. Thus, FER is the core module of automatic music playlist generation.

**Security:** One of key factors of security is lie detection which is highly related to facial micro expressions. Facial micro expressions are rapid involuntary facial expression that can reveal ones true intentions (Pfister *et al.*, 2011). Facial micro-expressions often last less than 1/3 of a second which is so rapid that only highly trained individuals can distinguish them. Unlike regular macro-expressions which can be easily hidden, micro expressions are nearly impossible to hide the reactions. Currently, facial micro-expression recognition is still at the early stages due to the challenging problems in both facial capturing and expression analysis. With the development of dynamic 3D capturing and analysis, facial micro-expression recognition is expected to be widely applied to security and negotiation.

**home-based services:** Home-based services contain many different applications in tutoring system, robotics, entertainment system etc. It is also a future vision where HCI is involved in our daily lives. Computers will be able to not only recognize linguistic commands, but also discover the intentions conveyed by non-verbal behaviours, especially facial expressions.

Many previously mentioned applications are home-based services, such as personalized Avatars, music for mood, automatic counselling, etc. In (Wu *et al.*, 2008), a new Intelligent Tutorial System (ITS) is developed to meet the emotional need of users. In this approach, FER is integrated in ITS to perform emotional tutorial according to learning emotions and intentions of students.

## 2.2 Facial Frontalization

As mentioned in Chapter 1, the effectiveness of deep learning methods relies on the massive data collections. But manually collecting and labelling tremendous number of faces is time-consuming, error prone and financially challenging. In (Masi *et al.*,

2016), Masi et al. firstly posed a question that Do we really need to collect millions of faces for effective face recognition? In answer to this question, face frontalization was posed and has shown that good results can still be achieved even without millions of manually collected faces. Face frontalization is used to synthesize realistic facial images, which can be seen as a pre-processing of face & facial expression recognition. The existing approaches focus on generating faces in different viewpoints, especially frontal view, from query images. Face frontalization is a comprehensive study, which is often associated with face alignment, face morphing and texture rectification. Based on the objective, face frontalization methods can be divided into two categories: face normalization and face generation.

### 2.2.1 Face Normalization

It has been reported that most popular methods had more than 10% superior performance on frontal-frontal verification over frontal-profile verification (Sengupta *et al.*, 2016). Based on this fact, normalization-based approaches are proposed to normalize facial images taken in the wild by recovering the frontal faces from non-frontal viewpoints. Encoder-decoder (Cole *et al.*, 2017; Kan *et al.*, 2014; Zhu *et al.*, 2013) is a good way for face synthesis, where encoder extracts pose-robust features while decoder recovers the frontal (small pose) faces and there is no need to perform pose estimation. However, the synthesized faces often involve large distortions, since the main objective of encoder-decoder model is not image synthesis but pose-invariant feature extraction.

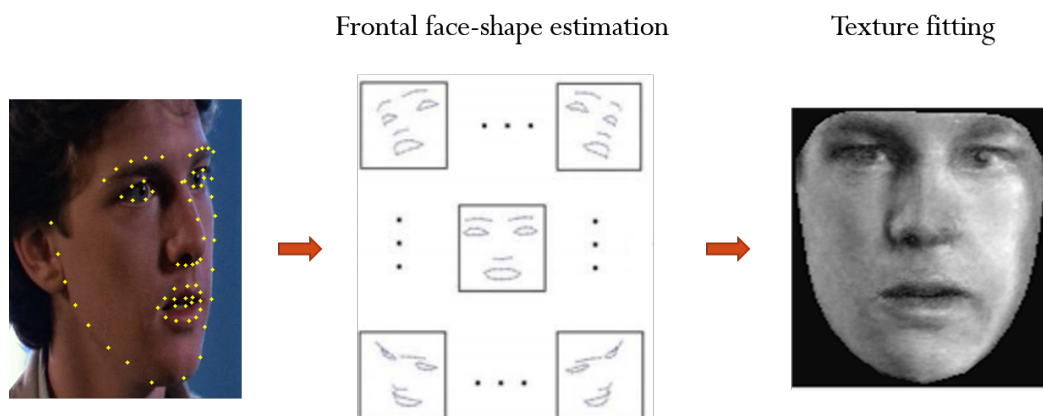


Figure 2.4: Face Frontalization = Frontal shape estimation + texture fitting



Most approaches posed face frontalization as an independent pre-processing or registration step that is separated from feature extraction and recognition steps. A typical pipeline of face normalization includes: face alignment, pose normalization of shape and texture fitting. Current research on face alignment has achieved significant progress (Alabort-i Medina & Zafeiriou, 2015; Asthana *et al.*, 2014; Kazemi & Sullivan, 2014; Ren *et al.*, 2014; Trigeorgis *et al.*, 2016; Tzimiropoulos & Pantic, 2014; Uříčář *et al.*, 2015; Xiong & De la Torre, 2013; Zhu *et al.*, 2015). With the facial shape obtained by face alignment, pose normalization maps the non-frontal shape to its frontal view. Zhu *et al.* (Zhu *et al.*, 2015) used 3D Morphable Model (3DMM) (Blanz & Vetter, 1999, 2003) to fit the 2D landmarks to 3D shape model and then rendered the 3D model to 2D frontal image. Considering that 3DMM fitting is computational expensive, hard frontalization (Hassner *et al.*, 2015; Sagonas *et al.*, 2015) was straightforward proposed by using an unmodified 2D/3D frontal shape template for all the query images. Although hard frontalization totally ignored facial shape features, it still showed good results on frontal face synthesis and face recognition/verification.

The final step is texture fitting that fits facial textures to the reconstructed frontal shape. For a non-frontal face of yaw angle, there will be some invisible regions due to self-occlusion. The main problem of texture fitting is how to fill in the missing pixels. The most commonly used strategy is symmetry-based method (Ding *et al.*, 2012; Hassner *et al.*, 2015; Zhu *et al.*, 2015) which borrows mirrored pixels or gradient to fill in the invisible regions. Another approach is 2D/3D model fitting (Li *et al.*, 2012; Sagonas *et al.*, 2015; Tzimiropoulos & Pantic, 2017) that approximates textures by linear combination of a set of pre-defined holistic facial texture models (orthonormal basis) acquired by Principal Component Analysis (PCA). Basically, model fitting methods are able to synthesize more realistic faces even across large head-pose whilst symmetry-based approaches often generate a non-smooth or weird face, but model fitting methods could hardly achieve real-time performance. The advantage of face normalization is that promising results can still be achieved even without sufficient training sets. However, the recognition accuracy of face normalization approaches has an upper bound due to the finite training instances. The great success of deep face recognition models has shown that face normalization is not always necessary for a face recognition system if there are huge training sets.

### 2.2.2 Face Generation

The performance of deep face recognition models will increase if more labelled faces are collected. But manually collecting and labelling millions of faces is labour intensive. Face generation was naturally posed in which large numbers of labelled faces can be obtained by automatic synthesis rather than manual collection. Face generation approaches always focus on generating faces in novel viewpoints, which can be considered as an extension of face normalization. A typical representation is (Masi *et al.*, 2016) in which 2.4 million faces are synthesized from 495 thousands images by generating individual faces across novel viewpoints and new mouth expressions. The face recognition performance on the synthesized faces shows certain improvement and even comparable to those methods trained on millions of manually collected faces (eg, FaceNet (Schroff *et al.*, 2015), VGG face (Parkhi *et al.*, 2015), DeepFace (Taigman *et al.*, 2014), Face++ (Zhou *et al.*, 2015)).

Many face generation methods depend on 3DMM in order to synthesize faces in novel viewpoints. 3D Morphable Model (3DMM) has been shown to be capable of reconstructing the entire 3D facial surface from a single input image (Banz & Vetter, 1999, 2003). It generates a group of shape models and appearance models in PCA space and then reconstructs 3D surface by linearly combining these models. But it suffers the so-called one-minute-per-frame problem that it is quite computationally expensive. Z-face (Levi & Hassner, 2015) presented a dense 3D registration method which estimated a dense group of 2D facial landmarks by cascade regression and registered a 3D dense model according to the 2D landmarks by using 3DMM model fitting. The output is a dense 3D mesh which can be used to synthesize faces in different viewpoints. Yin *et al.* (Yin *et al.*, 2017) incorporated 3DMM to Generative Adversarial Network (GAN) (Goodfellow *et al.*, 2014), which estimated 3DMM coefficients by a deep network (called deep 3DMM which showed significant acceleration of 3D model fitting) and its output helped GAN to synthesize identity-preserving faces. Although the main objective is to synthesize normalized frontal faces, it still enables 3D reconstruction which can be used to synthesize more faces in novel viewpoints.

Deep learning methods are also very popular for face generation. Zhu *et al.* (Zhu *et al.*, 2014) demonstrated that faces across different identities can be better distinguished by a group of multi-view facial features rather only frontal facial features and

then proposed Multi-view perceptron (MVP) which factorized identity features and view features by using deterministic hidden neurons and random hidden neurons, respectively. MVP can produce a full spectrum of views (multi-view facial features, as well as facial images) from a single 2D image. Yim et al. (Yim *et al.*, 2015) used multi-task deep neural network (DNN) to generate a facial image of any query head-pose from a single input image. Multi-task DNN included a main DNN that generated face of desired head-pose and an auxiliary DNN for the secondary task of identity maintenance. The output of this model is an identity-preserving face of desired head-pose. Tran et al. (Tran *et al.*, 2017) proposed a GAN-based framework that used an encoder-decoder structured generator and a multi-task CNN as discriminator for identity classification task and pose classification task. The output of the generator is the synthesized identity-preserving face of desired pose.

Most face generation methods are proposed for Pose-Invariant Face Recognition (PIFR), whose objective is to derive a single identity-preserving multi-view facial representation. Multi-view introduced new intra-class variance caused by pose variations whilst identity-preserving normalized the intra-class variances due to expressions, illuminations etc. There is theoretical conflict by both introducing and reducing intra-class variances. Considered DCNN is competent in modelling large intra-class variance, there is no need to introduce identity-preserving and face generation should be focusing on generating more faces with large (but reasonable) intra-class variances in order to approximate to in-the-wild conditions. Currently, only the work in (Masi *et al.*, 2016) addressed this problem, but it is still not adequate because the expression variance is only introduced by generating closed mouth expression when the subjects originally open their mouths. More expressions should be generated to satisfied with real-world conditions.

All the face frontalization methods mentioned above only focus on PIFR task. We proposed facial expression-aware face frontalization (Wang *et al.*, 2016, 2017) which was the first work on frontal facial expression reconstruction and achieved the state-of-the-art performance of FER even compared with deep learning methods (Mollahosseini *et al.*, 2016). But both methods belong to face normalization whose performance has an upper bound as is mentioned above. In order to break through this upper bound, we propose to develop a facial expression generation method to synthesize millions of re-

alistic facial expression images and establish an in-the-wild facial expression database containing both manually collected images and synthesized images.

There are currently no methods for large-scale facial expression generation. Theoretically, facial expression generation is much more challenging than facial identity generation which was done in (Masi *et al.*, 2016). For both artificial machine sensing and natural human sensing, facial appearances are always bias to discriminative information of identities but not expressions (Sariyanidi *et al.*, 2015; Zhu *et al.*, 2015). Facial identity variations are one of the most important issues for FER and sometimes they cause even larger intra-class variance than head pose changes. Therefore, generating millions of labelled facial expression images is quite challenging and our study in this thesis is still based on face normalization.

## 2.3 Relation to Our Work

Although a variety of approaches have been proposed for FER in-the-wild challenges, most of them focused on only one specific challenge while ignored others. We summarize the main challenges of FER in-the-wild as follow:

**Identity bias** is the fact that the facial features always contains discriminative cues of facial identity rather than expressions. This problem give birth to the research topic of person-independent FER. Two most effective strategies to this topic are part registration which only extracts facial features on salient regions, and dynamic analysis which extend image-based spatial feature representation to video-based spatio-temporal feature representation. As we show in the experimental analysis of Chapter 3, modelling both salient facial regions and spatio-temporal features improved performance compared to state-of-the-art approaches.

**Head pose and Occlusions** are the main problems of facial analysis which may produce large intra-class variance. Head pose is out-of-plane head rotation which often leads to self-occlusion. The problem of various head pose results in a hot research topic for multi-view or view-invariant FER. Occlusion is caused by the unknown objects locating between subject’s face and camera. This problem is not often addressed in FER research. Face frontalization could effectively solve both problem, but the existing face frontalization approaches only mentioned face recognition task but not FER. We develop a novel face frontalization method which not only solve the two occlusion

problems, but can also be effectively applied to FER task. The experimental results of Chapter 4 and 5 shows outstanding performance for FER in-the-wild.

**Spontaneous expressions** are natural human emotional response to context. Most existing FER research and experiments were conducted on posed facial expression datasets where the images were recorded when subjects were asked to act a certain expression. Spontaneous expressions are much more formless than posed expressions. Spontaneous FER is currently quite challenging. Most existing works solved this problem by simply downloading more facial expression images and training a DCNN model. We present a multi-template model in Chapter 4 to normalize the irregular spontaneous expressions into several uniform shape templates.

These three challenges are currently the main problem of FER in-the wild. Most existing approaches can only solve one or several problems among them. In this thesis, we jointly consider all the challenges and present a face frontalization-based method to normalize the non-linear factors caused by the three problems.

## 2.4 Summary

This chapter reviews some existing methods and techniques in facial analysis. Appearance-based multi-view FER approaches need very large training samples to ensure accuracy. Viewpoint normalization in (Zhao & Pietikainen, 2007) can overcome this problem, but it is based on geometric features so that the accuracy is low. There is no previous work that performs viewpoint normalization on facial appearance features. Little work has been done to address person-independent validation on multi-view FER.

The research on person-independent FER and multi-view FER are done separately. These two issues are the main problems of real-world application to facial expressions. There is no previous work that combines them together. It is necessary propose a unified model to deal with both identity bias and head pose variations.

## Chapter 3

# Dynamic FER Under Controlled Conditions Using Key Features

Both unconstrained and posed FER approaches have to face identity bias challenge. Therefore, we start from studying the common problem of identity bias before going into FER in the wild and presenting a novel method for person-independent FER. This chapter proposes to combine “interest regions” detection and spatio-temporal features for person-independent facial expression analysis under controlled conditions. We will also discuss the advantages and disadvantages of the “interest regions” detection, and then draw forth to FER in the wild.

### 3.1 Introduction

The problem of identity bias is that the whole facial appearances contain more discriminative cues in terms of identity rather than expressions. In many existing works of FER, the recognition rate can be improved by conducting person-specific experiments that trains the model per person/expression categories. The performance of these methods will drop significantly when they are used to recognize expression categories of novel subjects. Accordingly, there is a more practical and important evaluation protocol referring to person-independent validation, in which some of the images in the testing sets contain novel subjects. Currently, there are two effective strategies for person-independent FER: 1) extracting features on expression-enriched regions, 2) extending to dynamic FER.

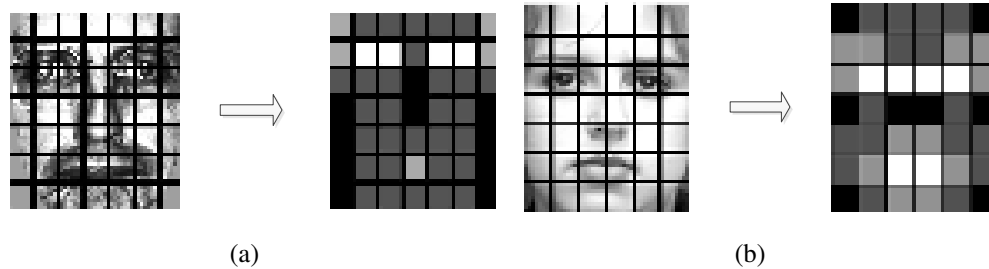


Figure 3.1: Black squares indicate weight 0.0, dark gray 1.0, light gray 2.0, and white 4.0. (a) Subregions were weighted for face recognition. (b) The blocks were weighted for facial expression recognition.

The first strategy aims to detect facial Region of Interest (ROI). The identity-related information often involve three factors: 1) outline of face, 2) different size of facial organs and 3) different displacement of facial organs. The expression-enriched regions often lies into eyebrows, eyes, nose and mouth. A typical example is shown in Fig. 3.1 where two similar region-enhanced approaches are presented for face recognition (Ahonen *et al.*, 2006) and facial expression recognition (Shan *et al.*, 2009), respectively. The regions of higher weighted value contribute more than the regions of lower weighted value. This figure illuminates that facial appearance is a compound containing both identity cues and expression cues, and their corresponding ROIs are different. Therefore, ROI-based approaches are currently an effective solution to identity bias problems.

The second strategy aims to extract temporal features from sequential video clips. Temporal features can better describe facial muscle movements and represent more subtle expression changes. It has been reported that facial temporal cues mostly related to expression changes rather than identity. Therefore, extracting spatio-temporal features is an effective strategy to reduce the influence of identity bias. Currently, the most commonly used spatio-temporal feature extraction method is local binary pattern on three orthogonal planes (LBP-TOP) (Zhao & Pietikainen, 2007).

Although there are some existing works on facial salient region detection and dynamic FER, there is no research that combines them together to reduce the influence of identity bias. Aiming at the these problems of LBP-TOP, we propose an approach to automatically recognize six basic emotions using local patch extraction and LBP-TOP

representation. First, we detect point-based facial landmark by means of Supervised Descent Method (SDM) (Xiong & De la Torre, 2013) which separately detects facial fiducial points in the first frame and tracks them in the following frames. Then, we extract several local patches according to fiducial points. This extraction method has two main advantages: (a) the selected patches lie in expression-enriched regions and the regions facial outlines are totally withdrawn, and (b) independent local patches break the order of the displacement of facial organs. In each patch of sequence, block-based approach is exploited where LBP-TOP features are extracted in each block and connected to represent facial motions. Finally, we perform SVM classifier for emotion classification. The overview of this system is presented in Fig. 3.2.

The main contributions of this work include:

(1) propose a novel method for emotion-enhanced feature extraction. The identity bias problem caused by the variations of facial outlines and different displacement of facial organs can be well solved.

(2) integrate the most effective methods, such as SDM and LBP-TOP, for facial registration and facial representation.

(3) the experimental results shows a good performance on person-independent facial expression recognition. Based on the experimental results, a detailed discussion and analysis is reported in the end of this chapter to address the problem of identity bias.

The rest of this chapter is organized as follows. Section 3.2 reviews several related works of dynamic FER. Theoretical components are described in Section 3.3. The description of data sets and experimental evaluation are given in Section 3.4. Section 3.5 concludes this chapter with a summary and discussion.

## 3.2 Related Works

Facial expression dynamics are essential for facial behavior understanding since they describe basic facial muscle movements. Compared to frame-by-frame spacial representations, spatial-temporal representations are considered to be more efficient and sensitive when representing subtle expressions (Sariyanidi *et al.*, 2015). They enable modelling temporal variance to capture higher level muscular activities and discriminate expressions that look similar in space. Much progress has been made to model



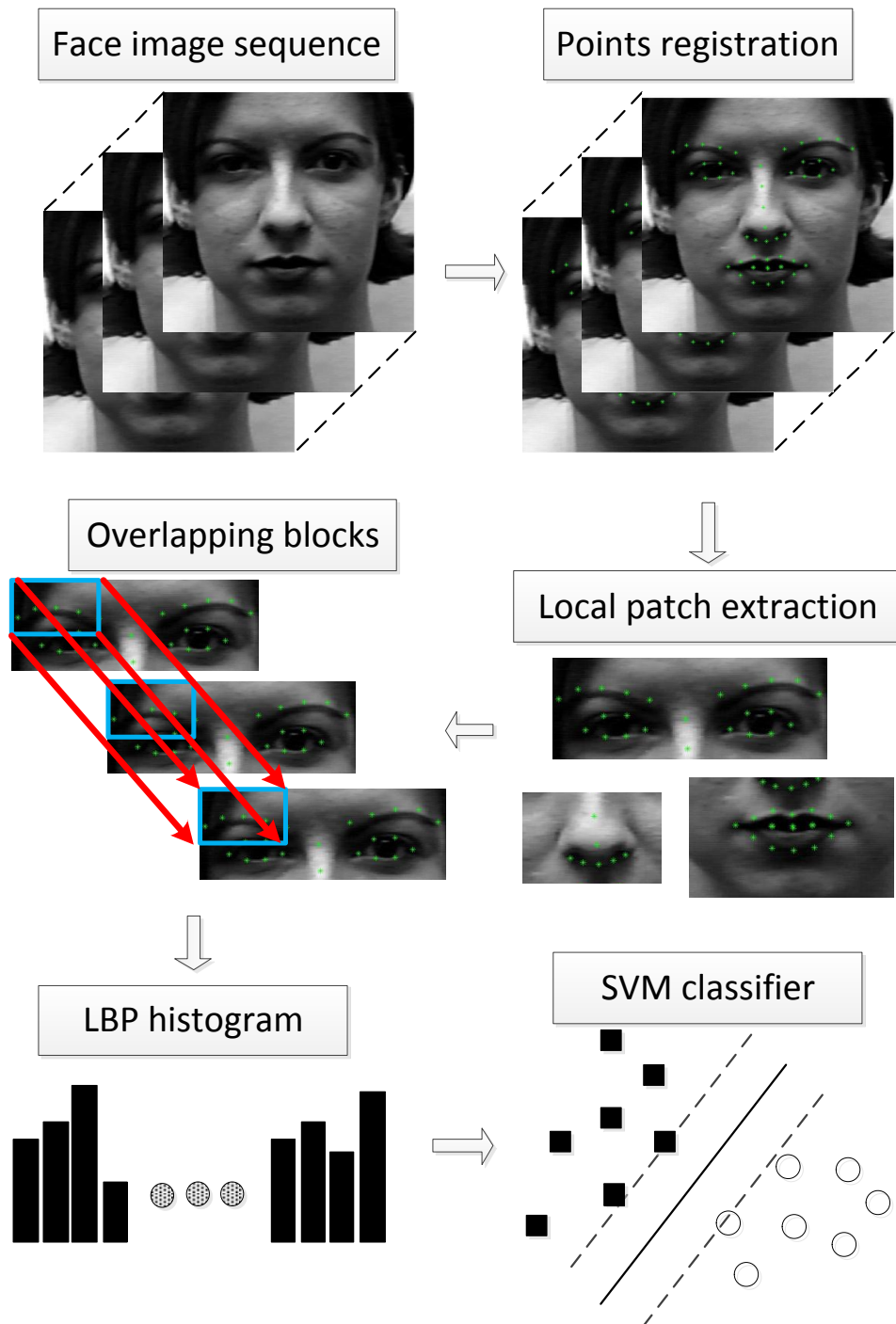


Figure 3.2: The outline of proposed system

spatial-temporal relations from facial image sequences (Jiang *et al.*, 2014; Koelstra *et al.*, 2010; Wang *et al.*, 2013; Zhao & Pietikainen, 2007).

Based on the feature extraction strategy, current FER approaches can be divided into two categories: geometric-based methods and appearance-based methods. Approaches that only use geometric features mostly rely on facial fiducial points (Rudovic *et al.*, 2013; Valstar & Pantic, 2012; Wang *et al.*, 2013). Geometric-based recognition can be viewed as facial points movement classification. Spatial relations and motions of these points are often used for expression analysis. In dynamic shape analysis, one can model the temporal characteristics into sequences of temporal segmentations: neutral, onset, apex and offset, which naturally solves the problem of unpredictability in temporal extent. Also, much research focus on pose-wise facial expression recognition (Rudovic *et al.*, 2013; Tong *et al.*, 2010) and have made great contributions. However, these methods totally ignore the information presented in skin texture changes, and it is reported that they are often outperformed by appearance-based approaches. So in this work, we only focus on appearance-based methods. Most appearance-based approaches applied low-level histogram representations. Popular low-level features include Local binary pattern (LBP), Local Phase Quantisation (LPQ) (Ojansivu & Heikkilä, 2008), Histogram of Gradients (HOG) (Dalal & Triggs, 2005), Scale-Invariant Feature Transform (SIFT) etc. Among them LBP and LPQ features extraction methods are extended to Three Orthogonal Plane for dynamic application.

Local binary pattern (LBP) was originally used to describe static local texture structures (Ojala *et al.*, 2002), and has shown advantages in the tolerance against illumination and its computational simplicity. Zhao *et al.* (Zhao & Pietikainen, 2007) applied LBP feature extraction method on three orthogonal planes (LBP-TOP), which successfully introduced temporal relations to spatial LBP features. They validated LBP-TOP on facial expression recognition as well as dynamic texture analysis, and results showed that LBP-TOP features outperformed spatial LBP features in the application to facial expressions. It has also been reported that LBP-TOP outperforms its spatial counterpart in both emotion and AU recognition (Sariyanidi *et al.*, 2015). However, key issues existing in LBP-TOP include challenging face registration and identity bias. Face registration is challenging because LBP-TOP needs each frame in an image sequence to be in same size, or at least the subregions of each frame to be in same size.

Any in-plane or out-plane rotation will degrade its performance. An effective LBP-TOP operator is highly dependent on face registration. The problem of identity bias generally exists in low-level features (Sariyanidi *et al.*, 2015). It means that the extracted features reserve more information about identity rather than expressions. Shan *et al.* proposed a person-independent approach using LBP (Shan *et al.*, 2009), which highly relies on face registration. And it has not yet been applied to facial expression dynamics.

Considering the co-occurrences in subregions and their locations, recent research on LBP often uses block-based techniques for texture analysis (Jiang *et al.*, 2011; Taheri *et al.*, 2014; Valstar *et al.*, 2011). This approach was first applied by Ahonen *et al.* in the application to face recognition (Ahonen *et al.*, 2006). They divided a face image into several non-overlapping regions from which LBP features were extracted and concatenated into a spatially enhanced histogram. Considering that some certain regions contribute more than others regarding identity variance, the regions were correspondingly weighted based on their contributions. The weighted Chi square distribution was employed for this application. Zhao *et al.* extended this method to overlapping block-based approaches in their experiments (Zhao & Pietikainen, 2007). In the dynamic application to facial expression, the best results were obtained with an overlapping ratio of 70% of original blocks. However, they did not apply any techniques to remove personal-related information. The derived LBP descriptors were highly affected by identity bias. Shan *et al.* was also inspired by (Ahonen *et al.*, 2006), they exploited an expression-based region weighted method that only considered the importance of expressions instead of identities (Shan *et al.*, 2009). Nevertheless, this method does not take advantage of overlapping blocks. The strict consistency of representative regions makes it difficult for face registration. Some works were done manually in preprocessing. Also, the experiment that was used to test person-independency did not achieve an ideal result.

### 3.3 Method

Our proposed method consists of three fundamental components. First, we localize the facial landmarks (49 fiducial points) using SDM. With the position of detected

landmarks, we then extract local patches around fiducial points and apply LBP-TOP in each patch. Finally, SVM classifier is employed for emotion recognition.

### 3.3.1 Facial Landmark Detection

In our study, We use SDM as facial landmark detection method. SDM is a non-parametric shape model for face alignment. Given an image with manually labeled landmarks which are referred to as  $x_*$ , face alignment can be defined as minimizing a Non-linear Least Square (NLS) function:

$$f(x_0 + \Delta x) = ||h(d(x_0 + \Delta x)) - \phi_*||_2^2 \quad (3.1)$$

where  $\phi_* = h(d(x_*))$  represents SIFT values and  $x_0$  is initial configuration of landmarks. During training, SDM learns a sequence of generic descent directions. For testing image, it minimizes NLS objective function using learned decent directions.

In practice, a two-step preprocessing need to be done first: (1) detect a face using openCV face detector, and (2) initialize shape estimation by centering the mean face at normalized square. Then SDM can be executed to detect facial landmarks. Sometimes, one need to perform Principle Component Analysis (PCA) for dimensionality reduction. For tracking task, SDM performs detection in each frame by initializing the landmark estimation from its previous frame. We employed this method in this work because it is extremely fast and accurate. Points registration which detect fiducial points in the first frame of the image sequence and track them in the rest is used. The interested reader can refer to (Xiong & De la Torre, 2013) for more details of SDM.

### 3.3.2 Extract Localized Patches

We firstly perform a preprocessing to normalize in-plane rotation, in which eyes positions are used for alignment in case of in-plane rotation, as is done by most papers (Jiang *et al.*, 2014; Taheri *et al.*, 2014; Valstar *et al.*, 2011; Zhao & Pietikainen, 2007). Considering the importance of expression, we extract three local patches around nose, mouth and eyes regions that are considered to make more contribution to expression variance. Following the registration of facial landmark points, local patches are extracted based on the position of these points. There are some certain points which

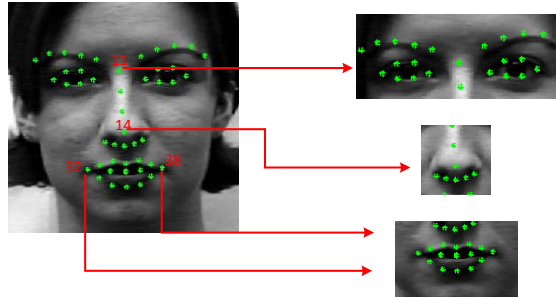


Figure 3.3: Local patch extraction: the first patch centers around point number 11, the second centers around point number 14, point number 32 and 38 are used to design the bounding box of the third patch.

are always in a fixed position regardless of any changes of facial muscles. Depending on these points, a localized texture patch is obtained by establishing a point-centered bounding box with a fixed size in each frame. An image sequence is therefore broken down into its fundamental patch sequences. In this experiment, three patches were extracted from the regions corresponding to eyes, nose and mouth, respectively, as is shown in Fig. 3.3. In each patch, we perform block-based technique to capture local facial motions. We chose  $2 \times 8$  blocks in eyes-related patch,  $2 \times 2$  for nose-related patch and  $5 \times 7$  for mouth-related patch. According to the research in (Zhao & Pietikainen, 2007), the overlapping ratio  $ra'$  in height is computed as followed:

$$ra' = \frac{ra \cdot h/r}{\frac{(ra \cdot h \cdot (r-1))/r + h}{r}} = \frac{ra \cdot r}{ra \cdot (r-1) + r} \quad (3.2)$$

where  $ra$  is original overlapping ratio,  $h$  is the hight of block, and  $r$  is row number of blocks.

### 3.3.3 Temporal Feature Extraction

Given the local patches and each blocks, the LBP-TOP histogram is computed over each block from each patch sequence. LBP is one of the most commonly used method because it is computational simple, powerful to present local structures and easily extended to its temporal model (Huang *et al.*, 2011). The original LBP operator encodes local texture variation with an integer which is derived by comparing each pixel with

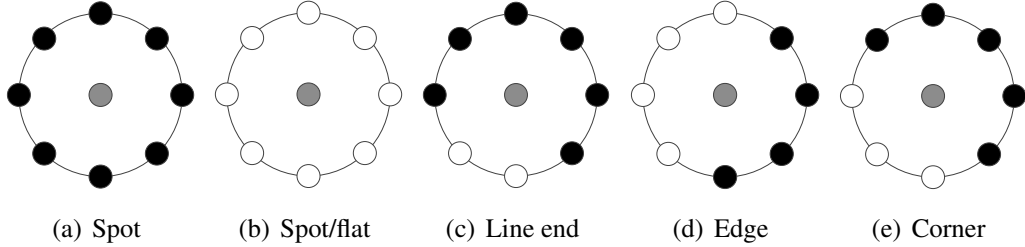


Figure 3.4: Examples of some uniform patterns that show meaningful local structures

its 8 neighbors in a  $3 \times 3$  neighborhood. LBP histograms simply count these integers, and therefore deriving a 256 dimensional histogram representation. Ojala et al. described the extended LBP using circular neighborhoods which allows any radius and numbers of pixels, and showed a more discriminant representation with 59-element subset in which only those meaningful patterns (uniform pattern), such as spot, edge, corner shown in Fig. 3.4, are concerned (Ojala *et al.*, 2002). Typically, we denote a LBP feature representation as  $LBP_{(P,R)}^{u2}$ , where the notation  $u2$  stands for using uniform patterns only and  $(P, R)$  defines a circular neighborhood of  $P$  sampling points on the radius of  $R$ .

LBP-TOP extend spatial LBP features to the spatio-temporal domain. The regular LBP features (using circular neighborhood) are extracted from local spatio-temporal neighborhoods over three orthogonal planes: spatial plane  $XY$ , horizontal spatio-temporal plane  $XT$  and vertical spatio-temporal plane  $YT$ . Uniform patterns still constitute dominant components in representing temporal co-occurrences, so it is appropriate to use uniform patterns in three planes.

The final histogram concatenate all these histograms into a single vector to represent the whole image sequence. Empirically, we will introduce uniform patterns to our LBP-TOP operators, which results in a histogram representation of 177 bins per sequence. The corresponding operator is denoted as  $LBP-TOP_{P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T}^{u2}$ , where the radius in axes  $X, Y$  and  $T$  can be marked as  $R_X, R_Y$  and  $R_T$ , and  $P_{XY}, P_{XT}$  and  $P_{YT}$  represent the number of neighborhood points in  $XY, XT$  and  $YT$ .

### 3.3.4 Emotion Recognition

SVM is considered as one of the most robust and accurate method for data classification. It provides a good balance between model complexity and generalization error. In a two-class learning task, SVM finds a hyperplane with maximal margin for linear separable data. In a geometrical view, this approach ensures both the accurate partition in training data and the best generalization ability on test data. For nonlinear data, SVM allows domain-specific kernel functions that map the input data to a different space of larger dimensions. The most frequently used kernel functions are polynomial and Radial Basis Function (RBF) kernels.

SVM only makes binary decisions. For multi-class problem, a simple but effective technique is one-vs-rest approach. It trains the binary classifiers to distinguish one class against all the others. Each class trains a corresponding binary classifier. The final decision for a test case is made according to the the geometrical distance between test point and decision boundary of all the binary classifiers (the classifier with the largest distance is chosen).

## 3.4 Experiment

The efficiency of the proposed method is tested on Cohn-Kanade Extended (CK+) database. The overall performance and person-independent performance will be test respectively.

### 3.4.1 Dataset

CK+ is a posed facial expression database (Lucey *et al.*, 2010). It consist of 210 adults whose age range from 18 to 50. It contains 123 subjects and 593 frontal image sequence in which the facial expressions of subjects are displayed from neutral to target emotions. From these, only 327 sequences from 118 subjects fit the prototypical definition of emotions and therefore, are annotated with seven emotional categories (six basic + contempt). Within an image sequence,68 landmark points are manually annotated in all images, and action units and their intensity are also provided for the peak frame. In our study, 309 sequences that are explicitly labeled one of the six basic emotions are selected.

### 3.4.2 Evaluation

In this experiment, an overlapping ratio of 70% of these blocks is selected, which shows an outstanding performance in previous work (Zhao & Pietikainen, 2007). Corresponding overlapping ratio in both height and width is about 43%.

We did the experiments using uniform patterns with a neighborhood of  $P = (8, 8, 8)$  and radii of  $R = (3, 3, 3)$ , denoted as  $LBP - TOP_{8,8,8,3,3,3}^{u2}$ . For evaluation, we trained SVM classifier with RBF kernel function using “leave one subject out” cross-validation.

As we can see in Table 3.1, happy, surprise and disgust are recognized with high accuracy (100%, 97.44% and 100%, each), whereas the recognition rates of the others are relatively low. Table 3.2 presents the comparison of different methods. (Wang *et al.*, 2013) is geometric-based approach, whilst the others are appearance-based approaches. It is clear that appearance-based methods outperform geometric-based model. In this stage, the proposed method does not perform so good as the other two appearance-based approaches. Unlike the whole face representation, our method achieves a partial face representation, which will cause the loss of information, as well as the reduction of accuracy. In the real-world applications, we believe it is reasonable to improve person-independent emotion recognition that partially sacrifices the accuracy.

In former experiments, person-independent cases are not emphasizes. Nearly every testing face (subject) could appear in training sets. In order to evaluate person-independent performance, we randomly choose half of the 118 subjects (individuals) for training and the other half for testing, associated emotion sequences are also divided. This scheme guarantees that the subjects used for training will never appear in testing. The testing on a totally new faces is therefore person-independent. This step is repeated ten times to achieve the average results.

As is described in table 3.3, our method achieves the best performance on person-independent experiments. The other two appearance-based methods (Shan *et al.*, 2009) (Zhao & Pietikainen, 2007) show a dramatic drop in person-independent performance. Meanwhile, our proposed method decrease very slightly (only 1%). The work from (Wang *et al.*, 2013) is not mentioned in this comparison. The authors did not release



### 3.4 Experiment

Table 3.1: Confusion matrix (%) of 6-class facial expression recognition for this work

	Angry	Disgust	Fear	Happy	Sadness	Surprise
Angry	<b>75.00</b>	4.17	0	20.83	0	0
Disgust	0	<b>100.00</b>	0	0	0	0
Fear	0	0	<b>57.14</b>	28.57	0	14.29
Happy	0	0	0	<b>100.00</b>	0	0
Sadness	16.67	0	0	16.67	<b>66.67</b>	0
Surprise	0	0	0	2.56	0	<b>97.44</b>

Table 3.2: Comparison of different methods for “leave one subject out” cross-validation

	Angry	Disgust	Fear	Happy	Sadness	Surprise	Total
(Shan <i>et al.</i> , 2009)	86.67	96.61	84.00	100.00	67.86	98.80	92.88
(Wang <i>et al.</i> , 2013)	91.1	94.0	83.3	89.8	76.0	91.3	86.3
(Zhao & Pietikainen, 2007)	86.67	100.00	92.00	100.00	90.91	98.80	94.50
Ours	75.00	100.00	57.14	100.00	66.67	97.44	87.74

Table 3.3: Comparison of different methods for “leave one subject out” cross-validation and person-independent validation

	Cross-validation	Person-independent
(Shan <i>et al.</i> , 2009)	92.88	85.16
(Zhao & Pietikainen, 2007)	94.50	85.89
Ours	87.74	86.56

their source code so that we cannot directly test this algorithm by person-independent evaluation protocol.

### 3.4.3 Discussion

According to the experimental results of cross validation, approaches based on whole face registration outperforms our method which is based on “interest regions” registration. But our method achieves the best results in person-independent validation, which suggests that ROI-based part registration is effective in extracting discriminative features of facial expressions. Compared with whole face registration, ROI-based registration addresses expression-enriched regions while identity-enriched regions, such as facial profile, are withdrawn. This strategy could effectively reduce the influence of identity bias. However, we can also find that part registration may lead to a loss of expression-related information due to the relatively low recognition rate of cross validation.

Facial features are a compound that consists of both identity- and expression-related cues. It would be better if we can normalize identity-related cues before capturing the whole facial features. Face normalization refers to spatial alignment that aligns all the faces into one or several common patterns. For example, as facial profile features contains more identity-discriminative features, the influence of identity bias is expected to be reduced if every face are normalized to share the same profile features by a well-designed rule of spatial alignment. Inspired by this idea, the following chapters will study face frontalization approaches and presents novel methods for real-world FER.

## 3.5 Conclusion

In this chapter, we focus on the task of facial expression dynamics learning and person-independent emotion estimation using localized, block-based LBP-TOP feature representation. The main contribution in this study is that we propose a local patch extraction method that extracts local patches from fiducial point-centered, fixed-sized bounding boxes and preserve information about their spatial relations. These patches

are considered to be more important for human beings to perform expressions. Thus, a person-independent, spatially enhanced feature representation is obtained.

Experiments on CK+ database with person-independent implementation show that our method is effective and favourable compared to other methods. Our approach is more flexible for real application problems, because it not only achieves a good solution to identity bias, but also allows tolerance against small changes in the facial image size and head-pose. The main problem of this approach is that there will be certain information lost if we only register ROIs and it is not suitable for FER under unconstrained conditions. In the following chapters, we will investigate novel spatial face normalization methods and go deeper into FER in the wild.

## **Chapter 4**

# **Facial Expression-Aware Face Frontalization for Static FER in the Wild**

In Chapter 3, we found that the whole facial regions contain redundant information which may lead to identity bias, but there will be certain information lost if we only extract features on the so-called “salient regions”. The identity-related shape information often lies into three factors: 1) outline of face, 2) different size of facial organs and 3) different displacement of facial organs. ROI-based strategy would solve the first problem by discarding regions of facial outlines and partially solve the third problems by dividing a face into independent regions. Obviously, ROI-based approaches cannot effectively remove all the factors caused by identity-bias. In this chapter, we present a frontalization-based method which considers the features on the whole facial regions and identity bias problems can be diminished by several normalization strategies. Furthermore, we start to consider all the challenges of FER in the wild from this chapter. The main goal of face frontalization is to synthesize realistic frontal face from non-frontal facial images. The existing methods are either computational expensive or fail to perform frontalization on novel subjects, none of them has been applied for FER. In this work, we present a novel facial expression-aware face frontalization method which jointly considers all the challenges of FER in the wild, as was discussed in Chapter 1.

## 4.1 Introduction

It is commonly accepted that universal facial expressions of emotions include six basic categories: angry, disgust, fear, happy, sad and surprise. Following the research work by Ekman *et al.* (Ekman & Friesen, 2011), facial expressions of emotions can be semantically described as a group of Action Units (AUs) defined in Facial Action coding System (FACS). FACS is the most comprehensive and anatomic system in which AUs are facial muscle action descriptor that encode the smallest visible facial muscle movements. Some excellent results have been achieved for AU detection (Eleftheriadis *et al.*, 2016; Koelstra *et al.*, 2010; Rudovic *et al.*, 2015; Walecki *et al.*, 2015). Although there are obvious connections between AUs and six facial emotions, there are only a few approaches proposed for AU-based facial emotion analysis (Liu *et al.*, 2013; Taheri *et al.*, 2014). This is because AUs are very small facial activities that are quite sensitive to the unconstrained conditions such as head-pose changes and occlusions. Therefore, AU-based approaches for facial emotion analysis has not yet been fully developed.

Recently, face frontalization has attracted wide attentions due to its effectiveness in facial analysis. It is commonly accepted that more robust features can be captured from frontal face rather than profile face. Thus, the main objective of face frontalization is to recover the frontal faces from non-frontal viewpoints. Meanwhile, there are also several extended face frontalization approaches that generate facial images in not only frontal view, but also other views in order to capture more facial features. In general, face frontalization includes two key components: frontal facial shape estimation and frontal facial texture fitting.

Frontal shape estimation starts from facial landmark detection. Recently, many breakthroughs have been made on automatic facial landmark detection (Ren *et al.*, 2014; Xiong & De la Torre, 2013; Zhu & Ramanan, 2012). The objective of frontal shape estimation is to align the non-frontal facial landmarks to their frontal positions. Then, frontal texture-fitting recovers facial appearances by texture warping and rectification. When the non-frontal facial textures are overlaid on a frontal face mesh, there will be one half face (left or right side of face) with rich pixels and some regions the other half without visible pixels. The task of texture fitting is to compensate these regions and rectify the whole facial textures.

There are currently two main problems in the existing face frontalization approaches: 1) it is difficult to achieve real-time performance 2) most methods could only be used for face recognition but not FER.

Face frontalization cannot achieve real-time performance because most approaches are unsupervised so that it takes a long time to do the optimization. Meanwhile, a majority of approaches achieve only person-specific face frontalization in which novel subjects cannot be normalized to their frontal view. Therefore, they are not suitable for FER because a good FER system should work well on any unseen faces. In the existing generic (person-independent) face frontalization methods, a rough solution to frontal shape estimation is presented in which an unmodified shape template (in frontal view) is used as reference for all the query images and then texture-fitting is performed based on this single template (Hassner *et al.*, 2015; Sagonas *et al.*, 2015). This strategy is called hard frontalization in which the reconstructed frontal faces will share a common 2D/3D face shape. The template is often made in a neutral shape as a compromise. However, the facial expression cues are ignored when reconstructing frontal face. It is a challenging task to remain or recover facial expressions during the process of face frontalization. As far as we know, there is no attempt so far that performs face frontalization with full considerations of facial expressions. To this end, we present a novel approach that develops supervised approach for real-time face frontalization and combine the expert knowledge of AUs to achieve a facial expression-aware approach for FER in the wild.

In this chapter, we propose an approach of Facial Expression-Aware face Frontalization (FEAF). This approach includes three main steps: multi-template design, template matching and texture fitting.

Firstly, multiple emotional shape templates are designed in order to fit in with more facial expression changes. Considering that the facial regions of eyes, eyebrows and mouth contain the most enriched facial expression cues (Shan *et al.*, 2009; Xue *et al.*, 2013), all the possible combinations of AUs on these salient regions will be collected to form the emotional templates. Obviously, not all the templates contain principal emotional cues. We then propose a templates measurement strategy based on information theory to measure their importance. The principal templates will be selected by this strategy. Their shapes are obtained by computing the mean landmarks of all the instances within each template category. In training data, we manually label each

image with one of the emotional templates according to the visible AU combinations of facial images.

The second step is template matching which automatically matches each query image with an appropriate template. With all the images from training data and their assigned templates as unique class label, template matching can be viewed as a classification problem which can be effectively solved by Support Vector Machine (SVM). By these two steps, a multi-template model is presented. There are two main advantages of this model: 1) the query facial images of arbitrary out-of-plane head rotations can be recovered to the frontal view; 2) the three non-linear factors of identity bias (different outlines, size and displacement of face) with regards to shape variations can be normalized to several unified shape templates. Theoretically, multi-template model provide a better solution of identity bias than ROI-based approaches since it considers all the three factors and registers the whole face without any information lost.

Finally, we rectify the textures into the selected shape templates by Active Appearance Model (AAM) instantiation (Sagonas *et al.*, 2015). AAM can effectively reconstruct facial textures by linearly combining a group of eigen faces whose coefficients are learned and optimized through an unsupervised way. The experimental results on a small-scale facial expression dataset demonstrate the effectiveness of FEAF. In order to test the performance on large-scale dataset, we present a novel FEAF-based deep learning model for interpersonal relation prediction (Zhang *et al.*, 2018), which achieves the state-of-the-art performance. The contributions of this work are summarized as follow:

- 1) We propose a novel facial expression-aware face frontalization method which reconstructs frontal faces with detailed facial expression cues from unconstrained facial images. This is the first work of its kind that jointly considers all the non-linear factors in FER task.

- 2) The proposed method is robust to head-pose variations and occlusions, and provides a better solution of identity bias than ROI-based approaches. It is proved to be an effective methods for the application of facial expression recognition in the wild.

- 3) FEAF provides a flexible platform for multiple facial template design for facial expression recovery during face frontalization. The multi-template design strategy is also fitted to other applications of FER. We demonstrate its flexibility by developing an

FEAF-based deep learning model and test it on interpersonal relation estimation task. It achieves the state-of-the-art performance.

## 4.2 Related Works

Most existing research on FER, as well as public facial expression database, processes faces in controlled environment where the subjects exhibit posed expressions and their facial images are captured in frontal or near-frontal view without occlusions (Zeng *et al.*, 2009). With the development of human-machine interaction system, there is a significantly increasing demand for FER in the wild.

The variations of out-of-plane head rotation is one of the key challenges for FER in the wild. There are only a few works on head-pose-invariant FER. Head-pose-invariant FER approaches focus on how to tackle different views. Moore *et al.* (Moore & Bowden, 2011) propose a multi-view approach which classifies all the possible head-poses into multiple discrete categories of yaw angles. For all the training images, pose estimation is performed and then head-pose angle is assigned to its closest category. Then a view-specific facial expression classifier is trained in each yaw angle category. This method requires a large amount of training data in each head pose in order to train the classifier. It will fail to perform recognition on novel head pose categories. Another multi-view FER method (Hesse *et al.*, 2012) suffers the same problems. In (Eleftheriadis *et al.*, 2015) and (Rudovic *et al.*, 2013), pair-wise view normalization are described. Different Gaussian Process Regression model are used to model the relations of landmarks between a non-frontal face and its pair-wise frontal face. They can deal with novel head-pose categories, but still need a very large amount of training samples.

It is obvious that head-pose-invariant FER highly relies on the performance of pose estimation method. Meanwhile, training data is required to be in high-quality and large-amount. The satisfied database is often not readily available. More importantly, head-pose variation is just one factor of unconstrained facial images. The problem of occlusion is not mentioned in these methods. Thus, head-pose-invariant FER approaches only partially solve the problem of FER in the wild.

The state-of-the-art performance on FER in the wild has been obtained by using DCNN (Kim *et al.*, 2015; Mollahosseini *et al.*, 2016; Tang, 2013). However, deep FER



models have not yet attracted wide attentions and their results were not such fabulous as other applications of DCNN (e.g. face recognition (Parkhi *et al.*, 2015; Taigman *et al.*, 2014), image recognition (Simonyan & Zisserman, 2014; Szegedy *et al.*, 2015) etc.). This is due to the very finite public resources of facial expression images, especially those captured in the wild. Meanwhile, it is also quite challenging to manually collect and label millions of facial expression images due to the ambiguously semantic nature of facial expressions. Human beings often exhibit expressions of mixed emotions so that it is nearly impossible to label a facial expression image with a clear and independent emotional category. Therefore, small-sample learning methods are still the mainstream of current FER research.

Face frontalization aims to reconstruct the frontal facing view from non-frontal viewpoints. It is a comprehensive research method, which is often associated with face alignment, face morphing and texture rectification. Most approaches were designed for face recognition task since it has been reported that most popular facial processing methods had more than 10% superior performance on frontal-frontal verification over frontal-profile verification. Face frontalization can be seen as an independent preprocessing or registration strategy. It is used to normalize unconstrained facial samples of various out-of-plane head rotations and occlusions into a controlled case of clean frontal face. Based on the generalization capability, face frontalization methods can be divided into two categories: person-specific approaches and generic approaches.

Frontal facial shape estimation is the fundamental step, but also very challenging. In (Jeni *et al.*, 2015) and (Roth *et al.*, 2015), two approaches of person-specific 3D model reconstruction are performed, in which several images captured from one person in different poses and expressions are used to reconstruct his/her 3D model. These methods can implement frontalization, but they are not practical since these methods a) are computational expensive to build 3D model, b) require a massive training data to learn shape models, and c) will fail in reconstructing 3D model of novel subjects.

AU	Name	AU	Name	Emotion	AUs
1	Inner Brow Raiser	17	Chin Raiser	Happy	6+12+25
2	Outer Brow Raiser	20	Lip Stretcher	Sad	1+4+15
4	Brow Lowerer	22	Lip Funneler		
5	Upper Lip Raiser	23	Lip Tightener	Surprise	1+2+5+26
6	Cheek Raiser	24	Lip Pressor	Fear	1+2+4+5+20+26
7	Lid Tightener	25	Lips Part		
9	Nose Wrinkler	26	Jaw Drop		
12	Lip Corner Puller	27	Mouth Stretch	Angry	4+5+7+23
15	Lip Corner Depressor	28	Lip Suck	Disgust	9+15+16
16	Lower Lip Depressor	43	Eyes Closed		

Figure 4.1: Relations between universal emotions and AUs

## 4.3 FEAF Model

### 4.3.1 Initial Design of Multi-template Model

Since FACS AUs have been effectively used in facial behaviour analysis, we design the multi-template model according to AU combinations. As is previously mentioned, there is a close relations between AUs and 6 universal facial emotions. As is decribed in Fig. 4.1, the left table shows the AU numbers and their description. The right table shows that each basic emotion can be described as a specific combination of AUs, which suggests a typical relation between universal emotions and AUs. However, spontaneous facial expressions of one specific emotion varies due to individual differences and expression intensities. Therefore, we design multiple templates by re-grouping AUs in salient regions in order to meet the requirements of modelling various expressions.

It is commonly accepted that the facial regions of eyebrows, eyes and mouth are regions of interests(ROI) which convey the most enriched cues of facial expressions. Thus, the emotional templates are designed according to the facial activities in ROI. But in practice, the corresponding AUs or AU combinations are sometimes not significantly visible or zre not associated with emotions. So we set an explicit exclude criteria: 1) The less obvious features are excluded (eg, AU7 and AU16) due to the ambiguous semantic nature of AUs. 2) Considered our template are shape models, the













ROI	AU	Shape	Examples
Eyebrow	1+2		
	1+4		
	4		
	Neutral		
Eyes	5		
	Neutral		

Figure 4.2: FACS AUs on eyebrow and eye regions















ROI	AU	Shape	Examples
Mouth	12+15		
	15		
	20		
	23		
	25		
	27		
	Neutral		

Figure 4.3: FACS AUs on mouth regions

AUs which are only reflected by texture features are excluded (eg, AU17 and AU24).  
3) Those that are not related to facial emotions are excluded (eg, AU43).

Based on the above mentioned rule, we first design the shape model in eyebrows, eyes and mouth regions, respectively. AU1, 2 and 4 encode the activities of eyebrows and their combinations generate four emotional eyebrow behaviours: 1+4, 1+2, 1+2+4 and 4 corresponding to sad, surprise, fear and angry, respectively. The expression of 2+4 is not considered since it happens occasionally. By observing that the combination 1+4 and 1+2+4 are visibly similar and difficult to distinguish them only according to shape features, we only maintain 1+4 and remove 1+2+4. Therefore, the results of eyebrow templates include 3 combination, as is shown in Fig. 4.2.

The AUs related to eye movements are AU5, 7 and 43. AU43 is not a discriminative feature for facial emotions, so it is excluded straightforwardly. AU7 is useful but its shape features are not so obvious as they are nearly the same as neutral state of eyes. So the eye behaviours include only two states: eyes widely open (AU5) and neutral, as is illustrated in Fig. 4.2.

The mouth activities are much more complicated. AU 12, 15, 16, 20, 22, 23, 24, 25, 26, 27 and 28 describe the mouth actions. Among them, AU 16, 24 and 28 are less significant shape features. The differences between 26 and 27, 25 and 22 cannot be reflected by landmark features, so AU22 and 26 are excluded. The remaining AUs are illustrated in Fig. 4.3.

Combining all these components will lead to 42 templates, but most of them are meaningless or non-existent (eg, 4+12+25 and 1+2+15). We check all these templates on a public facial expression database: Static Facial Expression in the Wild (SFEW), and manually assign each facial image with a template. If a template cannot match with any images in the database, we will think of it as an unreasonable template and remove it. After this process, there are 16 templates left as is shown in Fig. 4.5 and Fig. 4.6.

The 16 initial templates could cover nearly all the possible shapes facial expressions of emotions. But how to distinguish them is a challenging problem. In order to make it easier to classify these templates, we need to reduce the number of templates by merging some similar templates. Details of template design and template matching are discussed in Section 4.3.2.

Frequency and J-measure of eyebrow AUs on CK+ database

	angry	disgust	fear	happy	sad	surprise	J
AU1	0	0	0	0	2	0	0.0155
AU4	40	36	3	0	2	0	0.2405
AU1+2	0	0	4	0	3	80	0.3018
AU1+4	0	0	12	0	17	0	0.1661
AU1+2+4	0	0	6	0	4	1	0.0515
Neutral	5	23	0	69	0	2	0.2349

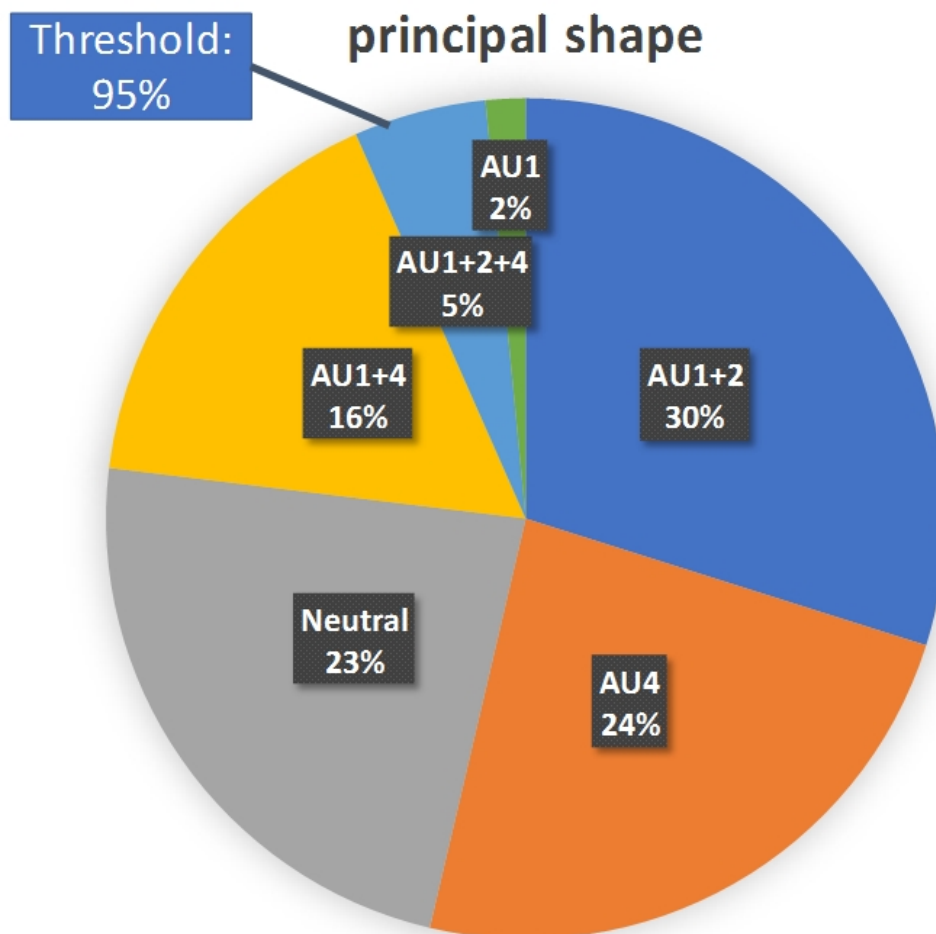


Figure 4.4: Principal shape template of eyebrow behaviours

### 4.3 FEAF Model












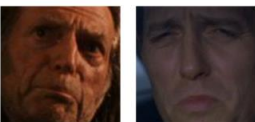




Template	AU	Shape	Examples	Related emotions
1	12+25			Ha:92.5%, Ne:1.25%, Sa:1.25%, Su:5%
2	1+2+5+25			An:8.54%, Di:10.98%, Fe:26.83%, Ha:6.1%, Ne:7.32%, Sa:8.54%, Su:31.73%
3	1+2+5+27			An:13.33%, Fe:16.67%, Ha:16.67%, Su:63.33%
4	1+4+5+20			An:7.14%, Fe:64.29%, Sa:28.57%
5	1+4+5+27			An:33.33%, Fe:33.33%, Sa:33.33%
6	1+4+15			Di:20%, Ne:20%, Sa:60%
7	1+4+25			An:10.34%, Di:17.24%, Fe:37.93%, Sa:27.59%, Su:6.9%
8	1+4			Di:25.81%, Ha:9.68%, Ne:16.13%, Sa:48.39%

Figure 4.5: Initial shape template (part 1)

### 4.3 FEAF Model






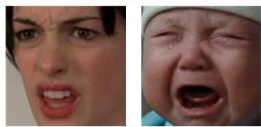

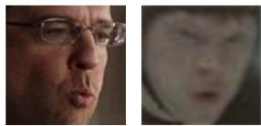

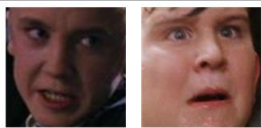





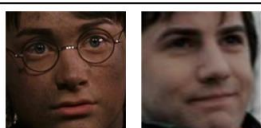
Template	AU	Shape	Examples	Related emotions
9	4			An:18.89%, Di:36.67%, Fe:10%, Ha:5.56%, Ne:6.67%, Sa:22.22%
10	4+25			An:48.98%, Di:12.24%, Fe:14.29%, Sa:10.2%, Su:14.29%
11	4+27			An:44.83%, Di:13.79%, Fe:20.69%, Su:20.69%
12	4+23			An:14.29%, Di:14.29%, Su:71.43%
13	4+5+25			An:35.29%, Fe:29.41%, Su:35.29%
14	4+20			An:14.29%, Di:7.14%, Fe:7.14%, Ha:14.29%, Sa:57.14%
15	4+15			An:53.85%, Di:15.38%, Fe:7.69%, Ha:7.69%, Sa:15.38%
16	Neutral			An:12.56%, Di:7.25%, Fe:10.63%, Ha:10.63%, Ne:39.13%, Sa:12.08%, Su:7.73%

Figure 4.6: Initial shape template (part 2)



### 4.3.2 Template Design and Template Matching

The goal of template matching is to match the facial shape of a query image with its most appropriate template. It, thus, starts from facial landmark detection. We employed Supervised Decent Method (SDM) (Xiong & De la Torre, 2013) for landmark detection due to its effectiveness and small computational cost. SDM can well detect 49 landmark points located in eyebrows, eyes, nose and mouth. As is discussed before, our template model only focuses on eyebrows, eyes and mouth. So, the 9 points located in the nose region are deleted, which results in 40 points for template matching (Wang *et al.*, 2016). As is shown in Fig. 4.7, the green points on nose region are excluded from the shape model. Then, the shape is regularized to a uniformed size by using procrustes analysis. With the obtained shape features, we classify an image into one of 16 templates. This task can be seen as a classification problem which can be solved by machine learning techniques.

The classification results on 16 templates can be seen in Table 4.1. It is obvious that the accuracy of template matching on initial templates is low (The overall accuracy is 47%). This is because there are some similar templates, which may result in mismatching. Apparently, the accuracy will improve if we reduce the number of templates by merging the similar templates. However, too few templates will oppositely cause large distortion of expression reconstruction. The goal is to minimize the error rate of template matching while retaining as many discriminative templates as possible. So we design the multi-template model according to two issues: discrimination of templates and error rate.

For the first problem, discrimination describes how much information a template may contain (how useful for learning). Although each template depicts a specific facial shape movement and is substantial for expression modelling, their contributions to facial emotion recognition task are quite different. The most informative templates will be selected to build our multi-template model.

J-measure (Smyth & Goodman, 1992) is an effective criterion which is introduced in information theory and usually uses rule-based learning approaches. Given two distributions  $Y$  and  $T$ ,  $y$  and  $t$  are their specific values, respectively. For our purpose,  $Y$  is the distribution of labels of six facial emotions and  $T$  is the template distribution

Table 4.1: Measurement of initial template matching on SFEW database

	t1	t2	t3	t4	t5	t6	t7	t8
accuracy(%)	61.25	45.12	53.33	7.14	0	40.00	34.48	3.23
j-measure	1.50	0.24	0.95	1.12	0.82	1.03	0.53	0.76
J-measure	0.17	0.03	0.04	0.02	0.004	0.007	0.02	0.03
T-measure	10.49	1.26	2.16	0.16	0	0.30	0.76	0.11
	t9	t10	t11	t12	t13	t14	t15	t16
accuracy(%)	57.78	24.49	51.27	0	17.65	0	23.08	58.94
j-measure	0.39	0.53	0.65	1.22	0.85	0.68	0.60	0.19
J-measure	0.05	0.04	0.03	0.01	0.02	0.01	0.01	0.06
T-measure	2.91	0.91	1.39	0	0.36	0	0.26	3.34

with  $\mathbf{T} = t$  being a particular input event that one of 16 templates is selected. J-measure, as well as one of its component j-measure, is expressed as follow:

$$j(\mathbf{Y}; \mathbf{T} = t) = \sum_y p(y|t) \log\left(\frac{p(y|t)}{p(y)}\right) \quad (4.1)$$

$$J(\mathbf{Y}; \mathbf{T} = t) = p(t) j(\mathbf{Y}; \mathbf{T} = t) \quad (4.2)$$

where  $p(y)$  and  $p(y|t)$  are priori and posteriori probabilities of  $\mathbf{Y}$ , respectively. And  $p(t)$  is the probability that event  $\mathbf{T} = t$  occurs. The j-measure can be viewed as the average mutual information of  $y$  and  $t$ , which defines the information theoretic similarity between emotion categories and template. The template  $\mathbf{T} = t$  with higher value of j-measure is better because it is more biased to one emotion and, thus, more discriminative. In J-measure,  $p(t)$  denotes how many training instances of a template covers over the whole database. For example, t6 is biased to Sad but only a few images are assigned to t6. Although t6 is discriminative, its contribution to average recognition rate of facial expressions is still finite. So, J-measure is the average expectation of discrimination with regards to each template.

Another problem of template matching is classification error. A well-designed template should be not only discriminative with regards to emotions but also easily distinguished with other templates. The classification accuracy of template matching can be denoted by  $A(t)$ .

We propose a new measurement for template matching, called T-measure:

$$T(\mathbf{Y}; \mathbf{T} = t) = A(t) J(\mathbf{Y}; \mathbf{T} = t) \quad (4.3)$$

The T-measure computed on SFEW training data is shown in Table 4.1. A well-designed template will have a higher value of T-measure. This kind of template should be selected by a threshold strategy as follow: 1) sort templates according to T-measure values from large to small, 2) calculate the summation of all the  $n$  T-measure values  $S(n) = \sum_{i=1}^n T(\mathbf{Y}; \mathbf{T} = t_i)$ , 3) calculate the summation  $S(j) = \sum_{i=1}^j T(\mathbf{Y}; \mathbf{T} = t_i)$  where  $j = 1, 2, \dots, n$ , 4) calculate the percentage  $S(j)/S(n)$  starting from  $j = 1$  and moving on calculating the percentage with  $j \leftarrow j + 1$  until the percentage is larger than the threshold.

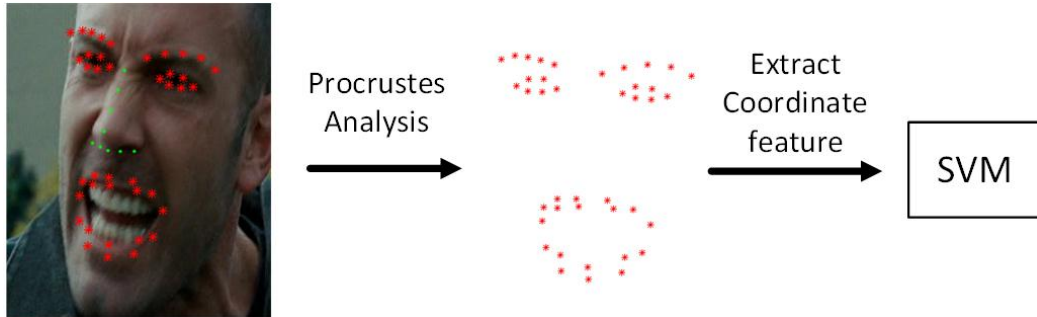


Figure 4.7: Template Matching

If we set the threshold by 85%, 6 template t1, t2, t3, t9, t11 and t16 will be selected. If we set the threshold by 95%, 8 templates t1, t2, t3, t7, t9, t10, t11 and t16 will be selected. In this set-up, we use 95% as the threshold, as illuminated in Fig. 4.4. Therefore, the corresponding 8 templates are selected as the best-designed fiducial templates. According to definition of T-measure, it is obvious that the selected templates have already covered most of the training instances over the whole database. So there is no need to merge the rest of the templates to these fiducial ones because their contributions to template design and matching is very small.

With all the 8 fiducial templates and their corresponding images, we manually select the frontal facial images without occlusions from training data. The reference templates are obtained by computing the mean shape of each template category. The results are shown in Fig. 4.8.

As is previous mentioned, each facial image will be matched to one of 8 fiducial templates and machine learning techniques can employed for this task. The above mentioned classification accuracy of initial template matching is obtained by SVM.

**Support Vector Machine:** SVM is one of the most effective and accurate methods for data classification (Heisele *et al.*, 2001). In a two-class learning task, SVM assumes that the best classification results are obtained by maximizing the margin of hyperplane between two classes. It can be expressed by:

$$\begin{aligned} & \underset{\mathbf{w}, b}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{w}\| \\ & \text{s.t. } y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, \dots, n \end{aligned} \quad (4.4)$$

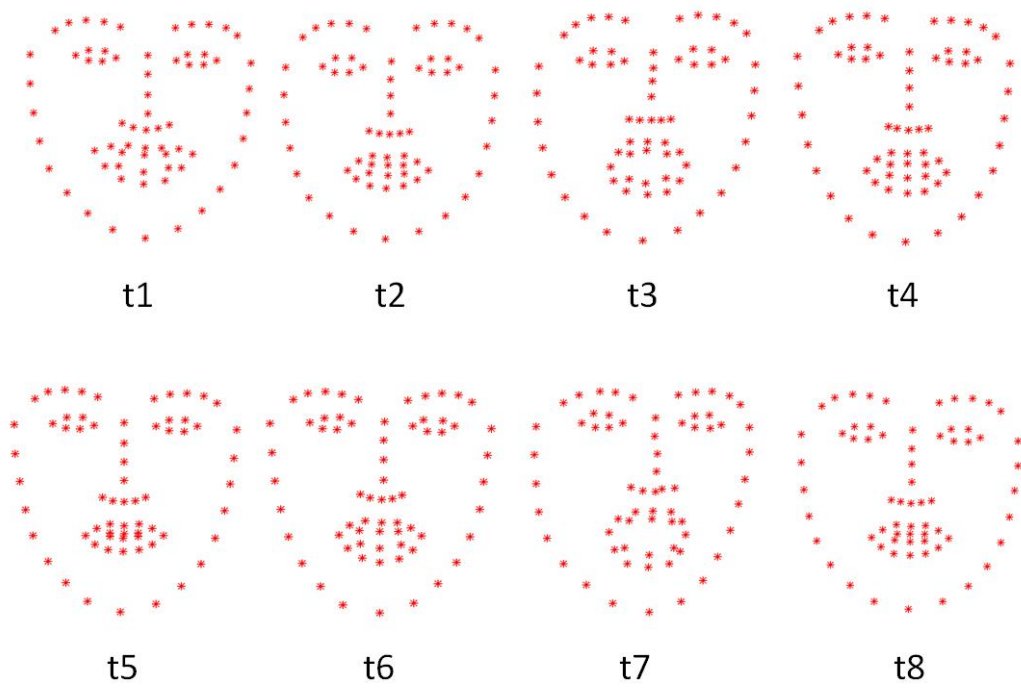


Figure 4.8: Eight templates of face shape

where  $(\mathbf{x}_i, y_i)$  is a training instance,  $\mathbf{w}$  and  $b$  denote weight vector and bias in the hyperplane, respectively. For non-linear problems, SVM allows domain-specific kernel functions which map the coordinate system to a new feature space. SVM achieves good trade-off between model complexity and generalization error. SVM shows peculiar advantages in small sample learning, nonlinear or high-dimensional data patterns recognition tasks.

In this approach, the coordinate values of landmarks are training instances and corresponding number of templates are class labels. In this previous step, we have already manually labelled the images by one of the templates, so a SVM model can be trained on this training data. Considered there are many non-linear factors caused by head-rotation and individual differences, we employed a non-linear kernel Radial Basis Function (RBF) for this task.

The whole process of template matching can be seen in Fig. 4.7. This is a flexible platform that any other applications of FER can employ the 8 fiducial templates obtained on SFEW training data, but not limited on these templates. The algorithm of multi-template design and matching are also suitable for other applications of FER, such as interpersonal relation prediction.

The best result of these three methods will be used for template matching.

### 4.3.3 Texture Reconstruction

Once the template is well matched, we should reconstruct realistic facial textures within this template. Texture reconstruction starts from filling in textures to the base mesh (selected shape template) by a two-step image warping: 1) a warp function  $\mathbf{W}(\mathbf{x}; p)$  is computed to associate the each pixel position from base mesh with the pixel positions of input image  $I$ , and 2) the value each pixel  $\mathbf{x}$  in the warped image  $I(\mathbf{W}(\mathbf{x}; p))$  is obtained by sampling the image  $I$  at that corresponding position. We employed piecewise affine warping method (Matthews & Baker, 2004) to calculate warp function  $\mathbf{W}(\mathbf{x}; p)$ . Piecewise affine warping is based on an assumption that image warping on a small local region can be seen as a linear transformation although whole face warping is nonlinear.

Given a base shape (one of the 8 principal templates), Delaunay triangulation is used to create multiple non-overlapping triangles whose vertices are facial landmark

points. All these triangles make up the mesh. Each triangle accounts for a fairly small region such that it is reasonable to use linearly affine warping.

Let  $s_0$  denote the base mesh whose pixels are denoted as  $\mathbf{x} = (x, y)$ . Assume a pixel  $(x^0, y^0)$  in the base mesh lie into a triangle whose vertices are  $(x_i^0, y_i^0)$ ,  $(x_j^0, y_j^0)$  and  $(x_k^0, y_k^0)$ , this pixel can be uniquely expressed as:

$$(x^0, y^0) = (x_i^0, y_i^0) + \alpha[(x_j^0, y_j^0) - (x_i^0, y_i^0)] + \beta[(x_k^0, y_k^0) - (x_i^0, y_i^0)]$$

$$\text{where } \begin{cases} \alpha = \frac{(x^0 - x_i^0)(y_k^0 - y_i^0) - (y^0 - y_i^0)(x_k^0 - x_i^0)}{(x_j^0 - x_i^0)(y_k^0 - y_i^0) - (y_j^0 - y_i^0)(x_k^0 - x_i^0)} \\ \beta = \frac{(y^0 - y_i^0)(x_j^0 - x_i^0) - (x^0 - x_i^0)(y_j^0 - y_i^0)}{(x_j^0 - x_i^0)(y_k^0 - y_i^0) - (y_j^0 - y_i^0)(x_k^0 - x_i^0)} \end{cases} \quad (4.5)$$

Let  $s$  denote the shape of input face where there is a unique triangle  $(x_i, y_i)$ ,  $(x_j, y_j)$  and  $(x_k, y_k)$  associated with the triangle  $(x_i, y_i)$ ,  $(x_j, y_j)$  and  $(x_k, y_k)$  from base mesh. The results of  $\alpha$  and  $\beta$  are used to calculate the associated pixel position  $(x, y)$  in the input image:

$$\mathbf{W}(\mathbf{x}; p) = (x, y) = (x_i, y_i) + \alpha[(x_j, y_j) - (x_i, y_i)] + \beta[(x_k, y_k) - (x_i, y_i)] \quad (4.6)$$

As the location of each pixel of the base mesh is assigned to the corresponding position in input image, the pixel value of base template is obtained by sampling pixel values from input image  $I$  at corresponding position. The most commonly used sampling strategy is bilinear interpolation.

After piece-wise affine warping, the selected template is filled in textures from input images. However, the warped image may not be realistic due to the self-occlusion caused by out-of-plane head rotation, as is illuminated in Fig. 4.10. There should be further processing to rectify the textures. Current techniques for texture fitting include AAM model instantiation and RSF. We will implement both of them and their comparison will be shown in the experiment section.

**Active Appearance Model:** AAM (Matthews & Baker, 2004) is well know for facial landmark detection. AAM model fitting is employed to reconstruct facial textures. For each query image  $I \in \mathbb{R}^{m \times n}$ , AAM model instantiation minimizes an objective function:

$$\operatorname{argmin}_{\lambda} \| A(\mathbf{x}) - I(\mathbf{W}(\mathbf{x}; p)) \|^2 \quad (4.7)$$

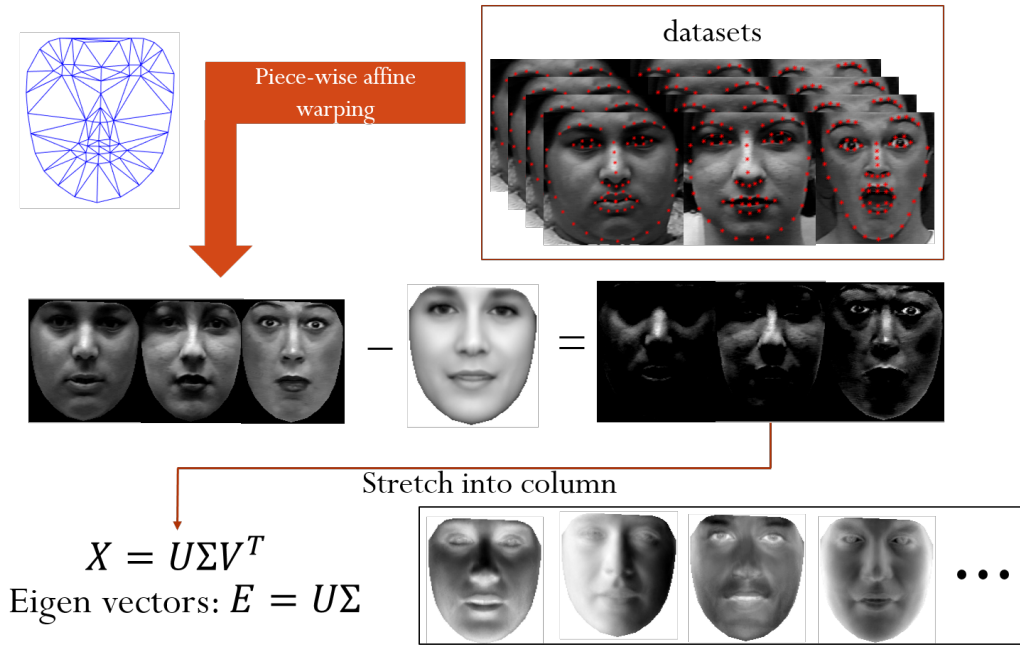


Figure 4.9: Eigen faces

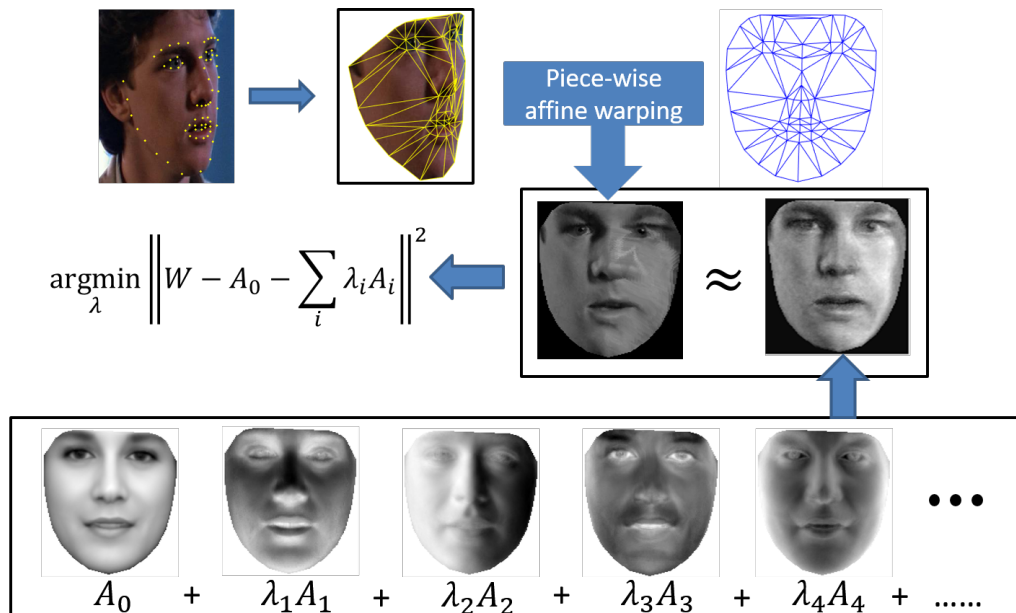


Figure 4.10: Texture Fitting



where  $A(x) = A_0(x) + \sum_{i=1}^m \lambda_i A_i(x)$  is the required frontal face in which  $A_0(x)$  is mean face and  $\sum_{i=1}^m \lambda_i A_i(x)$  is a linear combination of a set of pre-defined eigen faces  $M_A = [A_1(x)|A_2(x)|\dots|A_m(x)]$ , parameterized by  $\lambda$ . The eigen faces are computed by applying Principal Component Analysis (PCA) to a set of warped training images. The original training images should normally contain clean (no occlusion) and frontal faces. They are then shape normalized by piece-wise affine warping their facial shapes and appearances onto a base mesh (selected template). By applying PCA,  $A_i$  is manually set to be  $m$  eigen faces with regards of  $m$  largest eigenvalues. An overview of this process is illuminated in Fig. 4.9.

In our experiment, the shape of input face is automatically detected by SDM. The detected landmarks may sometimes incorrect, which may directly lead to a failure of frontalization result. Therefore, we continue employing AAM gradient search strategy to enhance the landmark detection results. The Equation 4.7 is modified by minimizing:

$$\operatorname{argmin}_{\lambda, \Delta p} \| A(x) - I(\mathbf{W}(x; p + \Delta p)) \|^2 \quad (4.8)$$

where  $p$  is updated by  $p \leftarrow p + \Delta p$ . The linear approximation is given by a Taylor series expansion:

$$I(\mathbf{W}(x; p + \Delta p)) = I(\mathbf{W}(x; p)) + \nabla I \frac{\partial \mathbf{W}}{\partial p} \Delta p \quad (4.9)$$

where  $\nabla I$  is the gradient image,  $\frac{\partial \mathbf{W}}{\partial p}$  is the warp Jacobian evaluated by  $p$ , and  $p$  is the parameter of current shape referred to the equation  $s = s_0 + \sum_{i=1}^n p_i s_i$  defined by Active Shape Model (ASM) (Cootes *et al.*, 1995). The base shape  $s_0$  is the mean shape of all shapes of training images and the eigenvectors  $s_i$  represent shape variance computed by applying PCA to the training shapes. Let us denote the landmark positions as  $s = (x_1, y_1, x_2, y_2, \dots, x_v, y_v)$ . The warp Jacobian is computed by applying chain rule:

$$\frac{\partial \mathbf{W}}{\partial p} = \sum_{i=1}^v \left[ \frac{\partial \mathbf{W}}{\partial x_i} \frac{\partial x_i}{\partial p} + \frac{\partial \mathbf{W}}{\partial y_i} \frac{\partial y_i}{\partial p} \right] \quad (4.10)$$

where

$$\begin{aligned} \frac{\partial \mathbf{W}}{\partial x_i} &= (1 - \alpha - \beta, 0) \quad \text{and} \quad \frac{\partial \mathbf{W}}{\partial y_i} = (0, 1 - \alpha - \beta) \\ \frac{\partial x_i}{\partial p} &= (s_1^{x_i}, s_2^{x_i}, \dots, s_n^{x_i}) \quad \text{and} \quad \frac{\partial y_i}{\partial p} = (s_1^{y_i}, s_2^{y_i}, \dots, s_n^{y_i}) \end{aligned} \quad (4.11)$$

The solution of Equation 4.13 for  $\lambda$  is given as:

$$\lambda = (I(\mathbf{W}(x; p + \Delta p)) - A_0(x))M_A^T \quad (4.12)$$

where  $M_A^T = M_A^{-1}$  since they are orthonormal eigenvectors. Then the solution to  $\Delta p$  is to use Gaussian Newton approximation as:

$$\Delta p = \sum_x \left( \left[ \nabla I \frac{\partial \mathbf{W}}{\partial p} \right]^T \left[ \nabla I \frac{\partial \mathbf{W}}{\partial p} \right] \right)^{-1} \left[ \nabla I \frac{\partial \mathbf{W}}{\partial p} \right]^T [A(x) - I(\mathbf{W}(x; p))] \quad (4.13)$$

The algorithm works iteratively with update rule  $p \leftarrow p + \Delta p$  until converge. The final parameter  $\lambda$  is used to calculate the frontal facial image  $A_0(x) + \sum_{i=1}^m \lambda_i A_i(x)$ . The overview is shown in Algorithm 1.

**Robust Statistical face Frontalization:** RSF can be seen as a variant of AAM. The main advantage is that there is an additional error matrix whose coefficient can be set manually to control the influence of occlusions. RSF is closely related to Transform Invariant Low-rank Texture (TILT) (Zhang *et al.*, 2012). RSF is based on the fact that frontal face image has the minimum rank (smallest value of nuclear norm) when compared to non-frontal face images. Although many different denotations of RSF and AAM represent the same meanings, we still follow the denotations from ???. So an optimization problem can be described as follow:

$$\begin{aligned} & \underset{L, e, c, \Delta p}{\operatorname{argmin}} \|L\|_* + \lambda \|E\|_1 \\ & s.t. \begin{cases} H^{(1)}(\Delta p, c, e) = x(p) + J(p)\Delta p - Uc - e = 0 \\ H^{(2)}(L, c) = L - \sum_{i=1}^k R(u_i)c_i = 0 \end{cases} \end{aligned} \quad (4.14)$$

where  $L$  is low-rank matrix which is expected to be the recovered frontal face.  $E$  is sparse error matrix.  $X(p)$  is the warped image and  $p$  is the parameter of its shape referred to the equation  $s = s_0 + U_s p$  defined by Active Shape Model (ASM) (Cootes *et al.*, 1995) where  $s_0$  is the reference template.  $J(p) = x(p) \frac{\partial W}{\partial p}$  is the Jacobian matrix.  $U = [u_1 | u_2 | \dots | u_k]$  is the pre-computed appearance model (eigen faces computed on only clean frontal faces). Equation  $H^{(1)}$  indicates that the addition of low rank texture  $L$  and sparse error  $E$  agrees with the warped image, such that  $X(p) = L + E$ . In equation  $H^{(2)}$ , the low rank matrix is represented as a linear combination of  $U$  where  $c$  is

its parameter.  $R(\bullet)$  is an operator that reshape a vector to its corresponding matrix. By introducing augmented Lagrangian method (ALM) and alternating directions method of multipliers (ADMM). The parameters can be optimized iteratively.

---

**Algorithm 1** Facial Expression-Aware face Frontalization

---

**Input:**

Test image  $I$ , orthonormal appearance model  $M_A = [A_1(x)|A_2(x)|\cdots|A_m(x)]$ , orthonormal shape model  $M_s = [s_1|s_2|\cdots|s_n]$ , eight templates  $\mathbf{T} = [t_1|t_2|t_3|t_4|t_5|t_6|t_7|t_8]$

- 1: Detect facial shape  $s^{49}$  of 49 landmarks of  $I$  using SDM
- 2: Assign  $I$  to one of the templates  $t_i$  by performing Procrustes analysis on  $s^{49}$ , extracting coordinate features and classifying feature representation using SVM
- 3: initial 66 landmarks  $s$  using SDM and compute  $p = M_s^T(s - t_i)$ .
- 4: **while** not converged **do**
- 5:  $I(\mathbf{W}(x; p)) \leftarrow$  Warp  $I$  onto  $t_i$
- 6:  $J \leftarrow$  compute Jacobian matrix
- 7: **while** not converged **do**
- 8: compute  $\Delta p$  and  $\lambda$  by Equation 4.12 and 4.13 or via inner loop of RSF
- 9: **end**
- 10:  $p = p + \Delta p$
- 11: **end**

**Output:**

frontal face  $A(x) = A_0(x) + \sum_{i=1}^m \lambda_i A_i(x)$

---

The whole RSF includes outer loop and inner loop. Inner loop solves the above optimization problem and returns  $\Delta p$ . Outer loop updates  $p$  by  $p = p + \Delta p$  and then use the new parameter to compute warp image  $X(p)$  and Jacobian matrix.

In (Koelstra *et al.*, 2010), the author does not provide the source code of RSF, so we implement it independently. The time complexity of RSF is relatively high. Both optimization in inner loop and piecewise affine warping in outer loop are time consuming. In practice, there will be more than 120 iterations in each inner loop and no less than 40 iterations in outer loop. This situation makes RSF computationally expensive.

Considering that RSF optimization is an unsupervised linear search, it will lead to

## 4.4 FEAF for Interpersonal Relation Estimation

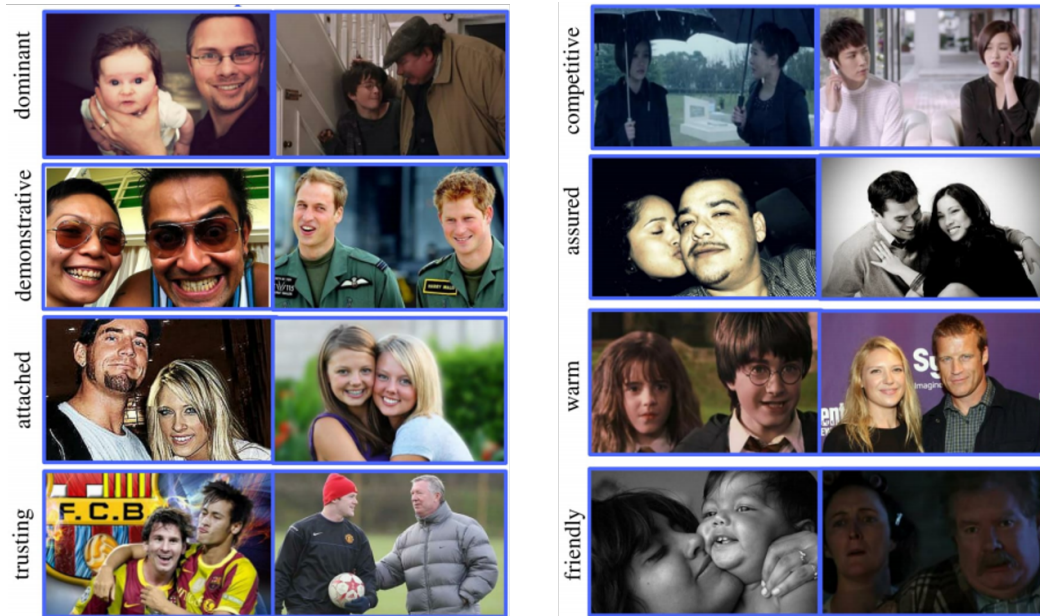


Figure 4.11: Interpersonal relation traits

very slow convergence and expensive time cost. If the initialization is localized near the optimal value, the number of outer loop will drop significantly. In order to achieve the expected initialization, accurate landmark detection is needed. SDM is adequate, but this time 66 points are needed to synthesize the face mesh instead of the original 49 points. The authors of SDM only provide a 49-points detection model and the training part is totally hidden. So we implement SDM and train the 66-points landmark model. The 66 points computed using SDM is not so accurate compared with its 49-points counterpart, but it is still very close to the optimal values. In this strategy, outer loop of RSF can converge within five rounds. By introducing this simple step, this method become much more efficiency. The whole approach is shown in Algorithm 1.

## 4.4 FEAF for Interpersonal Relation Estimation

FEAF is proposed for small-sample learning task. In order to validate its performance on large-scale dataset, we introduce interpersonal relation prediction, as well as the corresponding large-scale dataset, and present a novel approach that integrates FEAF with deep CNN for large-scale data processing.

## 4.4 FEAF for Interpersonal Relation Estimation

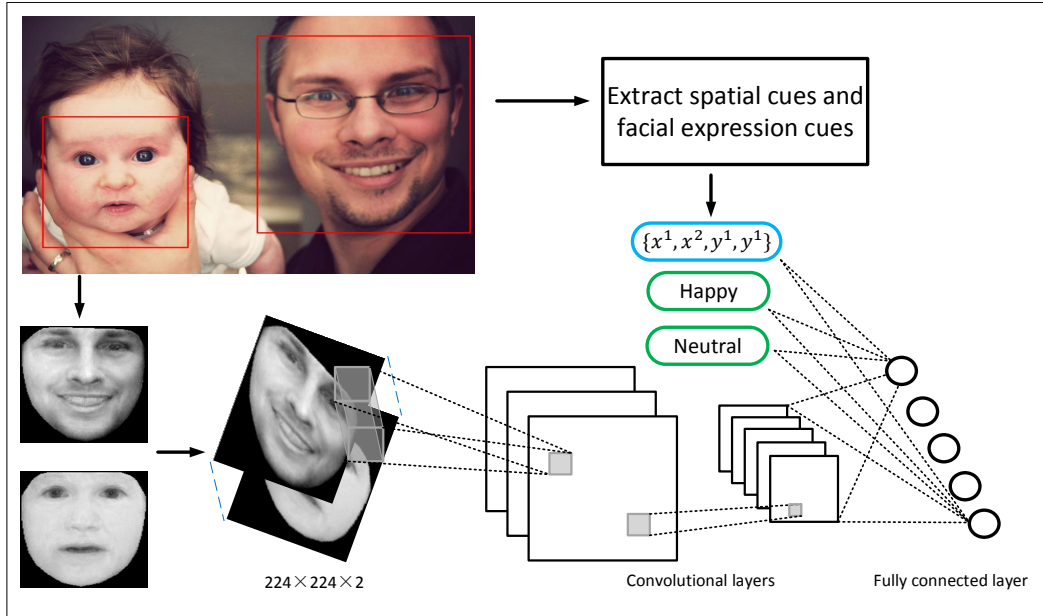


Figure 4.12: Approach for interpersonal relation prediction

In (Zhang *et al.*, 2015) and (Zhang *et al.*, 2018), Zhang *et al.* presented an interpersonal relation dataset and proposed to use a deep learning architecture to predict pair-wise interpersonal relation categories. The dataset is created based on the psychological study from (Kiesler, 1983). As is shown in Fig. 4.11, eight relation traits are defined, which are dominant, competitive, trusting, warm, friendly, involved, demonstrative and assured. Accordingly, their negative counterparts are submissive, deferent, mistrusting, cold, hostile, detached, inhibited and unassured. Each image contains two faces and is labelled by eight independent binary traits. Currently, (Zhang *et al.*, 2015, 2018) from Zhang *et al.* is the only research and benchmark for automatic facial expression-driven interpersonal relation prediction. However, those methods capture features of pair-wise faces separately and, therefore, neglect the mutual features. How to capture mutual facial features is still a gap. In this chapter, we fill this gap by presenting a FEAF-based deep learning approach.

In this work, we propose a new approach for interpersonal relation prediction, where a shared representation of each related two faces is achieved by 1) deep features learned by deep CNN, 2) spatial cues and 3) facial expression cues. Firstly, two faces are detected and normalized to frontal view using FEAF. Then, two frontal faces

are combined in parallel so that a two-channel map (or six-channel map for colour images) is generated, which goes through all the convolutional layers to extract image features. Then, we employed the work from (Zhang *et al.*, 2018) to represent spatial cues as scale-normalized positions of two faces. Facial expression cue is the output of individual emotion recognition predicted by FEAF model. Spatial and facial expression cues directly link to fully connected layer. Finally, we train a deep CNN model as shown in Fig. 4.12.

The configuration of CNN is shown in Table 4.2. A computational block normally includes an either convolutional or Fully Connected (FC) layer, a batch normalization (BN) layer, an Activation function (Rectified Linear Unit, ReLU) and a Max-Pooling layer. Stride describes how many pixel grids should be jumped over when sliding the kernel windows on the image. Padding describes how many additional rows and columns can be added to border of an image and empirically, all the additional elements should be zero. The spatial and facial expression cues are quantified into a six dimensional feature vector and added to the input of block 7. We choose Sigmoid function + Cross Entropy as the loss function because the eight interpersonal relation traits are assumed to be mutually independent.

## 4.5 Experiment

### 4.5.1 Database and experimental design

The proposed method will be evaluated on a public database: Static Facial Expression in the Wild (SFEW) (Dhall *et al.*, 2011). SFEW contains 700 spontaneous facial expressions images annotated by human experts with seven categories: six universal emotions and neutral. The images cover different real-world conditions such as occlusion, low resolution and variations in illumination and head pose. SFEW include 2 image sets: one for training and the other for testing and vice versa. It provides a clear person-independent evaluation protocol in which the images of one specific person with one specific emotion can only exist in one image set, whatever training or testing set. This setup ensures that the experiment is strictly person-independent. SFEW is currently the only in-the-wild facial expression database which was fully annotated by human experts. So, it is the only database that satisfied our needs.



Figure 4.13: Face frontalization on unconstrained images

The experiment is designed as follow. With seven categories of templates, we first train an SVM model on the training data. On the testing data, this SVM model is used for template matching which returns the number of predicted template. Then we perform facial texture fitting on both training and testing images according to the assigned templates. The reconstructed frontal face will be used for the final expression recognition which include feature extraction and emotion classification. In Section 4.5.2, the visual effects of face frontalization will be displayed. In Section 4.5.3, we will show the results of recognition rate of facial expressions and compare our method with the state-of-the-art FER approaches.

## 4.5.2 Face frontalization

In Fig. 4.13, we illuminate the result of face frontalization in different real-world conditions. Fig. 4.13 (a) and (b) shows that out-of-plane head rotation in pan and tilt angles can be well recovered to frontal view. Fig. 4.13 (c) shows that large variations of facial expressions can be well maintained on the reconstructed frontal faces. Fig.6 (d), (e) and (f) displayed the facial images with the problem of illumination changes, low resolution and occlusions, respectively. Our method shows robustness to these problems.

As discussed in Section 4.2, both soft symmetry (Hassner *et al.*, 2015) and RSF (Sagonas *et al.*, 2015) represent the state-of-the-art results of generic face frontalization. Their performances of visual effects on unconstrained images are displayed in Table 4.3. In this experiment, we directly use the open source library of (Hassner *et al.*, 2015) provided by the authors. For RSF, the basic idea is also based on AAM model instantiation and the unique reference template is often neutral. So we implement RSF and use T8 (neutral template) as reference template. In Table 4.3, we select some typical real-world conditions and show the visual results of frontalization methods. For illumination, RSF and FEAF remain the real illumination changes, while soft symmetry may produce fake illuminations when the illumination conditions in the left face sides are different from the right side. Take the second image of illumination column from Table 4.3 as an example, the original image only shows one lighting source in the left side. But in the synthesized image of soft symmetry, it looks like there exists two lighting sources in both left and right sides. For occlusion problem, the visual effects shows RSF and FEAF outperforms soft symmetry. If the occlusion occurs on the pixel enriched face side, occlusion objects may also be used compensate the other face part. So, soft symmetry is quite sensitive to occlusions. For out-of-plane head rotation, all of these methods perform well on small-angle rotation. There are some large distortions of the synthesized images using soft symmetry. RSF and FEAF achieve better results but may fail to reconstruct the illumination changes on the originally invisible facial regions. For the problem of face alignment, face frontalization must start from landmark detection. Although there are many successful landmark detection methods, misalignment is inevitable. Misalignment will cause distortions when using soft symmetry. RSF and FEAF is robust to the landmark location because they are both based on AAM model instantiation which can search for appropriate landmark locations automatically. Actually, RSF and FEAF can achieve good results even if they start from mean shape. But we still use the detected landmarks as initialization because this set-up will significantly improve the efficiency.

Expression reconstruction is the most important issue in our work. In Table 4.3, we can see that all the reconstructed faces by RSF share the same landmark locations, in which many significant facial expressions are lost. Soft symmetry can well reconstruct expressions on some images, but there is very large distortions, which will lead to failure in recognizing facial expressions. FEAF works well on both frontalization





Figure 4.14: Face frontalization

and expression reconstruction. In general, soft symmetry is sensitive to illumination changes, occlusions and misalignment. Its advantage is that it works well on maintaining expressions to some extent. RSF and FEAF performs well on nearly all the unconstrained real-world conditions. The problem of RSF is that it totally failed in maintaining the facial expressions. FEAF is the most effective method in reconstructing facial expressions.

### 4.5.3 Facial expression recognition

In this section, we evaluate the frontalization model on facial expression recognition tasks. For each frontalized face, we use two popular local descriptors: Local Binary Pattern (LBP) (Huang *et al.*, 2011) and Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) for feature extraction and SVM for emotion classification.

**Histogram of Oriented Gradients** is popular feature descriptor to describe local spatial relations, which has been successfully used for object detection. It divides an image into small adjacent regions and counts occurrences of gradient orientation in each region. The gradient orientation is achieved by sliding a derivative mask (weight kernel) through the whole image and calculating weighted summation within each sliding window. Since both LBP and HOG descriptors are in histogram manner, we can combine the two features together by cascading the two histograms, denoted as LBP+HoG. The detailed introduction of LBP and SVM can be referred to Section 3.3.3 and Section 3.3.4, respectively.

Table 4.4 shows the confusion matrix for FER attained by cascaded histogram of LBP+HoG features and SVM classifier. As can be seen, Happy achieves the best performance while Neutral are easily confused with other categories. This is because Happy is closely related to template T1 which is the most discriminative template. Meanwhile Neutral faces are not often related to neutral template T8 due to individual differences, which results in a low recognition rate of Neutral category.

In Table 4.5, we show FER performance of the three generic frontalization approaches and make a comprehensive comparison. LBP+SVM representation performs better than single LBP feature. The results of single HOG representation is not listed because of its poor performance. Soft symmetry based approach performs worse due to the large distortion of the face reconstruction and sensitivity to occlusions. RSF performs a little better because of its robustness to occlusions and realistic face reconstruction results. But the accuracy is still low because it often fails to reconstruct detailed facial expressions. Our methods outperforms the other two methods, which demonstrates its effectiveness of facial expression reconstruction.

Table 4.6 shows the comparison with the state-of-the-art approaches for FER in the wild. The result of baseline is obtained from database creators (Dhall *et al.*, 2011). It neither employs view-invariant approaches nor addresses the problem of discriminative

feature learning. So the baseline shows inferior performance than most view-invariant FER methods. In (Eleftheriadis *et al.*, 2015), the latest approach of view-normalization model is presented. It achieves relatively high accuracy in Disgust and Sadness. Nevertheless, its recognition rates of Neutral and Surprise are extremely poor, even lower than random guess. As previously discussed, its performance highly relies on a large amount high-quality training data. Therefore, we can conclude that this method is unstable in small sample learning tasks. In (Liu *et al.*, 2013), a dictionary-learning model is displayed and its result is a little better than (Eleftheriadis *et al.*, 2015). The results on all of the emotions categories are better than baseline, which reveals its good stability. Our method achieves an outstanding performance when compared with these methods. The average recognition rate improves over 20%, which indicates a significant improvement in the research of FER in the wild. The accuracy of both Angry and Happy is more than 50% accuracy, while the others also achieves a considerable improvements. It can be seen that the other two frontalization methods are also superior over the state-of-the-art approaches according to Table 4.5 and Table 4.6. Although face frontalization was originally proposed for face recognition, it is proved that face frontalization is also valuable for FER. Compared with traditional multi-view FER approaches, face frontalization is also a good choice for FER in the wild. What is worth mentioning is that although great breakthrough has been made in this research, the accuracy of FER using deep learning methods is still slightly higher than our methods. But deep learning must be based on extremely large amount of extra training data and its experiments are not strictly person-independent, which does not comply with the evaluation protocol of SFEW. Our research is still currently the best work for small sample learning task.

### 4.5.4 Interpersonal Relation Prediction

The experiment in this section is designed to demonstrate the flexibility of FEAF for different applications of facial expression analysis. As is described in Section 4.4, we develop a novel deep structure that flexibly integrates FEAF with DCNN. In this Section, we evaluate this approach on interpersonal relation prediction dataset (Zhang *et al.*, 2018).

This is a recent released public dataset containing 8016 unconstrained facial images chosen from web and movies. Each image is annotated with faces' bounding boxes and labelled by 8 independent binary categories (positive or negative) of relation traits. Different from SFEW, this is a large-scale dataset which enable training a deep CNN model.

To evaluate the performance, we follow the evaluation metric from (Zhang *et al.*, 2018) where the accuracy is defined by:

$$accuracy = 0.5 \times \left( \frac{n_p}{N_p} + \frac{n_n}{N_n} \right) \quad (4.15)$$

where  $N_p$  and  $N_n$  represent the numbers of positive and native samples, respectively, and  $n_p$  and  $n_n$  are the numbers of true positive and true negative samples. This metric accounts for the imbalanced attribution of samples.

The result is shown in Table 4.7. As this dataset is released recently, (Zhang *et al.*, 2018) is the only result that can be introduced for comparison. Our method has a 10% superiority than (Zhang *et al.*, 2018). Furthermore, the authors of (Zhang *et al.*, 2018) collected over 318K external images for training while we only use the images from relation trait dataset without introducing any other images. This result demonstrates that FEAF can be flexibly integrated FEAF with DCNN and work well on large-scale dataset.

## 4.6 Summary

In this chapter, to the best of our knowledge, we present the first work that jointly considers all the non-linear factors for FER in the wild. For the problem of identity bias, we develop a multi-template model to normalize the various facial geometric differences (the outline of face, different size and displacement of facial organs). Then, we employed the latest techniques of face frontalization to solve the problem of head-pose changes and occlusions. Finally, T-measure was proposed to carefully select principal templates, which helped to reduce the large intra-class variance caused by the irregularity of spontaneous expressions.

The state-of-the-art performance is achieved in the task of static facial expression recognition in the wild. In the next chapter, we aim to capture more subtle facial

## **4.6 Summary**

---

expressions and further improve the face frontalization-based model for dynamic FER.

Table 4.2: The configuration of CNN

	Layers	Kernel Size	Stride	Padding
Block 1	Convolution	$11 \times 11 \times 2 \times 64$	4	0
	BN	-	-	-
	ReLU	-	-	-
	Max-Pooling	$3 \times 3$	2	1
Block 2	Convolution	$5 \times 5 \times 64 \times 256$	1	2
	BN	-	-	-
	ReLU	-	-	-
	Max-Pooling	$3 \times 3$	2	0
Block 3	Convolution	$3 \times 3 \times 256 \times 256$	1	1
	BN	-	-	-
	ReLU	-	-	-
Block 4	Convolution	$3 \times 3 \times 256 \times 256$	1	1
	BN	-	-	-
	ReLU	-	-	-
Block 5	Convolution	$3 \times 3 \times 256 \times 256$	1	1
	BN	-	-	-
	ReLU	-	-	-
	Max-Pooling	$3 \times 3$	2	0
Block 6	FC	$6 \times 6 \times 256 \times 4096$	1	0
	BN	-	-	-
	ReLU	-	-	-
Block 7	FC	$1 \times 1 \times 4102 \times 4096$	1	0
	BN	-	-	-
	ReLU	-	-	-
Block 8	FC	$1 \times 1 \times 4096 \times 8$	1	0
Loss	Sigmoid + Cross Entropy	1		

	expressions	illumination	occlusion	head pose	misalignment
query image					
soft symmetry					
RSF					
Ours					

Table 4.3: Comparison of generic Frontalization methods

Table 4.4: Confusion matrix (%) on SFEW databse

	Angry	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Angry	51.79	2.68	5.36	10.71	12.50	7.14	9.82
Disgust	9.41	41.18	14.12	11.76	15.29	7.06	1.18
Fear	9.09	5.05	42.42	7.07	11.11	7.07	18.18
Happy	4.39	4.39	4.39	76.32	1.75	4.39	4.39
Neutral	16.00	10.00	9.00	6.00	36.00	12.00	11.00
Sadness	17.17	7.07	8.08	17.17	9.09	38.38	3.03
Surprise	15.38	3.30	15.38	10.99	14.29	3.30	37.36



Table 4.5: Comparison of recognition rate (%) of generic frontalization methods on SFEW database

Frontalization	feature extraction	Angry	Disgust	Fear	Happy	Neutral	Sadness	Surprise	Average
RSF	LBP	24.11	28.24	24.24	41.32	31.00	25.25	28.57	29.15
	LBP+HOG	30.57	25.88	27.18	42.50	34.00	25.25	29.67	31.19
Soft symmetry	LBP	22.32	21.18	23.23	44.74	23.00	28.28	19.78	26.57
	LBP+HOG	26.79	17.65	30.30	44.74	26.00	21.21	17.58	27.00
ours	LBP	45.54	32.94	45.45	74.56	32.00	31.31	36.26	43.57
	LBP+HOG	51.79	41.18	42.42	76.32	36.00	38.38	37.36	47.14

Table 4.6: Comparison of recognition rate (%) of the state-of-the-art methods on SFEW database

	Angry	Disgust	Fear	Happy	Neutral	Sadness	Surprise	Average
Baseline	23.00	13.00	13.90	29.00	23.00	17.00	13.50	18.90
(Liu <i>et al.</i> , 2013)	25.89	28.24	17.17	42.98	14.00	33.33	10.99	24.70
(Eleftheriadis <i>et al.</i> , 2015)	24.11	14.12	20.20	50.00	23.00	23.23	21.98	26.14
FEAF	<b>51.79</b>	<b>41.18</b>	<b>42.42</b>	<b>76.32</b>	<b>36.00</b>	<b>38.38</b>	<b>37.36</b>	47.14
(Mollahosseini <i>et al.</i> , 2016)	-	-	-	-	-	-	-	<b>48.6</b>

Table 4.7: Comparison of relation traits prediction performance

	HOG+SVM	(Zhang <i>et al.</i> , 2018)	ours
Accuracy	62.1%	70.0%	80.47%

## Chapter 5

# Cascade Regression-Based Face Frontalization for Dynamic FER

Facial expressions are inherently subtle facial muscle movements which can be better analyzed in a dynamic manner rather static. Dynamic facial expression recognition methods capture facial movements by means of modelling spatio-temporal features. The influence of identity bias can be significantly diminished by dynamic approaches because facial temporal features have nothing to do with identity-related cues but are completely associated to expression changes. Although dynamic approaches provide a good solution to identity bias challenge, the problems of occlusions and head rotation variations still remain. Currently, most existing approaches only considered the registration of time changes but ignored the above two problems. In Chapter 4, face frontalization has been proved to be quite effective for head pose normalization and occlusion obliteration. However, FEAF is based on several meaningful templates, which is not able to track facial changes in an image sequence. In this chapter, we presents a novel sequential face frontalization method for dynamic facial expression analysis.

### 5.1 Introduction

FER has various applications including, but not limited to, human computer interaction (HCI) (Dureha, 2014; Wang *et al.*, 2015; Wu *et al.*, 2008), animation (Aneja *et al.*, 2016; Ichim *et al.*, 2015; Thies *et al.*, 2016) and security (Pfister *et al.*, 2011). The

existing works on expression analysis often focus on static facial images processing. However, facial expressions are inherently dynamic actions which can be better described as several sequential pieces of facial motions in a time interval. Although static FER methods have achieved considerable results, they completely ignored discriminative features conveyed by subtle facial muscle movements. Dynamic FER on the whole image sequence is more natural and reasonable.

The aim of dynamic FER is to predict facial expression categories from an image sequence in which expression evolutionary process often evolves initial neutral state, onset, apex phase, offset and final neutral state. The existing dynamic facial feature descriptors can be classified into two categories: low-level feature representation and high-level facial motion representation. The best known low-level feature representations are LBP-TOP (Zhao & Pietikainen, 2007) and LPQ-TOP (Jiang *et al.*, 2014) which capture the local gradient features over both spatial neighbourhoods in one frame and temporal neighbourhoods between frames. High-level semantic methods aim to derive a meaningful facial motion representation in which temporal alignment is often presented to different sequences with different time interval into a uniform temporal space (five states of expression evolutionary process) (Guo *et al.*, 2016). As both two categories address the problem of temporal modelling, they have been proved to be effective on the posed facial expression data sets but poor on unconstrained image sequences where there is large head-pose changes and occlusions. Currently, there is no attempt that focuses on head-pose normalization for dynamic facial expression recognition.

Intuitively, the problem head-pose changes and occlusions from unconstrained facial images can be well normalized through face frontalization (Ferrari *et al.*, 2016; Zhu *et al.*, 2016). The main objective of generic face frontalization is to automatically recover the non-frontal face to its frontal view from a single image. In general, face frontalization includes two key components: frontal face-shape estimation and frontal face-texture fitting. Frontal shape estimation localizes facial key point positions and aligns them to the frontal positions. The objective of frontal shape estimation is to align the non-frontal facial landmarks to their frontal positions. The task of texture fitting is to fit textures to the predicted shape by texture warping and rectification. It has been reported that frontal shape estimation is quite challenging, the mainstream

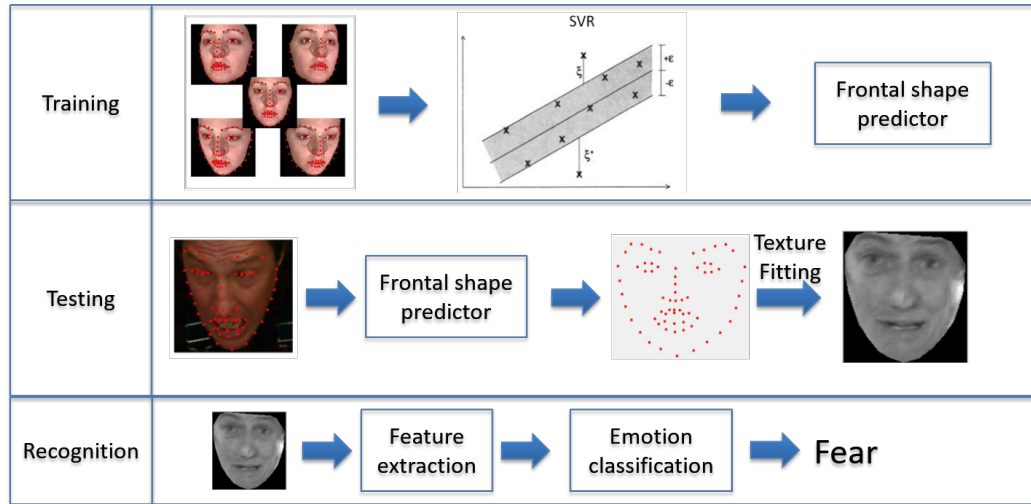


Figure 5.1: Overview of the proposed method

approaches focus on hard frontalization that an unmodified frontal shape template (often made in neutral) is used as reference and facial textures of all the query faces will be fitted to the template (Zhu *et al.*, 2015). By this strategy, the reconstructed faces will lose of facial expression related information. There is only a few approaches so far that performs face frontalization with full considerations of facial expressions.

In this chapter, a new facial expression-aware face frontalization method is proposed for dynamic analysis of facial expressions. The key issue is how to predict the frontal position of facial key points given detected non-frontal shape. Inspired by the success of regression based 2D face alignment and morphable model in (Asthana *et al.*, 2011; Guo *et al.*, 2016; Xiong & De la Torre, 2013), we propose a novel cascade regression model for 2D frontal shape estimation. As is shown in Fig. 5.1, a set of facial images in different viewpoints is collected and each non-frontal face is associated with its frontal counterpart. Several regression approaches are chosen to learn the pair-wise relations between non-frontal shape and frontal shape. It is obvious that pair-wise changes in head-pose, expressions and individual differences are non-linearly coupled in 2D shape. So one step regression cannot well model this relation. We propose an adaptive cascade regression model so that the non-frontal shape can be gradually approximated to its frontal shape. In this training stage, each cascade learns a function that maps the input shape to the most approximated groundtruth position.

During testing, the cascade regression model will be used as frontal face-shape predictor that estimates the key point positions in frontal view. The obtained frontal shape will be viewed as the based mesh. Then in the texture-fitting step, we employ Active Appearance Model (AAM) instantiation (Matthews & Baker, 2004) to reconstruct the facial appearances. The reconstructed face will be: (a) in frontal view, (b) remains deformations of detailed facial expressions, (c) remove occlusions.

The contributions of the paper are summarized as follows:

- 1) We propose a cascade regression-based face frontalization approach. Different from computational expensive 3D solutions which suffers from one-minute-per-frame problem, this method is based on 2D face reconstruction and works in real-time.
- 2) Different from the existing 2D face frontalization methods that loses deformations of expressions, the proposed frontalization method is expression-aware. The vivid expression changes are remains. Meanwhile, the occlusions are removed.
- 3) We demonstrate, for the first time, that 2D face frontalization is also effective for dynamic analysis facial expressions in-the-wild.

## 5.2 Related Works

Dynamic FER models capture spatio-temporal features which represent a range of frames within a time interval. As is mentioned above, The existing methods can be divided into low-level feature based and motion based approaches.

Low-level spatio-temporal representations can be seen as an extension of low-level spatial representations. Shape features are described by tracked facial fiducial points. The location of each pints, as well as the length and angle of pair-wise points connection, forms the basic shape features. Till now, shape representations are less common because it has been reported and validated over and over again that appearance models outperform shape models (Sariyanidi *et al.*, 2015). Appearance representation are the mainstream for dynamic FER. LBP-TOP is a popular method that extracts Local Binary Pattern (LBP) features from Three Orthogonal Planes (TOP) (Zhao & Pietikainen, 2007). The original LBP features extracted from a single spatial plane are extended to two more spatio-temporal plane, which enables extracting gradient features between frames. LPQ-TOP follows the same principle and is used to Action Unit (AU) recognition (Jiang *et al.*, 2014).

Obviously, low-level features didn't consider the specific knowledge in facial expression domain. Recent research focuses more on capturing high-level semantic features represented by facial motions. It is commonly accepted that facial expression process of human beings includes five states: initial neutral, onset, apex, offset and final neutral. This standard process can be seen as a template and motion based methods are actually a time alignment strategy that normalizes the input sequence to the five reference states. Koelstra et al.(Zhao & Pietikainen, 2007) used free-form deformations (FFD) (Rueckert *et al.*, 1999) based nonrigid registration to capture motions for AU recognition. Guo et al.(Guo *et al.*, 2016) used diffeomorphic transformation for time alignment and proposed atlas construction to capture facial appearance movements. Wang et al. (Wang *et al.*, 2013) assumed each local facial points movement as a local facial event and learned the motion dependency by modelling temporally overlapping facial events and their temporal relations by interval temporal Bayesian network. In (Liu *et al.*, 2014a), a Universal Manifold Model (UMM) is learned to statistically unify each input video (modelled as spatio-temporal manifold via low-level features) to the standard expression evolutionary process.

All the methods mentioned above only address the problem of temporal state alignment, but ignore spatial texture alignment. If the subjects of the video clips move their head frequently, the appearance changes caused by head-pose will be much larger than subtle expressions changes. This is why most dynamic FER methods performs good on the posed expression data sets but poor on the unconstrained image sequences. A well designed dynamic descriptor should consider both time alignment and facial appearance normalization. The problem of facial appearance normalization for suitable expression registration is quite challenging that no previous work has been done on it.

Face frontalization aims to recover the frontal facial appearances from non-frontal images. It is a comprehensive research topic which is often associated with face alignment, face deformation and texture rectification. It provides a good way to normalize facial appearances, in which the research output of face frontalization seems to be satisfied with the demand of dynamic FER mentioned above. But current methods are not suitable for this task due to many different problems.

If a facial image is captured from non-frontal view, there will be one face side containing rich pixels while the other side missing some pixels. Direct interpolation will cause large distortions on the reconstructed images. Therefore, the main problem



of face frontalization is how to fill in the invisible part. Current methods for face frontalization includes 2D based model and 3D based model.

In (Jeni *et al.*, 2015) and (Roth *et al.*, 2015), two approaches of person-specific 3D model reconstruction are performed, in which several images captured from one person in different poses and expressions are used to reconstruct his/her 3D model. The main drawback of these methods are they are unable to reconstruct 3D surface of novel faces. In order to deal with this problem, many methods were proposed based on 3D Morphable Model (3DMM), which is, theoretically, capable of reconstructing a full 3D facial surface from a single input image (Ferrari *et al.*, 2016; Zhu *et al.*, 2015, 2016). Although 3D based methods can implement frontalization, they are not practical since a) they are computational expensive to build 3D model, b) a massive training data to learn shape models are required, and c) some of them will fail in reconstructing 3D model of novel subjects.

In (Xue *et al.*, 2013) and (Shan *et al.*, 2009), two effective 2D frontalization approaches are presented and both of them belong to hard frontalization which employs a single 2D/3D reference template as base shape and all the query images will fit their facial textures to it. Separately, Soft symmetry (Xue *et al.*, 2013) fills in the non-visible regions by the corresponding symmetric visible parts of face. Apparently, this approach is sensitive to occlusions and it only enables tilt head rotation recovery but fails in recovering the faces in pan angles. The texture-fitting strategy of Robust Statistical face Frontalization (RSF) (Zhu *et al.*, 2016) is based on Active Appearance Model (AAM) instantiation (Matthews & Baker, 2004) which reconstruct the appearances by combining a group of eigen faces. It is robust to occlusions and capable of recovering the faces in whatever pan or tilt angles. Therefore, RSF is more stable than soft symmetry.

Recently, deep learning methods are also used face frontalization. In (Tran *et al.*, 2017), Generative Adversarial Network (GAN) is used to generate frontal faces. The results showed that even very large head-pose can be recovered.

These methods have been proved to be effective in face recognition task. However, the reconstructed faces are expected to approximate to real frontal faces in only the identity and the expressions are more or less removed. As far as we know, there is only one work on facial expression-aware face frontalization (FEAF) (Wang *et al.*, 2016). In this approach, multiple emotional shape templates are designed instead of single shape

template and it achieves good results in static FER. Inherently, it is still hard frontalization and not suitable for dynamic FER because all the frames of image sequences will be arbitrarily normalized to finite templates so that the dynamic information of subtle expression changes, as well as shape changes, will be lost. Dynamic FER requires a novel face frontalization method that is able to not only recover appearances to frontal view, but also distinguish very subtle changes in facial shape and appearances.

## 5.3 Methodology

We propose a new face frontalization method which synthesizes subtle expression-aware frontal faces. There are mainly two problems: cascade regression-based frontal face-shape estimation and AAM-based frontal face-texture fitting. There is no need to further introduce texture fitting strategies since we have elaborated on them in Section 4.3.3. For the first problem, we firstly introduce several regression approaches, and then present a novel cascade regression-based model to solve this problem.

### 5.3.1 Regression Approaches

In this stage, we propose the problem of learning an associated pattern between the non-frontal facial shape and its corresponding frontal counterpart in a regression manner. Given a pair of shape vectors:

$$\mathbf{x} = [x^1, y^1, x^2, y^2, \dots, x^n, y^n] \quad \text{and} \quad \mathbf{x}_0 = [x_0^1, y_0^1, x_0^2, y_0^2, \dots, x_0^n, y_0^n] \quad (5.1)$$

representing non-frontal and frontal facial annotations, respectively, the problem is to learn a regression function  $\mathcal{R}$  that makes  $\mathcal{R}(\mathbf{x})$  most approximate to  $\mathbf{x}_0$ . We implement three popular regressors for this task: Linear Regression, Support Vector Regression (SVR) and Gaussian Process Regression (GPR).

**Linear Regression:** The regression function  $\mathcal{R}$  refers to:

$$\mathcal{R}(x) = \langle \omega, x \rangle + b \quad (5.2)$$

where  $\omega$  is the regressor expected to be learned from training data and  $b$  is bias. Consider the training data  $\{(x^1, y^1), \dots, (x^l, y^l)\}$ , where  $x = x^1, \dots, x^l \in \mathbb{R}^n$  is the  $n$ -dimensional feature vector and  $y = y^1, \dots, y^l \in \mathbb{R}$  is the response, linear regression

solve a least square problem by minimizing:

$$\operatorname{argmin}_{\omega} \| y - \omega^T x \|^2 + \lambda \| \omega \|^2 \quad (5.3)$$

where  $\lambda > 0$  is a regularization coefficient used to avoid over-fitting. The solution for  $\omega$  is given as:

$$\omega = yx^T(x x^T + \lambda I)^{-1} \quad (5.4)$$

where  $I$  is the identity matrix.

**Support Vector Regression:** SVR also Considers the linear function  $y = \langle \omega, x \rangle + b$ . It aims to learn a soft margin instead of a solid boundary so that the prediction becomes less sensitive to noise and errors. The SVR function can be expressed as:

$$\begin{aligned} \min \quad & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l (\xi_i^- + \xi_i^+) \\ \text{s.t.} \quad & \begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \epsilon + \xi_i^+ \\ \langle \omega, x_i \rangle + b - y_i \leq \epsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \end{cases} \end{aligned} \quad (5.5)$$

where  $C > 0$  is a constant which make the balance between maximum margin and tolerance  $\epsilon$ , and  $\xi$  is the slack variable which suggests that part of error is tolerated.

After introducing dual problem and Lagrangian multipliers. The optimization problem becomes:

$$\begin{aligned} \max \quad & \begin{cases} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ - \epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \end{cases} \\ \text{s.t.} \quad & \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \quad (5.6)$$

where  $\alpha_i$  and  $\alpha_i^*$  are the Lagrangian multipliers.

Kernel function is a trick in SVR to solve non-linear problems. The model is obtained by replacing the dot product  $\langle x_i, x_j \rangle$  in Equation 5.6 by a kernel function  $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ , where  $\phi(x)$  is a human-designed transformation rule that

map  $x$  to a transformed feature space. Radial Basis Function (RBF) is a commonly used kernel where  $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$  and this function is applied in this chapter. There are two different models of SVR in used for facial spatial alignment task.

**Gaussian Process Regression:** Similar to non-linear SVR, GPR also starts from linear function  $y = \langle \omega, x \rangle + b$  goes into kernel-based model. GPR is a probabilistic model that introduces an unobserved variable  $f(x)$  in a spatial inference process. The latent variable  $f(x_i)$  is selected from a Gaussian Process (GP) which is a set of random variables whose any arbitrary subsets complying with a joint Gaussian distribution. By this assumption, we suppose a new testing instance  $x_{new}$  interpreted as  $f(x_{new})$  will fall into the same joint Gaussian distribution. So the problem is how to achieve this extended distribution. This is formalized by:

$$f(x) \sim GP(m(x), K(x, x')) \quad (5.7)$$

where  $m(x)$  is the mean of  $f(x_1), f(x_2), \dots, f(x_l)$  and  $K(x, x')$  is the covariance matrix. Given an observed training set  $(x_1, f(x_1)), \dots, (x_l, f(x_l))$ , we suppose they have been modelled by  $f \sim N(\mu, K)$ . According to GP assumption, we have a new instance of  $x^*$  following:

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N \left( \begin{pmatrix} \mu \\ \mu^* \end{pmatrix}, \begin{bmatrix} K & K^* \\ K^T & K_{**} \end{bmatrix} \right) \quad (5.8)$$

The posterior probability of  $f^*$  can be computed as:

$$P(f^* | x^*, x, f) \sim N(\mu^*, \Sigma^*) \quad (5.9)$$

where  $\mu^* = \mu(x^*) + K_*^T K^{-1}(f - \mu(x))$  and  $\Sigma^* = K_{**} - K_*^T K^{-1} K^*$ . The estimated result is obtained by  $y = f(x) + \epsilon$ , where  $\epsilon \sim N(0, \sigma_y^2)$ .

### 5.3.2 Cascade Regression Model

The problem of facial spatial alignment is obviously non-linear. Although the above mentioned regression approaches are able to solve non-linear problem, they are still based on only a single heuristic. The presented problem should model facial changes under various head-pose and expressions, which is too complex to be solved in one-step

regression. Therefore, we develop an adaptive cascade regression model that learns the frontal-profile associations in a cascade manner that gradually approximates to optimum in several steps of regression rather than only one step.

In the frontalization task, the landmark positions are normalized through Procrustes analysis in which the in-plane rotation and size of the face are adjusted. Given  $M$  annotated facial image pairs of non-frontal and corresponding frontal faces, the linear function can be defined as  $\mathbf{x}_0 + \Delta\mathbf{x} = \langle \omega, \mathbf{x}_0 \rangle + b$ , where  $\mathbf{x}_0$  and  $\mathbf{x}$  represents the shape vector for non-frontal and frontal images, respectively. So  $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$  is known. Consequently, the final objective of regression can be expressed as:

$$\Delta\mathbf{x} \leftarrow \langle \omega, \mathbf{x}_0 \rangle + b \quad (5.10)$$

This equation can be referred to as the linear function of all the three regression methods mentioned above. This function is illuminated in Fig. 5.2.

Then we introduce the cascade regression manner regarding  $\Delta\mathbf{x}^i$  representing the obtained  $\Delta\mathbf{x}$  in the  $i$ th cascade. In each cascade, we revise Equation 5.10 into:

$$\Delta\mathbf{x}^i \leftarrow \langle \omega^i, \mathbf{x}_0^i \rangle + b^i \quad (5.11)$$

and train a regression model at current stage. In the new round,  $\mathbf{x}_0^{i+1} = \Delta\mathbf{x}^i + \mathbf{x}_0^i$ ,  $\Delta\mathbf{x}^{i+1} = \mathbf{x} - \mathbf{x}_0^{i+1}$  and they will be used to train a regression model at this round.

With the cascade regression model works iteratively. The algorithm will stop when these parameters and  $\Delta\mathbf{x}^i$  turn zero. Empirically, the algorithm converges in 4 or 5 steps.

During testing, the non-frontal facial landmarks should be localized first. There are many existing facial landmark detection methods that has been proved to be effective. With the obtained facial landmarks, frontal face-shape will be estimated using Equation 5.11 sequentially.

By performing cascade on three regression approaches, we achieved three independent cascade regression models. As is mentioned in Section 5.1, a single heuristic is not sufficient to model such a complex problem of spatial alignment. We further introduce ensemble to combine all the regression models together and make a comprehensive decision. Ensemble learning is a model combination strategy that combines all the models together and then votes for the final prediction. We applies the ensemble and makes the prediction by averaging each results predicted by the cascade regression models.

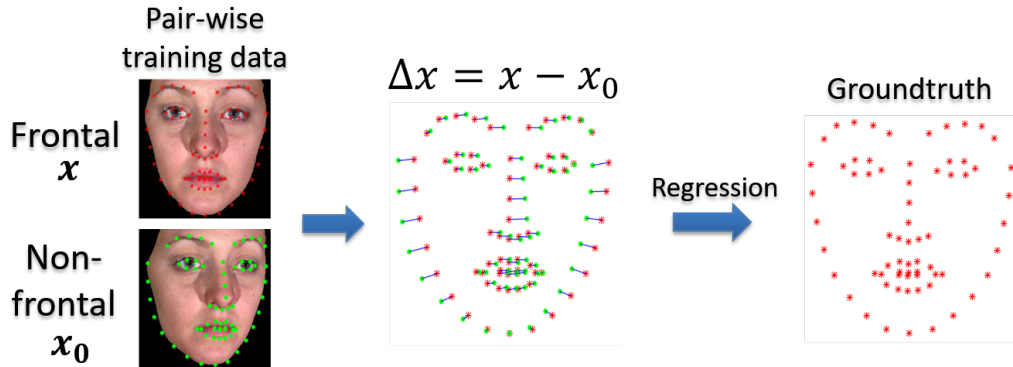


Figure 5.2: Cascade regression for frontal shape estimation

## 5.4 Experiment

The performance of the proposed method has been validated in three tasks: 1) regression error of spatial alignment, 2) static FER in the wild, 3) dynamic FER in the wild. For two FER tasks, the feature extraction and classification methods for static and dynamic FER are different. So, we will give a brief introduction, separately.

### 5.4.1 Training data collection

Binghamton University 3D Facial Expression (BU3DFE) is a static 3D facial expression database which include 100 subjects with 2500 3D facial expression models. The training data are captured by rendering 2D images using 3D models. Images are captured at 7 pan angles ( $-45^\circ$ ,  $-30^\circ$ ,  $-15^\circ$ ,  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$ ,  $45^\circ$ ) and 5 tilt angles ( $-30^\circ$ ,  $-15^\circ$ ,  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$ ), which results in totally 35 different viewpoints, as is shown in Fig. 5.4. Each training instance includes the position landmark points in one of the 34 non-frontal rotations and the corresponding points in frontal pose.

BU3DFE provide the 3D position of 83 landmarks for each 3D facial model. When the 3D landmark points were projected to 2D plane, there would be misalignments especially when there was large out-of-plane rotation. As is shown in Fig. 5.3, the red points of 3D projection are obviously misaligned. So we use OpenFace to automatically detect landmark points. OpenFace (Baltrušaitis *et al.*, 2016) is a very simple and effective tool for facial landmark detection. Most landmarks can be well detected

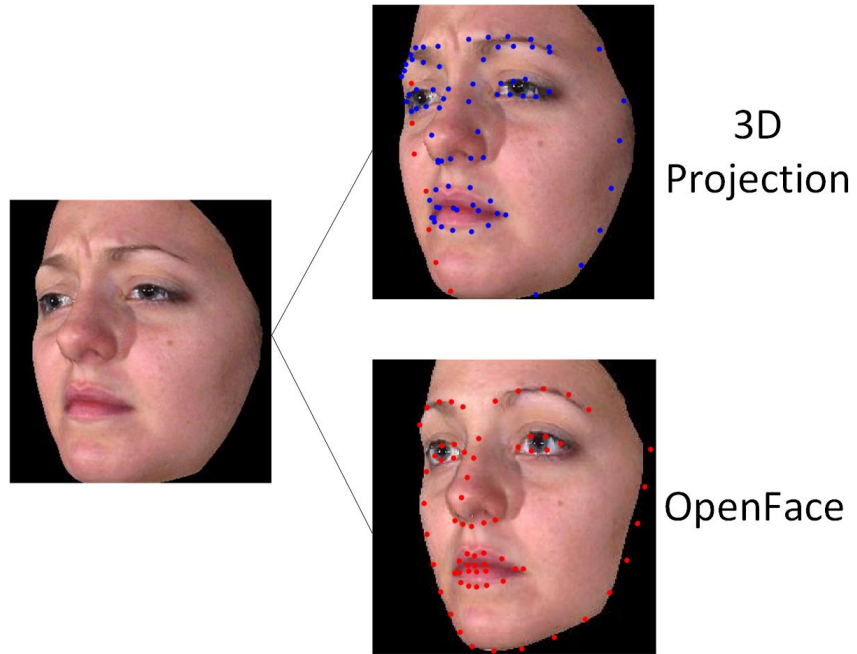


Figure 5.3: Landmark positions: 2D vs 3D

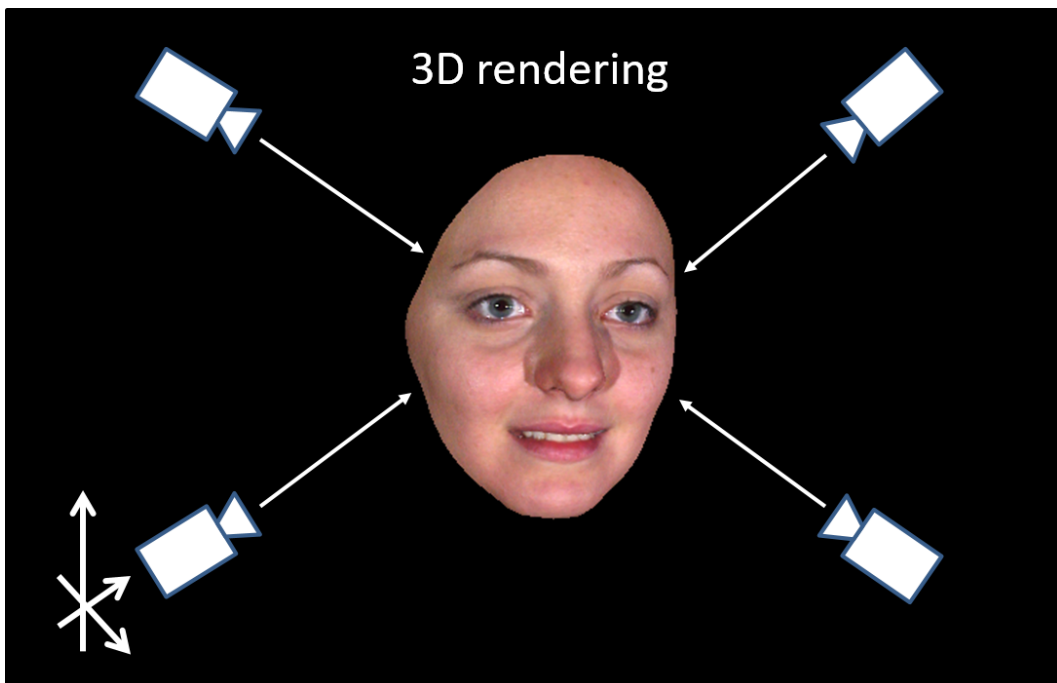


Figure 5.4: 3D rendering

Table 5.1: Alignment Error (%) of different regression methods

	Linear	SVR(linear)	SVR(RBF)	GPR	Ensemble
Single-step Regression	18.87	18.44	16.99	11.13	14.88
Cascade Regression	13.82	12.92	12.40	10.84	10.53

using this software. Misaligned points were manually revised. This process generates 1887 training examples in pairs of  $[\mathbf{x}, \mathbf{x}_0]$ .

### 5.4.2 Spatial Alignment

In this section, we measure the accuracy of frontal facial shape prediction based on the generated 1887 example pairs. We perform a 10-folder cross-validation strategy to evaluate the performance of different regression model. The alignment error is measured by:

$$err = \sqrt{(x - x_0)^2 + (y - y_0)^2} / l \quad (5.12)$$

where  $(x, y)$  and  $(x_0, y_0)$  are aligned position and groundtruth, respectively, and  $l$  is the width of facial bounding box which is calculated by  $(\max(x_0) - \min(x_0) + \max(y_0) - \min(y_0)) / 2$ .

Table 5.1 shows the comparison of alignment error using the regression approaches mentioned in Section 5.3.1. We can several cues from this table:

1) Non-linear regression approaches (GPR and SVR with RBF kernel) performs better than linear models (linear regression and linear SVR). GPR shows significant superiority than other regressors.

2) Each cascade regression model has an around 5% superior performance the corresponding single-step regression model. Furthermore, the error of cascade linear model is smaller than single-step RBF-based SVR models, which indicates that a linear model embedded in a cascade manner could also achieve an effective non-linear solution.

3) Ensemble of four cascade regressors could improve the performance over each base regressor. On the contrary, the combination of four single-step regression models has a higher error than GPR. This is because GPR is quite outstanding over the other



regressors. Ensemble learning can boost the performance only when the base learners have similar generalization abilities.

This result demonstrates the statement in Section 5.1 that combining multiple heuristics is a better solution to facial spatial alignment rather than using a single heuristic. In this experiment, the best result is obtained by the ensemble of four cascade regression models and this approach will be used to facilitate the next step of texture fitting.

### 5.4.3 Static FER on SFEW

The proposed method is used to solve dynamic FER problems. It is also appropriate to be applied for static FER. In this experiment, we firstly perform cascade ensemble regression-based method on each input image to achieve the normalized face. Then, Local Binary Pattern (LBP) and Support Vector Machine (SVM) are used for feature extraction and emotion classification, respectively. The detailed introduction of LBP and SVM can be referred to Section 3.3.3 and Section 3.3.4.

Statistical Facial Expression in the Wild (SFEW) (Liu *et al.*, 2014a) is a static spontaneous facial expression database which contains 700 images captured from movies labelled by seven categories: six universal emotions and neutral. There is a standard evaluation protocol provided by the authors of SFEW. The evaluation is strictly person-independent. In Table 5.2, the methods of (Thies *et al.*, 2016) and (Jeni *et al.*, 2015) are the state-of-the-art approaches. It is clear that our method outperforms the others. The overall recognition rate of the proposed method is 10% higher than the others, which suggests a considerable improvement. Based on this result we can conclude that the proposed method is effective for static FER in the wild. In chapter 4, FEAF is specifically proposed for static FER task. So FEAF outperforms this method.

In this comparison, we did not mention deep learn because our work focus on small sample learning task which is quite different from deep learning. Meanwhile, deep learning methods must use a large volume of external training data, which do not totally comply with the evaluation protocol of SFEW.

### 5.4.4 Dynamic FER on AFEW

In this section, the dynamic FER experiment and evaluation are presented. The experimental process still start from the proposed cascade ensemble regression-based

Table 5.2: Comparison of recognition rate (%) of the state-of-the-art methods on SFEW database

	Angry	Disgust	Fear	Happy	Neutral	Sadness	Surprise	Overall
Baseline	23.00	13.00	13.90	29.00	23.00	17.00	13.50	18.90
(Shan <i>et al.</i> , 2009)	25.89	<b>28.24</b>	17.17	42.98	14.00	33.33	10.99	24.70
(Jeni <i>et al.</i> , 2015)	24.11	14.12	20.20	50.00	23.00	23.23	21.98	26.14
Proposed	<b>40.18</b>	25.88	<b>48.48</b>	<b>55.26</b>	<b>37.00</b>	<b>36.36</b>	<b>37.36</b>	<b>40.71</b>

Table 5.3: Comparison of recognition rate (%) of the state-of-the-art methods on AFEW database

	Angry	Disgust	Fear	Happy	Neutral	Sadness	Surprise	Overall
Baseline(Dhall <i>et al.</i> , 2014)	50.00	25.00	15.21	57.14	34.92	16.39	21.73	33.15
(Liu <i>et al.</i> , 2014b)	<b>84.75</b>	17.95	27.27	<b>82.54</b>	<b>70.49</b>	22.03	6.52	<b>48.52</b>
(Guo <i>et al.</i> , 2016)	-	-	-	-	-	-	-	48.3
Proposed	65.63	<b>27.50</b>	<b>32.61</b>	74.07	57.14	<b>32.79</b>	<b>36.96</b>	48.40

method to normalize each facial image into frontal facing view. Then for the FER methodology, we apply LBP-TOP and SVM on the reconstructed facial image sequences for feature extraction and emotion classification, respectively. In Chapter 3, we have shown that ROI-based part registration may cause information loss. The proposed method could well solve the problem of spatial alignment. So, there is no need to apply ROI-based part registration and we directly use LBP-TOP to extract dynamic features on the whole reconstructed faces. The detailed introduction of LBP-TOP can be referred to Section 3.3.3.

In order to evaluate the performance on video sequences, we use another database: Acted Facial Expressions In The Wild (AFEW) (Dhall *et al.*, 2012). The AFEW is an unconstrained facial expression database whose video clips are collected from movies. It contains 1368 video clips which are divided into three parts: 578 for training, 383 for validation and 407 for testing. Considering that the groundtruth of testing images is still unreleased, we follow the evaluation protocol of Emotion Recognition in the Wild Challenge 2014 (EmotiW 2014) (Dhall *et al.*, 2014) but only compare the performance on validation data.

In Table 5.3, the baseline result is achieved by the database creators who used traditional LBP-TOP + SVM strategy. The winner of EmotiW 2014 competition is (Liu *et al.*, 2014b) who used both audio and video signals, and combined SIFT, HOG and DCNN with external training data. Due to the very large amount of external training examples, (Liu *et al.*, 2014b) still keeps the record of EmotiW 2014 competition. The algorithm of (Guo *et al.*, 2016) is a typical representation of time alignment which aims to model the variations of time extent. There is no more external training data used in this method. Our method achieves a comparable result, especially the recognition rate of sad and surprise have over 10% superior than the other methods.

The result of ours and (Guo *et al.*, 2016) are both based on reasonable heuristics and no external training data is used. They both achieved high accuracy which is only 0.2% inferior than the winners (Liu *et al.*, 2014b) who introduced large amount of external training data. By observing this fact, we can see that a reasonable heuristic is also important for a learning task although downloading more data may improve the performance.

We finally compare the two traditional methods between ours and (Guo *et al.*, 2016). Our method applies spatial alignment without temporal alignment while (Guo

*et al.*, 2016) goes opposite. Our method has a minor superiority than (Guo *et al.*, 2016), which demonstrates that both temporal alignment and spatial alignment are important to dynamic analysis of facial expressions. Although many researchers focus on modelling temporal relations, the importance of spatial relations problems caused by head-pose and occlusions still remain.

## 5.5 Summary

In this paper, we have presented a novel cascade regression-based approach for accurate frontal facial shape estimation method applied it to dynamic FER in the wild. It successfully fills the gap that there is no dynamic FER approaches for spatial alignment. To this point, FEAF has been proved to be effective on both static and dynamic FER tasks.

# Chapter 6

## Conclusions

### 6.1 Overview

This thesis aims to improve the recognition performance of facial expression analysis in the wild. We start from the literature review of the state-of-the-art FER approaches, and then present the main challenges and gaps in this research field. Correspondingly, we propose new methods to close these gaps.

Firstly, a patch-based method is developed to deal with identity bias problem and shows good results on person-independent dynamic FER. Secondly, a novel facial expression-aware face frontalization method is proposed, which is the first approach concerning all the challenges of FER in the wild. Then, Chapter 5 goes back to the inherent of facial expressions by further extending FEAF to fit in with dynamic FER. These two models of FEAF has two main contributions: 1) all the challenges of FER in the wild is considered and processed, which close the gap mentioned in Chapter 1; 2) FEAF dramatically improves the performance of FER in the wild and it outperforms the state-of-the-art approaches, especially when there are finite training examples. Finally, an FEAF-based deep learning framework is developed for interpersonal relation prediction, which demonstrates that FEAF is a flexible platform that can be applied to various applications of facial expression analysis.

## 6.2 Contributions

In this thesis, we propose novel methods for facial expression analysis in the wild. By reviewing the state-of-the-art approaches, we summarized that the performance of FER approaches are subject to five non-linear factors: identity bias, head pose variations, occlusions, illumination changes and irregularity of spontaneous expressions. Accordingly, the proposed methods are targeting to deal with these problems.

### 6.2.1 Patch-based Person-independent FER

In Chapter 3, facial salient region detection and spatio-temporal feature extraction are combined to reduce the influence of identity bias. Firstly, we detect point-based facial landmark by means of SDM which separately detects facial fiducial points in the first frame and tracks them in the following frames. Then, we extract local patches according to fiducial points. This extraction method has two main advantages: (a) the effect of identity bias can be better mitigate since the regions around fiducial points preserve more expression-related cues, and (b) within all the frames in a sequence, the location of subjects (e.g. eyes, nose) are more stable and facial texture movements are more smooth. In each patch of sequence, block-based approach is exploited where LBP-TOP features are extracted in each block and connected to represent facial motions. Finally, we perform SVM classifier for emotion classification. The main contributions of this work include: (1) propose a novel method for emotion-enhanced feature extraction, (2) integrate the most effective and latest methods, such as SDM and LBP-TOP, for facial registration and facial representation. (3) the experimental results shows a good performance on person-independent facial expression recognition.

### 6.2.2 Static Model of Facial Expression-Aware Face Frontalization

A novel facial expression-aware face frontalization architecture is proposed for static FER. This approach includes three main steps: multi-template design, template matching and texture fitting. The contributions of this work are summarized as follow: 1) We propose a novel facial expression-aware face frontalization method which reconstructs frontal faces with detailed facial expression cues from unconstrained facial images.

This is the first work in its kind that fully considers facial expression recovery for face frontalization; 2) This is the first work that is able to process all the five non-linear factors; 3) For the application of interpersonal relation estimation, FEAF is further improved and combines with deep learning structure. The main contributions of this work is that a) FEAF is proved to be a flexible platform for various applications of facial expression analysis, and b) FEAF can well collaborate with deep neural network and is also effective for large-scale facial image processing.

### 6.2.3 Dynamic Model of Facial Expression-Aware Face Frontalization

In Chapter 5, we presents a novel face frontalization method for dynamic facial expression analysis. In this method, we firstly collect facial images in a pair of a non-frontal face and its corresponding frontal image, and the pair-wise relation between non-frontal face-shape and frontal counterpart will be learned through a regression model. Considered such relation is highly non-linear, a sequentially cascade manner is proposed to iteratively fulfill this task. The obtained cascade regression models are combined to make a comprehensive decision and the output will be used as frontal face-shape predictor which transforms the non-frontal face-shape to its frontal view. Finally, the estimated frontal face-shape can be seen as a base mesh and Active Appearance Model instantiation is employed to reconstruct facial appearances. The contributions include: 1) We propose a cascade regression-based framework for face frontalization. The whole process is in 2D without using any 3D information; 2) Different from the static face frontalization methods that loses deformations of expressions, the proposed method can track vivid expression changes; 3) Compare with static FEAF model, dynamic variant is more professional at dealing with identity bias problem.

## 6.3 Future Work

In the following section, the limitations of this thesis and some directions for future work are discussed.



### 6.3.1 Facial Expression Applications

In this thesis, the proposed system and methods have achieved the state-of-the-art performance in recognition of 6 basic emotions and interpersonal relation traits. According to Chapter 2, the future research is expected to extend FEAF to several other facial expression analysis tasks, such as compound FER or valence & arousal detection. Furthermore, we will also consider the applications in pain detection, intelligent transportation system and intelligent tutorial system.

### 6.3.2 Facial Expression synthesis

FEAF can be seen as a branch of realistic facial synthesis. As deep learning requires extremely large data collections for training, many works have been done on synthesizing realistic images instead of manually collecting them. We have discussed that spontaneous expressions are so irregular that it is difficult to model them. Therefore, current approaches for facial expression synthesis can only generate expressions in several uniformed patterns. Spontaneous facial expression generation is expected to be an interesting and challenging topic in the future.

According to Chapter 1, most face generation methods are based on GAN or 3DMM model fitting. Both of the two types have limitations. Model-based methods often suffers from the problem of high computational cost. The problem of GAN is that the variance of its models is usually quite large, which may generate unrealistic, or even wired faces (e.g. a woman with beard).

FEAF is a model-based method. We plan to combine FEAF with GAN in the future. FEAF model can be used as an constrain condition to balance the high variance of GAN. Meanwhile, we expect to solve the problem of high computational cost by integrating supervised GAN training paradigm instead of the unsupervised FEAF model fitting. Beside facial expression images, we also consider introduce more constrain conditions during the GAN training process. For example, facial attributes dataset can be used for jointly synthesizing facial expressions and attributes, so that GAN model may learn a rule from the dataset that women and beard are mutually exclusive.



## References

- AHONEN, T., HADID, A. & PIETIKAINEN, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2037–2041. 21, 33, 37
- ALABORT-I MEDINA, J. & ZAFEIRIOU, S. (2015). Unifying holistic and parts-based deformable model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3679–3688. 27
- ANEJA, D., COLBURN, A., FAIGIN, G., SHAPIRO, L. & MONES, B. (2016). Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, 136–153, Springer. 23, 85
- ASHRAF, A.B., LUCEY, S. & CHEN, T. (2008). Learning patch correspondences for improved viewpoint invariant face recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8, IEEE. 21
- ASTHANA, A., LUCEY, S. & GOECKE, R. (2011). Regression based automatic face annotation for deformable model building. *Pattern Recognition*, **44**, 2598–2613. 87
- ASTHANA, A., ZAFEIRIOU, S., CHENG, S. & PANTIC, M. (2014). Incremental face alignment in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1859–1866. 27
- BALTRUŠAITIS, T., ROBINSON, P. & MORENCY, L.P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–10, IEEE. 95

## REFERENCES

---

- BARRETT, L.F. (2006). Are emotions natural kinds? *Perspectives on psychological science*, **1**, 28–58. 16
- BENITEZ-QUIROZ, C.F., SRINIVASAN, R., FENG, Q., WANG, Y. & MARTINEZ, A.M. (2017). Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*. 16
- BLANZ, V. & VETTER, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194, ACM Press/Addison-Wesley Publishing Co. 27, 28
- BLANZ, V. & VETTER, T. (2003). Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, **25**, 1063–1074. 27, 28
- CELANI, G., BATTACCHI, M.W. & ARCIDIACONO, L. (1999). The understanding of the emotional meaning of facial expressions in people with autism. *Journal of autism and developmental disorders*, **29**, 57–66. 24
- CHEW, S.W., LUCEY, P.J., LUCEY, S., SARAGIH, J., COHN, J. & SRIDHARAN, S. (2011). Person-independent facial expression detection using constrained local models. *Proceedings of FG 2011 Facial Expression Recognition and Analysis Challenge*, 915–920. 9
- COLE, F., BELANGER, D., KRISHNAN, D., SARNA, A., MOSSERI, I. & FREEMAN, W.T. (2017). Synthesizing normalized faces from facial identity features. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 3386–3395, IEEE. 26
- COOTES, T.F., TAYLOR, C.J., COOPER, D.H. & GRAHAM, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, **61**, 38–59. 67, 68
- DALAL, N. & TRIGGS, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, 886–893, IEEE. 36, 76

## REFERENCES

---

- DARWIN, C. & PRODGER, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA. 15
- DHALL, A., GOECKE, R., LUCEY, S. & GEDEON, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2106–2112, IEEE. 24, 72, 76
- DHALL, A., GOECKE, R., LUCEY, S., GEDEON, T. *et al.* (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, **19**, 34–41. 101
- DHALL, A., GOECKE, R., JOSHI, J., SIKKA, K. & GEDEON, T. (2014). Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, 461–466, ACM. 100, 101
- DING, L., DING, X. & FANG, C. (2012). Continuous pose normalization for pose-robust face recognition. *IEEE Signal Processing Letters*, **19**, 721–724. 27
- DU, S., TAO, Y. & MARTINEZ, A.M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 201322355. 2, 16
- DUREHA, A. (2014). An accurate algorithm for generating a music playlist based on facial expressions. *International Journal of Computer Applications*, **100**, 33–39. 25, 85
- EKMAN, P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**, 169–200. 16
- EKMAN, P. (2002). Facial action coding system (facs). *A human face*. 16
- EKMAN, P. & FRIESEN, W. (2011). Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists, San Francisco*. 47
- EKMAN, P. & FRIESEN, W.V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, **17**, 124. 16

## REFERENCES

---

- ELEFThERiADiS, S., RUDOVIC, O. & PANTIC, M. (2015). Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, **24**, 189–204. 7, 50, 76, 77, 83
- ELEFThERiADiS, S., RUDOVIC, O. & PANTIC, M. (2016). Joint facial action unit detection and feature fusion: A multi-conditional learning approach. *IEEE transactions on image processing*, **25**, 5727–5742. 47
- FERRARI, C., LISANTI, G., BERRETTI, S. & DEL BIMBO, A. (2016). Effective 3d based frontalization for unconstrained face recognition. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 1047–1052, IEEE. 86, 90
- GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIr, S., COURVILLE, A. & BENGIO, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680. 28
- GRAY, H.M. & TICKLE-DEGNEN, L. (2010). A meta-analysis of performance on emotion recognition tasks in parkinson?s disease. *Neuropsychology*, **24**, 176. 24
- GU, W., XIANG, C., VENKATESH, Y., HUANG, D. & LIN, H. (2012). Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *Pattern recognition*, **45**, 80–91. 9
- GUO, Y., ZHAO, G. & PIETIKÄINEN, M. (2016). Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Transactions on Image Processing*, **25**, 1977–1992. 86, 87, 89, 100, 101, 102
- HASSNER, T., HAREL, S., PAZ, E. & ENBAR, R. (2015). Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4295–4304. 10, 27, 48, 74
- HEISELE, B., HO, P. & POGGIO, T. (2001). Face recognition with support vector machines: Global versus component-based approach. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, 688–694, IEEE. 62

## REFERENCES

---

- HESSE, N., GEHRIG, T., GAO, H. & EKENEL, H.K. (2012). Multi-view facial expression recognition using local appearance features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, 3533–3536, IEEE. 7, 50
- HUANG, D., SHAN, C., ARDABILIAN, M., WANG, Y. & CHEN, L. (2011). Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **41**, 765–781. 39, 76
- ICHIM, A.E., BOUAZIZ, S. & PAULY, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, **34**, 45. 23, 85
- JACK, R.E., GARROD, O.G., YU, H., CALDARA, R. & SCHYNS, P.G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, **109**, 7241–7244. 16
- JENI, L.A., GIRARD, J.M., COHN, J.F. & DE LA TORRE, F. (2013). Continuous au intensity estimation using localized, sparse facial feature space. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, 1–7, IEEE. 8
- JENI, L.A., COHN, J.F. & KANADE, T. (2015). Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, 1–8, IEEE. 24, 51, 90, 98, 99
- JIANG, B., VALSTAR, M.F. & PANTIC, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 314–321, IEEE. 37
- JIANG, B., VALSTAR, M.F., MARTINEZ, B. & PANTIC, M. (2014). A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Trans. Cybernetics*, **44**, 161–174. 9, 20, 22, 36, 38, 86, 88

## REFERENCES

---

- KAN, M., SHAN, S., CHANG, H. & CHEN, X. (2014). Stacked progressive auto-encoders (spae) for face recognition across poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1883–1890. 26
- KAZEMI, V. & SULLIVAN, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874. 27
- KIESLER, D.J. (1983). The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological review*, **90**, 185. 71
- KIM, B.K., LEE, H., ROH, J. & LEE, S.Y. (2015). Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 427–434, ACM. 5, 50
- KOELSTRA, S., PANTIC, M. & PATRAS, I. (2010). A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, **32**, 1940–1954. 9, 36, 47, 69
- KUMANO, S., OTSUKA, K., YAMATO, J., MAEDA, E. & SATO, Y. (2009). Pose-invariant facial expression recognition using variable-intensity templates. *International journal of computer vision*, **83**, 178–194. 7
- LEVI, G. & HASSNER, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 503–510, ACM. 28
- LI, S., LIU, X., CHAI, X., ZHANG, H., LAO, S. & SHAN, S. (2012). Morphable displacement field based image matching for face recognition across pose. In *European conference on computer vision*, 102–115, Springer. 27
- LI, S., DENG, W. & DU, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2584–2593, IEEE. 16
- LISETTI, C. (1998). Affective computing. 16



## REFERENCES

---

- LIU, M., LI, S., SHAN, S. & CHEN, X. (2013). Au-aware deep networks for facial expression recognition. In *FG*, 1–6. 47, 76, 83
- LIU, M., SHAN, S., WANG, R. & CHEN, X. (2014a). Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1749–1756. 89, 98
- LIU, M., WANG, R., LI, S., SHAN, S., HUANG, Z. & CHEN, X. (2014b). Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, 494–501, ACM. 100, 101
- LUCEY, P., COHN, J.F., KANADE, T., SARAGIH, J., AMBADAR, Z. & MATTHEWS, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 94–101, IEEE. 41
- LUCEY, P., COHN, J.F., MATTHEWS, I., LUCEY, S., SRIDHARAN, S., HOWLETT, J. & PRKACHIN, K.M. (2011a). Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **41**, 664–674. 2
- LUCEY, P., COHN, J.F., PRKACHIN, K.M., SOLOMON, P.E. & MATTHEWS, I. (2011b). Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Face and Gesture 2011*, 57–64, IEEE. 2
- MANDAL, M.K., PANDEY, R. & PRASAD, A.B. (1998). Facial expressions of emotions and schizophrenia: a review. *Schizophrenia bulletin*, **24**, 399–412. 24
- MASI, I., TR?N, A.T., HASSNER, T., LEKSUT, J.T. & MEDIONI, G. (2016). Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, 579–596, Springer. 25, 28, 29, 30
- MATTHEWS, I. & BAKER, S. (2004). Active appearance models revisited. *International journal of computer vision*, **60**, 135–164. 64, 65, 88, 90

## REFERENCES

---

- MATTIVI, R. & SHAO, L. (2009). Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *International Conference on Computer Analysis of Images and Patterns*, 740–747, Springer. 20, 21
- MEHRABIAN, A. & FERRIS, S.R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, **31**, 248. 1
- MEHRABIAN, A. & WIENER, M. (1967). Decoding of inconsistent communications. *Journal of personality and social psychology*, **6**, 109. 1
- MOHAMMADI, M., FATEMIZADEH, E. & MAHOOR, M.H. (2014). Pca-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, **25**, 1082–1092. 9
- MOLLAHOSSEINI, A., CHAN, D. & MAHOOR, M.H. (2016). Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–10, IEEE. 29, 50, 83
- MOLLAHOSSEINI, A., HASANI, B. & MAHOOR, M.H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*. 17
- MOORE, S. & BOWDEN, R. (2011). Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, **115**, 541–558. 6, 50
- NICOLLE, J., RAPP, V., BAILLY, K., PREVOST, L. & CHETOUANI, M. (2012). Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 501–508, ACM. 8
- OJALA, T., PIETIKAINEN, M. & MAENPAA, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, **24**, 971–987. 36, 40
- OJANSIVU, V. & HEIKKILÄ, J. (2008). Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, 236–243, Springer. 36

## REFERENCES

---

- PARKHI, O.M., VEDALDI, A., ZISSERMAN, A. *et al.* (2015). Deep face recognition. In *BMVC*, 6. 28, 51
- PENG, H., LONG, F. & DING, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, **27**, 1226–1238. 21
- PEREIRA, L., OBARA, K., DIAS, J., MENACHO, M., LAVADO, E. & CARDOSO, J. (2011). Facial exercise therapy for facial palsy: systematic review and meta-analysis. *Clinical rehabilitation*, **25**, 649–658. 24
- PFISTER, T., LI, X., ZHAO, G. & PIETIKÄINEN, M. (2011). Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1449–1456, IEEE. 25, 85
- PLUTCHIK, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion*, **1984**, 197–219. 17
- PRINCE, S.J., ELDER, J.H., WARRELL, J. & FELISBERTI, F.M. (2008). Tied factor analysis for face recognition across large pose differences. *IEEE Transactions on pattern analysis and machine intelligence*, **30**, 970–984. 21
- REN, J., JIANG, X. & YUAN, J. (2015). Learning lbp structure by maximizing the conditional mutual information. *Pattern Recognition*, **48**, 3180–3190. 21, 22
- REN, S., CAO, X., WEI, Y. & SUN, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1685–1692. 27, 47
- ROTH, J., TONG, Y. & LIU, X. (2015). Unconstrained 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2606–2615. 24, 51, 90
- RUDOVIC, O., PANTIC, M. & PATRAS, I. (2013). Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE transactions on pattern analysis and machine intelligence*, **35**, 1357–1369. 7, 36, 50

## REFERENCES

---

- RUDOVIC, O., PAVLOVIC, V. & PANTIC, M. (2015). Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE transactions on pattern analysis and machine intelligence*, **37**, 944–958. 47
- RUECKERT, D., SONODA, L.I., HAYES, C., HILL, D.L., LEACH, M.O. & HAWKES, D.J. (1999). Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, **18**, 712–721. 89
- RUSSELL, J.A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, **39**, 1161. 17
- SAGONAS, C., PANAGAKIS, Y., ZAFEIRIOU, S. & PANTIC, M. (2015). Robust statistical face frontalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 3871–3879. 10, 27, 48, 49, 74
- SARIYANIDI, E., GUNES, H. & CAVALLARO, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, **37**, 1113–1133. 8, 20, 21, 22, 30, 34, 36, 37, 88
- SCHROFF, F., KALENICHENKO, D. & PHILBIN, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823. 28
- SENGUPTA, S., CHEN, J.C., CASTILLO, C., PATEL, V.M., CHELLAPPA, R. & JACOBS, D.W. (2016). Frontal to profile face verification in the wild. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–9, IEEE. 26
- SHAN, C., GONG, S. & MCOWAN, P.W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, **27**, 803–816. 8, 9, 20, 33, 37, 42, 43, 48, 90, 99
- SHOJAEILANGARI, S., YUN, Y.W. & KHWANG, T.E. (2011). Person independent facial expression analysis using gabor features and genetic algorithm. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, 1–5, IEEE. 9

## REFERENCES

---

- SIMONYAN, K. & ZISSERMAN, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 51
- SMYTH, P. & GOODMAN, R.M. (1992). An information theoretic approach to rule induction from databases. *IEEE transactions on Knowledge and data engineering*, **4**, 301–316. 59
- SZEGEDY, C., LIU, W., JIA, Y., Sermanet, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V. & RABINOVICH, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9. 51
- TAHERI, S., QIU, Q. & CHELLAPPA, R. (2014). Structure-preserving sparse decomposition for facial expression analysis. *IEEE Transactions on Image Processing*, **23**, 3590–3603. 19, 37, 38, 47
- TAIGMAN, Y., YANG, M., RANZATO, M. & WOLF, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1708. 28, 51
- TANG, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*. 5, 50
- TARIQ, U., YANG, J. & HUANG, T.S. (2014). Supervised super-vector encoding for facial expression recognition. *Pattern Recognition Letters*, **46**, 89–95. 20, 22
- TEKGUC, U., SOYEL, H. & DEMIREL, H. (2009). Feature selection for person-independent 3d facial expression recognition using nsga-ii. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, 35–38, IEEE. 9
- THIES, J., ZOLLHOFER, M., STAMMINGER, M., THEOBALT, C. & NIESSNER, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2387–2395. 23, 85, 98

## REFERENCES

---

- TONG, Y., CHEN, J. & JI, Q. (2010). A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE transactions on pattern analysis and machine intelligence*, **32**, 258–273. 36
- TRAN, L., YIN, X. & LIU, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 7. 29, 90
- TRIGEORGIS, G., SNAPE, P., NICOLAOU, M.A., ANTONAKOS, E. & ZAFEIRIOU, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4177–4187. 27
- TZIMIROPOULOS, G. & PANTIC, M. (2014). Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1851–1858. 27
- TZIMIROPOULOS, G. & PANTIC, M. (2017). Fast algorithms for fitting active appearance models to unconstrained images. *International journal of computer vision*, **122**, 17–33. 27
- UŘIČÁŘ, M., FRANC, V., THOMAS, D., SUGIMOTO, A. & HLAVÁČ, V. (2015). Real-time multi-view facial landmark detector learned by the structured output svm. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 2, 1–8, IEEE. 27
- VALSTAR, M.F. & PANTIC, M. (2012). Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **42**, 28–43. 9, 36
- VALSTAR, M.F., JIANG, B., MEHU, M., PANTIC, M. & SCHERER, K. (2011). The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 921–926, IEEE. 20, 23, 37, 38
- WALECKI, R., RUDOVIC, O., PAVLOVIC, V. & PANTIC, M. (2015). Variable-state latent conditional random fields for facial expression recognition and action unit

## REFERENCES

---

- detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, 1–8, IEEE. 47
- WANG, S., WANG, J., WANG, Z. & JI, Q. (2015). Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions. *IEEE Transactions on Multimedia*, **17**, 2185–2197. 24, 85
- WANG, Y., YU, H., DONG, J., STEVENS, B. & LIU, H. (2016). Facial expression-aware face frontalization. In *Asian Conference on Computer Vision*, 375–388, Springer. 29, 59, 90
- WANG, Y., YU, H., DONG, J., JIAN, M. & LIU, H. (2017). Cascade support vector regression-based facial expression-aware face frontalization. In *Image Processing (ICIP), 2017 IEEE International Conference on*, 2831–2835, IEEE. 29
- WANG, Z., WANG, S. & JI, Q. (2013). Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3422–3429. 22, 36, 42, 43, 89
- WENG, C.H., LAI, Y.H. & LAI, S.H. (2016). Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision*, 117–133, Springer. 2
- WU, Y., LIU, W. & WANG, J. (2008). Application of emotional recognition in intelligent tutoring system. In *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, 449–452, IEEE. 25, 85
- XIONG, X. & DE LA TORRE, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 532–539. 27, 34, 38, 47, 59, 87
- XUE, M., LIU, W. & LI, L. (2013). Person-independent facial expression recognition via hierarchical classification. In *Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on*, 449–454, IEEE. 8, 9, 48, 90

## REFERENCES

---

- YANG, Y., SHEN, H.T., MA, Z., HUANG, Z. & ZHOU, X. (2011). l<sub>2,1</sub>-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI proceedings-international joint conference on artificial intelligence*, 1589. 20, 21
- YANG, Y., MA, Z., HAUPTMANN, A.G. & SEBE, N. (2013). Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia*, **15**, 661–669. 21, 22
- YAO, A., SHAO, J., MA, N. & CHEN, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 451–458, ACM. 5
- YEASIN, M., BULLOT, B. & SHARMA, R. (2006). Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, **8**, 500–508. 9
- YIM, J., JUNG, H., YOO, B., CHOI, C., PARK, D. & KIM, J. (2015). Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 676–684. 29
- YIN, L., WEI, X., SUN, Y., WANG, J. & ROSATO, M.J. (2006). A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, 211–216, IEEE. 2, 17
- YIN, L., SUN, X.C.Y., WORM, T. & REALE, M. (2008). A high-resolution 3d dynamic facial expression database. In *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, vol. 126. 2, 18
- YIN, X., YU, X., SOHN, K., LIU, X. & CHANDRAKER, M. (2017). Towards large-pose face frontalization in the wild. In *Proc. ICCV*, 1–10. 28
- ZAFEIRIOU, S., PAPAIOANNOU, A., KOTSIA, I., NICOLAOU, M. & ZHAO, G. (2016). Facial affect“in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 36–47. 2, 17



## REFERENCES

---

- ZENG, Z., PANTIC, M., ROISMAN, G.I. & HUANG, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, **31**, 39–58. 50
- ZHANG, L. & TJONDRONEGORO, D. (2011). Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, **2**, 219–229. 8
- ZHANG, Z., GANESH, A., LIANG, X. & MA, Y. (2012). Tilt: Transform invariant low-rank textures. *International journal of computer vision*, **99**, 1–24. 68
- ZHANG, Z., LUO, P., LOY, C.C. & TANG, X. (2015). Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, 3631–3639. 71
- ZHANG, Z., LUO, P., LOY, C.C. & TANG, X. (2018). From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, **126**, 550–569. 49, 71, 72, 77, 78, 84
- ZHAO, G. & PIETIKAINEN, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, **29**, 915–928. 8, 20, 21, 31, 33, 36, 37, 38, 39, 42, 43, 86, 88, 89
- ZHAO, G., WU, Y., CHEN, F., ZHANG, J. & BAI, J. (2015). Effective feature selection using feature vector graph for classification. *Neurocomputing*, **151**, 376–389. 21, 22
- ZHOU, E., CAO, Z. & YIN, Q. (2015). Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*. 28
- ZHU, X. & RAMANAN, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2879–2886, IEEE. 47
- ZHU, X., LEI, Z., YAN, J., YI, D. & LI, S.Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 787–796. 27, 30, 87, 90

## REFERENCES

---

- ZHU, X., LEI, Z., LIU, X., SHI, H. & LI, S.Z. (2016). Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 146–155. 86, 90
- ZHU, Y., DE LA TORRE, F., COHN, J.F. & ZHANG, Y.J. (2011). Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *IEEE transactions on affective computing*, **2**, 79–91. 8
- ZHU, Z., LUO, P., WANG, X. & TANG, X. (2013). Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*, 113–120. 26
- ZHU, Z., LUO, P., WANG, X. & TANG, X. (2014). Multi-view perceptron: a deep model for learning face identity and view representations. In *Advances in Neural Information Processing Systems*, 217–225. 28

# Appendix A

## Publications

1. Wang Y., Yu H., Dong J., Stevens B., Liu H. Dynamic facial expression recognition using local patch and LBP-TOP. International Conference on Human System Interaction. 2015.
2. Wang Y., Yu H., Dong J., Stevens B., Liu H. Facial Expression-Aware Face Frontalization. Asian Conference on Computer Vision. 375-388, 2016.
3. Wang Y., Yu H., Dong J., Jian M., Liu H. Cascade support vector regression-based facial expression-aware face frontalization. IEEE International Conference on Image Processing. 2831-2835, 2017.

# **Appendix B**

## **Research Ethics**

# FORM UPR16

## Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information)



<b>Postgraduate Research Student (PGRS) Information</b>		<b>Student ID:</b>	up749434	
<b>PGRS Name:</b>	Yiming Wang			
<b>Department:</b>	School of Creative Technologies	<b>First Supervisor:</b>	Hui Yu	
<b>Start Date:</b> (or progression date for Prof Doc students)	01/10/2014			
<b>Study Mode and Route:</b>	Part-time <input type="checkbox"/>	MPhil <input type="checkbox"/>	MD <input type="checkbox"/>	
	Full-time <input checked="" type="checkbox"/>	PhD <input checked="" type="checkbox"/>	Professional Doctorate <input type="checkbox"/>	

<b>Title of Thesis:</b>	Face Frontalization for Facial Expression Recognition in the Wild
<b>Thesis Word Count:</b> (excluding ancillary data)	29,474

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

**UKRIO Finished Research Checklist:**  
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: <http://www.ukrio.org/what-we-do/code-of-practice-for-research/>)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES <input checked="" type="checkbox"/>	NO <input type="checkbox"/>
b) Have all contributions to knowledge been acknowledged?	YES <input checked="" type="checkbox"/>	NO <input type="checkbox"/>
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES <input checked="" type="checkbox"/>	NO <input type="checkbox"/>
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES <input checked="" type="checkbox"/>	NO <input type="checkbox"/>
e) Does your research comply with all legal, ethical, and contractual requirements?	YES <input checked="" type="checkbox"/>	NO <input type="checkbox"/>

**Candidate Statement:**

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

<b>Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):</b>	Completed in September 2015
---	-----------------------------

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

<b>Signed (PGRS):</b>		<b>Date:</b>	02/10/2018
-----------------------	--	--------------	------------

