ÉCOLE CENTRALE LYON

LIRIS

THESE

pour obtenir le grade de

DOCTEUR DE L'ECOLE CENTRALE DE LYON

Spécialité: Informatique

Combining 2D facial texture and 3D face morphology for estimating people's soft biometrics and recognizing facial expressions

---

dans le cadre de l'Ecole Doctorale InfoMaths

présentée et soutenue publiquement par

Huaxiong DING

December 2016

Directeur de thèse: Prof. Jean-Marie Morvan

Co-directeur de thèse: Prof. Liming Chen

JURY

Prof. Alice CAPLIER	Grenoble INP	Rapporteur
Prof. Boulbaba BEN AMOR	TELECOM Mine Lille 1	Rapporteur
Prof. Jean-Luc DUGELAY	Eurecom	Examinateur
DR1. Frédéric CAZALS	Université Nice Sophia Antipolis	Examinateur
Prof. Jean-Marie MORVAN	Université Lyon 1 / KAUST	Directeur de thèse
Prof. Liming CHEN	Ecole Centrale de Lyon	Co-directeur de thèse

Contents

List of Tables

List of Figures

Abstract

Since soft biometrics traits can provide sufficient evidence to precisely determine the identity of human, there has been increasing attention for face based soft biometrics identification in recent years. Among those face based soft biometrics, gender and ethnicity are both key demographic attributes of human beings and they play a very fundamental and important role in automatic machine based face analysis. Meanwhile, facial expression recognition is another challenge problem in face analysis because of the diversity and hybridity of human expressions among different subjects in different cultures, genders and contexts.

This Ph.D thesis work is dedicated to combine 2D facial Texture and 3D face morphology for estimating people's soft biometrics: gender, ethnicity, etc., and recognizing facial expression.

For the gender and ethnicity recognition, we present an effective and efficient approach on this issue by combining both boosted local texture and shape features extracted from 3D face models, in contrast to the existing ones that only depend on either 2D texture or 3D shape of faces. In order to comprehensively represent the difference between different genders or ethnics groups, we propose a novel local descriptor, namely local circular patterns (LCP). LCP improves the widely utilized local binary patterns (LBP) and its variants by replacing the binary quantization with a clustering based one, resulting in higher discriminative power as well as better robustness to noise. Meanwhile, the following Adaboost based feature selection finds the most discriminative gender- and ethnic-related features and assigns them with different weights to highlight their importance in classification, which not only further raises the performance but reduces the time and memory cost as well. Experimental results achieved on the FRGC v2.0 and BU-3DFE data sets clearly demonstrate the advantages of the proposed method.

For facial expression recognition, we present a fully automatic multi-modal 2D +

3D feature-based facial expression recognition approach and demonstrate its performance on the BU–3DFE database. Our approach combines multi-order gradient-based local texture and shape descriptors in order to achieve efficiency a nd robustness. First, a large set of fiducial facial landmarks of 2D face images along with their 3D face scans are localized using a novel algorithm namely incremental Parallel Cascade of Linear Regression (iPar–CLR). Then, a novel Histogram of Second Order Gradients (HSOG) based local image descriptor in conjunction with the widely used first-order gradient based SIFT descriptor are employed to describe the local texture around each 2D landmark. Similarly, the local geometry around each 3D landmark is described by two novel local shape descriptors constructed using the first-order and the second-order surface differential geometry quantities, i.e., Histogram of mesh Gradients (meshHOG) and Histogram of mesh Shape index (curvature quantization, meshHOS). Finally, the Support Vector Machine (SVM) based recognition results of all 2D and 3D descriptors are fused at both feature-level and score-level to further improve the accuracy. Comprehensive experimental results demonstrate that there exist impressive complementary characteristics between the 2D and 3D descriptors. We use the BU–3DFE benchmark to compare our approach to the state-of-the-art ones. Our multi-modal feature-based approach outperforms the others by achieving an average recognition accuracy of 86.32%. Moreover, a good generalization ability is shown on the Bosphorus database.

Keywords: facial soft-biometrics recognition, facial expressions recognition, multi-modal classification

Résumé

Puisque les traits de biométrie douce peuvent fournir des preuves supplémentaires pour aider à déterminer précisément l'identité de l'homme, il y a eu une attention croissante sur la reconnaissance faciale basée sur les biométrie douce ces dernières années. Parmi tous les biométries douces, le sexe et l'ethnicité sont les deux caractéristiques démographiques importantes pour les êtres humains et ils jouent un rôle très fondamental dans l'analyse de visage automatique. En attendant, la reconnaissance des expressions faciales est un autre challenge dans le domaine de l'analyse de visage en raison de la diversité et de l'hybridité des expressions humaines dans différentes cultures, genres et contextes.

Ce thèse est dédié à combiner la texture du visage 2D et la morphologie du visage 3D pour estimer les biométries douces: le sexe, l'ethnicité, etc., et reconnaître les expressions faciales.

Pour la reconnaissance du sexe et de l'ethnicité, nous présentons une approche efficace en combinant à la fois des textures locales et des caractéristiques de forme extraites à partir des modèles de visage 3D, contrairement aux méthodes existantes qui ne dépendent que des textures ou des caractéristiques de forme. Afin de souligne exhaustivement la différence entre les groupes sexuels et ethniques, nous proposons un nouveau descripteur, à savoir local circular patterns (LCP). Ce descripteur améliore Les motifs binaires locaux (LBP) et ses variantes en remplaçant la quantification binaire par une quantification basée sur le regroupement, entraînant d'une puissance plus discriminative et une meilleure résistance au bruit. En même temps, l'algorithme Adaboost est engagé à sélectionner les caractéristiques discriminatives fortement liés au sexe et à l'ethnicité. Les résultats expérimentaux obtenus sur les bases de données FRGC v2.0 et BU-3DFE démontrent clairement les avantages de la méthode proposée.

Pour la reconnaissance des expressions faciales, nous présentons une méthode

automatique basée sur les multi-modalité 2D + 3D et démontrons sa performance sur la base des données BU-3DFE. Notre méthode combine des textures locales et des descripteurs de formes pour atteindre l'efficacité et la robustesse. Tout d'abord, un grand ensemble des points des caractéristiques d'images 2D et de modèles 3D sont localisés à l'aide d'un nouvel algorithme, à savoir la cascade parallèle incrémentielle de régression linéaire (iPar-CLR). Ensuite, on utilise un nouveau descripteur basé sur les histogrammes des gradients d'ordre secondaire (HSOG) en conjonction avec le descripteur SIFT pour décrire la texture locale autour de chaque point de caractéristique 2D. De même, la géométrie locale autour de chaque point de caractéristique 3D est décrite par deux nouveaux descripteurs de forme construits à l'aide des quantités différentielle de géométries de la surface au premier ordre et au second ordre, à savoir meshHOG et meshHOS. Enfin, les résultats de reconnaissance des descripteurs 2D et 3D fournis par le classifier SVM sont fusionnés à la fois au niveau de fonctionnalité et de score pour améliorer la précision. Les expérimentaux résultats démontrent clairement qu'il existe des caractéristiques complémentaires entre les descripteurs 2D et 3D. Notre approche basée sur les multi-modalités surpasse les autres méthodes de l'état de l'art en obtenant une précision de reconnaissance $86,32\%$. De plus, une bonne capacité de généralisation est aussi présentée sur la base de données Bosphorus.

Mots clés: reconnaissance des biométries douces, reconnaissance des expressions faciales, classification sur les multi-modalités

Chapter 1

Introduction

1.1 Context

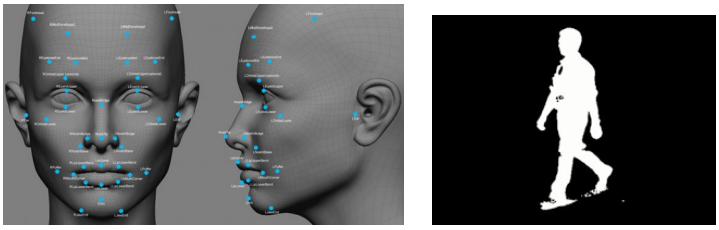
1.1.1 What is soft biometric?

Primary biometric systems have been widely used in many applications (e.g., identification, access control) in the past 20 years. Basically, biometric systems make use of physiological and behavioral characteristics to identify individuals, which are commonly known as fingerprint, face, iris, hand-geometry, etc., as shown in Fig.1.1. Unlike others techniques of access control (e.g, PIN code, password, tokens), biometric identifiers are unique to individuals which lead to an excellent reliability in service use. However, nothing is perfect in the real world. With the rapid and large-scale employment of biometric systems in public, such as business office and customs, more and more people are concerned about the robustness of those systems. The limitations of classic biometric systems are usually caused by the integrity of their identifier like noisy sensor data, non-universality, or spoof attacks.

Some techniques have been proposed to improve the performance of traditional biometric systems. One of those possible solutions is to use soft biometrics. The concept of soft biometric was initially proposed by Alphonse Bertillon in the 19th century. In the following years, more and more soft biometric traits have been proposed. The most famous examples of them are gender, ethnicity, age, height, weight, etc. In 2011, [Reid & Nixon 2011] redefined soft biometrics as any characteristic which can be naturally described by humans.

Since soft biometric traits can theoretically provide additional information about the identity of human, many primary biometric systems used them to improve the recognition accuracy. Thus, there has been increasing attention for soft biometric

(a) (b) (c)



(d) (e)

Figure 1.1: Various recognition systems based on different biometric traits. From (a) to (f), they are fingerprint, iris, voice, face, action, respectively.

identification in recent years.

1.1.2 Why facial soft biometric?

As said in [Michaeli 2013], *"We are all individuals with different physical and psychological make up. If the eyes are a window to the soul, how we perceive their faces is a key to understanding the souls of our fellow human beings"*, The human face plays a very fundamental and important role for human beings to understand each other in daily life. On a technical point of view, face biometrics are also effective to provide essential informations to identify individuals. Moreover, comparing to other popular biometric recognition approaches (e.g, fingerprint, iris, voice) in past few years, face based recognition approach has its own advantages:

1. **Natural**. In fact, like face biometric systems, human beings (even some animals) also distinguish and recognize individuals in communications by comparing the traits of faces, not the traits of fingerprint or iris.

2. **Non-invasive**. Face biometric systems can capture face images by visible lights, while fingerprint and iris recognition usually need professional senors,

which are much easier to be sensed. However, a non-invasive way at most times means a friendly way in biometric systems. The person will not have any feeling of invasion when it is being identified.

3. **Non-contact**. Unlike fingerprint or iris biometrics, face images can be captured from a distance without touching the person being identified. Meanwhile, the whole recognition procedure also doesn't need the interaction of person. This invisibility is really important in some uncontrolled environment.

4. **Easy to be accessed**. With a rapid development of photography equipment, the barrier to take a photo has become lower and lower. Now days, we can easily use our smart phone to take some photos with high qualities for face analysis.

However, Facial biometric systems show good future applications in many domains. They are not only widely deployed in traditional applications like video surveillance and access control, but also quickly applied in new domains: mobile payment, bank systems, criminal investigations, etc.

Up to now, there are a lot of outperforming and convincing results of facial recognition have been proposed in controlled environment. However, the application of face recognition in a real world is still restricted by several challenges, e.g, occlusion, pose, and expression. Inspired by the fact that soft biometric traits could strengthen primary biometric systems and facial biometric has rational advantages than other biometrics, more and more researchers pay attention to study approaches of recognizing soft biometric using information presented by faces.

Moreover, among those facial soft-biometrics, gender and ethnicity are both key demographic attributes of human beings and they play a very fundamental and important role in automatic machine based face analysis. As a pattern recognition issue, gender and ethnicity are normally considered being able to be categorized for the purpose of research. Gender can be categorized by biological differences, male and female, while categorization of different ethnic groups adopts the definitions applied to the 2000 census carried out by the federal government of the United

States [eth 8 31]. According to their statement, there have ethnic groups as follows: White, Asian, Black or African American, American Indian and Alaska Native, etc. Thus, gender and ethnicity recognition is a typical binary classification or multi-class classification issue in the field of pattern recognition.

However, in addition to provide extra information and criteria for systems of facial recognition, gender and ethnicity recognition have their own applications, i.e, providing the most appropriate services based on different cultures of gender and ethnic groups.

1.1.3 Why facial expression recognition?

"Words are only 7% *of the communication."* said the psychologist, Albert Mehrabian, in [Mehrabian 1972]. As a form of non-verbal communications, just like other gestures, facial expression plays a huge role in conveying social information between human beings. In the natural social communications, it is normally not hard for human to read expressions shown on faces. But how does a machine understand human's facial expressions? To answer this question from the perspective of non-verbal communications: facial expressions are the results caused by motions or positions of facial muscles. Inspired by those theories, [Ekman & Friesen 1978] proposed a scientific system, Facial Action Coding System (FACS), to measure facial behaviors. FACS is a powerful tool to describe nearly all possible observable components (named as Action Units (AUs)) of facial movement. In others words, from the perspective of FACS, every facial expression could be regarded as a consequence of a combination of action units. The effectiveness of FACS have been demonstrated over past few years. At the same time, [Ekman & Friesen 1978] also introduced six universal prototypical facial expressions: Neutral, Anger, Disgust, Fear, Happiness, Sadness, and Surprise along with neutral.

Facial expression recognition has a wide range of application, particularly in human-machine interaction, computer facial animation and the analysis of conversation structure. Meanwhile, a excellent technique of recognizing expressions is also able to provide additional information for enforcing primary facial recognition system.

1.1.4 Multi-modal facial recognition

From the perspective of sources, facial modalities can usually be divided into two main streams: 2D texture and 3D shape. Traditional 2D texture images are numerous, easy to be captured, and even with high qualities, while 3D shape models are on the opposite side: few, need expensive scanners, sometimes with poor qualities. But 3D shape model has his own advantages because it is more stable and robust to illumination variance, small head pose changes, and facial cosmetics. Theoretically, those two modalities are relatively complemented for face analysis. Thus, multi-modal face recognition based on coming 2D texture and 3D shape information has shown its effective power in past years.

1.2 Objectives and challenges

In the previous section, we have already introduced some related context in the field of face analysis. Among various aspects of face analysis, we mainly focus on two problems in this thesis:

1. Firstly, in order to make facial biometric system more robust, a flexible way is using traits of facial soft biometric. Thus, we want to study reliable approaches of recognizing facial soft biometrics. More details will be discussed in Section 1.2.1.

2. Secondly, considering facial expression recognition (FER) is a great challenge in face analysis and it also can enrich facial biometric system, our another task is studying FER problem and proposing our proper methods to recognize expressions. More details will be discussed in Section 1.2.2.

1.2.1 Facial Soft biometric recognition: gender and ethnicity

As mentioned previously, gender and ethnicity have attracted a number of researchers in part years due to its effectiveness for strengthen primary facial recognition systems. Many efforts have been made in the literature. Looking back to

historical development in this domain, we could roughly conclude challenges of gender and ethnicity classification in following aspects:

1. Gender and ethnicity recognition is fundamentally a pattern classification problem. How to find discriminative characteristics to distinguish gender and ethnic groups is the hottest research topic in this field. From a physiological point of view, the male and female faces differ in local details: male foreheads are usually wider than female foreheads, male have wide, longer noses, whereas female noses are more narrow and shorter, etc. Meanwhile, there are also some small differences among ethnic groups [Farkas *et al.* 2005]: the nose of East Asians was much broader than North American whites, the eyes were spaced farther apart in East Asians and some Middle Eastern populations, and the face was broader in East Asians. Now the question is how to design a flexible way by which machines could understand faces by taking into those physiological findings above. In the language of machine learning, that means we need find significantly effective feature descriptors to highlight differences for gender and ethnicity classification.

2. Considering the fact that each organ of faces, such as eye, nose, forehead, etc., provides different cues in facial gender and ethnicity classification, local feature based approaches generally tend to be more reasonable and thereby more efficient than the holistic ones. Thus, it leads to another challenge: which parts of face are the most discriminative gender-related or ethnic-related? Using features extracted from various facial regions highly related and discriminative to the task can also efficiently decrease the dimension of feature space.

In this thesis, our goal is to design facial recognition algorithms effective enough to distinguish people in gender and ethnic groups.

1.2.2 Facial expression recognition

In general, difficulties of facial expression recognition can be summarized as follows:

1. Similar as other recognition problem, to extract effective characteristic to distinguish different expressions is always the aims of researchers. Usually, there exist 2 ways to construct the feature descriptors for universal expressions. The premier one directly extracts discriminative clues from facial modalities as its feature descriptors. Another one firstly calculate feature vectors for recognizing related Action Units defined by FACS, and then, a fuzzy set of Action Units can be used to identify facial expressions. There are distinct gaps of performance between two ways above over past few years.

2. Actually, in the social communications, a facial appearance does not strictly correspond to an unique expression of emotions. It may be a result of combinations of diver expressions with different intensities. Thus, separating emotions from other facial expressions is a challenge while developing an automated FER system.

3. As opposed to facial identification, the diversity of individuals is a bad factor for FER. The ways of individuals in various cultures, genders or contexts to express their emotions are quite personal. Thus, how to separate respectively the traits caused by identity and expression is another interesting topic. And for FER, we need to design effective facial representations only highlighting the discriminative of expressions.

4. People prefer to express their emotional feelings along with some specific gestures or poses, which also adds significant complexity to algorithms. The influence of uncontrolled pose is sometimes terrible for traditional algorithms on 2D facial appearance. For 3D model, the variation of pose also could lead to missing data when it is being captured.

5. From the perspective of temporal dynamics, FER approaches can be categorized into static (still images) and dynamic (image sequences). Unlike static FER only extracts characteristics from single image, dynamic FER need additionally code facial deformations between continues frames. Thus, in general, the accuracy of prediction of dynamic FER is higher than that of static FER.

In this thesis, we focus on the problem of recognizing the six basic facial expressions using multimodal 2D + 3D static images

1.3 Our Contributions

These two main contributions of this PhD thesis involve multi-modal facial soft biometric recognition (gender and ethnicity) and multi-modal facial expression recognition. In the following, we summarize the main contributions of our thesis work.

a. Combining 2D facial texture and 3D facial shape for soft biometric estimation

To approach the issue previously mentioned, we first proposed a multi-modal facial biometric (gender and ethnicity) recognition system. In this work, two feature descriptors, which are able to accurately extract discriminative characteristics respectively for gender and ethnic groups, were introduced and employed.

The specific contributions of this work is as follows:

1. We introduced two facial representations to highlight the discriminative power of local texture and geometry clues of human faces for gender and ethnicity analysis. Oriented Gradient Maps (OGMs) simulate the response of complex neurons. It is based on a convolution of gradients in specific directions in a given circular neighborhood, and its advantage is insensitive to affine illumination and geometrical transformations. Meanwhile, another feature descriptor we used to represent local shape and texture traits is Local Circular Patterns (LCP), which can be regarded as a variant of Local Binary Patterns (LBP). It replaced the binary quantization scheme of LBP, which is very sensitive to noise, by a clustering based quantization. The effectiveness of both two facial representations will be demonstrated in Chapter 2.

2. We investigated the importance of each facial regions respectively for gender and ethnicity. It was achieved by make use of the wide-spread Adaboost algorithm, to select a compact subset of facial features from the entire multi-modal feature set. The features extracted from various facial regions (such as the ones of eyes, nose, forehead) represented as textures or shapes, highly related

and discriminative to the task of facial gender and ethnicity classification, can be determined and assigned with different weights.

3. We evaluated our proposed approaches on various datasets to benchmark its effectiveness and generalization to gender and ethnicity classification.

b. Multimodal facial expression recognition

Overall, we present an efficient multimodal (2D and 3D) and multiple-order (first and second) feature-based fully automatic FER approach, and validate it trough comprehensive experiments on the BU–3DFE database. Considerable complementary characteristics between the features of different orders and different modalities are highlighted either by early fusion or late fusion of 2D, 3D, as well as 2D and 3D descriptors. The generalization capability of our approach is further evaluated on the Bosphorus database.

The specific contributions of this work is as follows:

1. Our approach used automatically located landmarks each 2D face image and its corresponding 3D mesh scan, which also showed the superiority for multimodal 2D+3D FER. The majority of existing feature-based 3D FER approaches reported their results on the BU–3DFE benchmark based on a large set of (typically 83) 3D facial landmarks manually localized by the database providers. Therefore, the proposed frame-work presents a promising way to these landmark-based approaches so that they can be made automatic using the iPar–CLR algorithm in 2D and 3D multimodal face space.

2. We introduced a novel second-order image gradient based local texture descriptor (HSOG), a novel first-order mesh gradient (i.e., surface normal) based local shape descriptor (meshHOG), as well as a second-order mesh gradient (i.e., surface curvature) based local shape descriptor (meshHOS) are adapted in FER to comprehensively encode the expression variations in both the 2D and 3D modalities.

3. We analyzed both the early fusion (i.e., feature-level) and late fusion (i.e., score-level) strategies of 2D descriptors, 3D descriptors, as well as 2D and

3D descriptors are comprehensively demonstrated and their complementary characteristics are well revealed, which is our third contribution.

1.4 Organization of the Thesis

This thesis is organized as follows:

Chapter 2 makes a brief review of the state of the art on two aspects: facial soft biometric recognition and facial expression recognition. Some representative approaches will be presented and discussed.

Chapter 3 introduces our proposed soft biometric system for gender and ethnicity. Firstly, it introduces two facial features in detail and highlights its improvements to LBP based ones. Meanwhile, the feature selection process as well as the decision-level fusion will be presented. Experiments and results are displayed and discussed in the next section.

Chapter 4 presents our approaches for facial expression recognition. Firstly, we introduce the iPar–CLR based 2D and 3D facial landmark localization procedure. And then, we present the construction details of the HSOG, meshHOG and meshHOS descriptors. In the following section, we list and compare the accuracies of each single 2D and 3D descriptor, and the ones of their fusion. The generalization capability of the proposed approach is discussed in Section 6. Finally, we conclude the paper and point out the limitations and future directions.

Chapter 5 summarizes the thesis results and the dissertation contributions. Finally further research suggestions are given.

Chapter 2

Literature Review

In this chapter, firstly, we review and categorize the proposed approaches in the literature respectively for our tasks: facial soft biometric recognition and facial expression recognition. Then, several basic techniques and most representative methods are presented in the following. Meanwhile, we also summarize available facial databases for face analysis in both two domains.

2.1 Soft biometrics: gender and ethnicity classification

Since soft biometrics traits have been proposed for several years, there are already a number of facial soft biometric systems to identify gender and different ethnic groups.

From the perspective of type of data, the proposed approaches can be roughly categorized into 3 classes: 2D texture based algorithms, 3D shape based algorithms and multi-modal approaches using both of them. From 2D facial appearance, we can extract detailed texture information provided by skin, eye, mouth, etc., and simple geometric information like contour, gradient, etc. Meanwhile 3D facial can provide more precise geometric information, e.g, gradient, normal, curvature. In fact, those information of 2D texture and 3D shape are though of be complementary in theory. Thus, more and more researchers pay their attentions to develop facial soft biometric systems based dual-modalities. In the following sections, we will introduce the methods of the state of the arts respectively on those tree categories.

Basically, facial gender and ethnicity classification is a problem of pattern classification. And for a pattern recognition system, two main steps, features and classifiers, are usually the keys to the effectiveness and robustness of system. Thus,

we review features and classifiers of proposed methods in the literature.

To be more specific, we can mainly classify facial features for this problem in two categories: holistic based and local feature-based. Holistic based methods use global facial information as their soft biometric traits while local feature based methods prefer to make use of hand-crafted descriptors to highlight discriminative in local regions of faces. Additionally, some methods based on learned features are also presented.

2.1.1 2D facial appearance based approaches

Since 2D images and videos are easy to be accessed, a number of features have been proposed for facial biometric system in past 20 years. As mentioned previously, 2D facial appearance based features are designed to describe texture details. In this section, we introduce respectively raw image based, features based approaches.

2.1.1.1 Raw image based approaches

In the early 1990s, as shown in Fig. 2.1, [Golomb *et al.* 1990] trained a two-stage neutral network to discriminate sex in human faces, which consisted of a compression network and a prediction network. The compression network is a standard auto-encoder and the prediction network employed learned intermediate output of compression network as its input feature. Finally, their experiments achieved an average error rate of 8.1% on a set of 90 images.

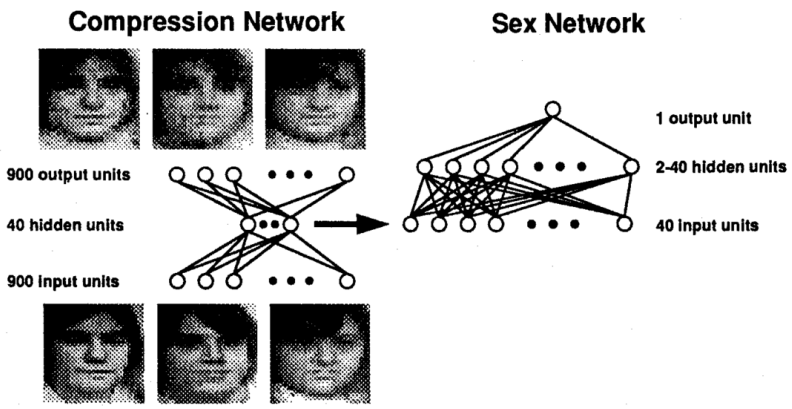


Figure 2.1: Two-stage network for discriminating sex. [Golomb *et al.* 1990]

Another network based approach has also been proposed. [Poggio *et al.*] extracted a set of geometrical features and then used it to train two competing HyperBF network proposed by their previous work [Poggio & Girosi 1990] to classify gender. A correct rate of 79% was achieved on a set of 20 males and 20 females.

In the systems of face recognition, raw image based techniques are always along with several subspace projection techniques: e.g, Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA). [Lu & Jain 2004] applied the LDA analysis to raw images at different scales, then integrated their results to further improve the classification performance, as shown in Fig. 2.2. The pipeline of their method is as follows: Firstly, they re-scaled input frontal facial image into different size: 42×42, 32×32, 24×24. Then, the LDA classifier was trained on each individual scale under a Bayesian statistical decision framework. Lastly, they integrated the classification results to arrive at the final decision using the product rule. Experimental results based on a face database containing 263 subjects (2630

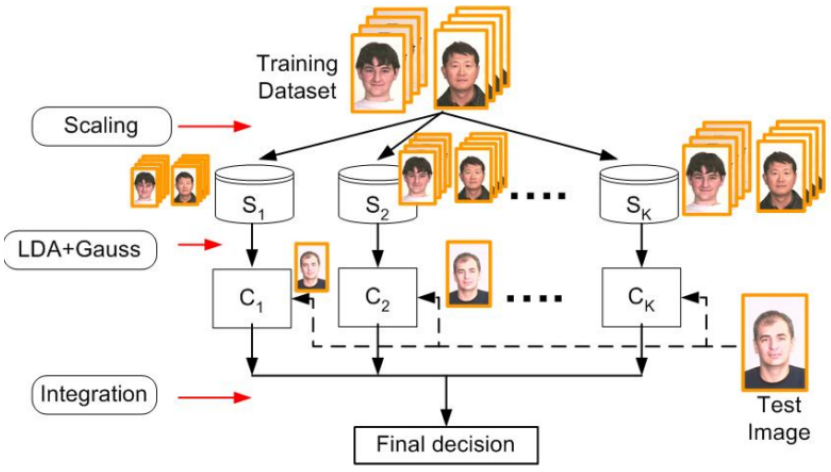


Figure 2.2: A framework for ethnicity classification integrating the results of LDA classifier on raw images with different scales. [Lu & Jain 2004]

images) achieved finally 96.3% for ethnicity classification.

Another work using holistic features is proposed by [Moghaddam & Yang 2000], which compared the performance of SVM and traditional classifiers (LDA, NN) for gender classification with low resolution facial images. Their experimental results demonstrate the effectiveness of SVM for gender classification. Mean-

while, [Kim *et al.* 2006] compared two different classifier Gaussian Process Classifiers (GPCs) and Support Vector Machine (SVM) for gender classification. The experimental results show that the performance of GPCs outperformed SVM on raw images.

More recently, [Chu *et al.* 2013] proposed an approach for identifying the gender based on a set of patches randomly cropped around detected face regions. More specifically, firstly, after the image sets generated by randomly cropping, orthogonal basis matrices were computed by eigenvalue decomposition to represent image sets. Then, each face image set was further projected into the linear subspace constructed by computed orthogonal basis. Secondly, canonical correlations were employed for measuring the similarities between projected sub-spaces. Based on similarity matrix, A FSD kernel [Wolf & Shashua 2003] was computed and adopted as the kernel of SVM. The proposed approach was evaluated on two public databases, and the recognition rates were achieved respectively on FERET (above 89%) and MORPH (above 93%) for gender classification.

2.1.1.2 Features based approaches

In general, local features usually outperforms holistic features in the facial biometric systems. Reviewing the literature of facial gender and ethnicity classification, the most commonly used features are variants of Gabor filter and Local binary pattern (LBP). Thus, we introduce firstly Gabor filter based and LBP based methods, and then some other particular features will be also presented.

a. Gabor filter based approaches

From the perspective of biology, 2D Gabor wavelets are similar as the activation of simple cells of visual cortex in human brains. Thus, 2-D Gabor filters have been widely used in image prepossessing, especially in feature extraction for texture analysis and segmentation. Besides, Gabor wavelet transform has both the multi-resolution and multi-orientation properties and are optimal for measuring local spatial frequencies. Mathematically, the kernel function of Gabor filter is defined in [Prasad & Domke 2005] as follows:

$$g(x, y, \theta, \phi) = \exp(-\frac{x^2 + y^2}{\sigma^2}) \exp(2\pi\theta i(x\cos\theta + y\sin\theta)) \tag{2.1}$$

where θ is the spatial frequencies, ϕ is the orientation, and σ is the standard deviation of the Gaussian kernel.

The response of a Gabor filter to an image is obtained by a 2D convolution operation. Let $I(x, y)$ denote the image and $G(x, y, \theta, \phi)$ denote the response of a Gabor filter with frequency θ and orientation ϕ to an image at point (x, y) on the image plane. $G(.)$ is obtained as:

$$G(x, y, \theta, \phi) = \int\int I(p, q)g(x - p, y - q, \theta, \phi)\mathrm{d}p\mathrm{d}q \tag{2.2}$$

One of the most representative method based on Gobar filter is the work of [Guo & Mu 2010]. The main contribution of their work is that they introduced a set of biologically-inspired features and verified its performance for ethnicity classification. Meanwhile, they produced a benchmark result for large-scale ethnicity estimation on a database with more than 55,000 face images.

The biological-inspired features (BIF) were initially proposed by [Riesenhuber & Poggio 1999], derived from a feed-forward model of the primate visual object recognition pathway. The initial model consisted of two layers simulating respectively simple and complex cell units, and the propagation within the model indicated the progress from the primary visual cortex to inferior temporal cortex, as shown in Fig. 2.3.

The first layer convolved input image with a bank of Gabor filters of four orientations and 16 scales. And the second layer max-pooled values within a local spatial neighborhood and across the scales (two consecutive scales). The advantage of taking the "max" operation is though of to tolerate small shifts and changes in scale.

After the feature extraction, they used some subspace analysis techniques, e.g, PCA and Orthogonal Locality Preserving Projections (OLPP) to preserve the local structure of the face manifold. In more detail, OLPP is defined as follows: Given a set of samples $X = \{x_1, x_2, \ldots, x_n\}$, we firstly define a symmetric matrix of weights

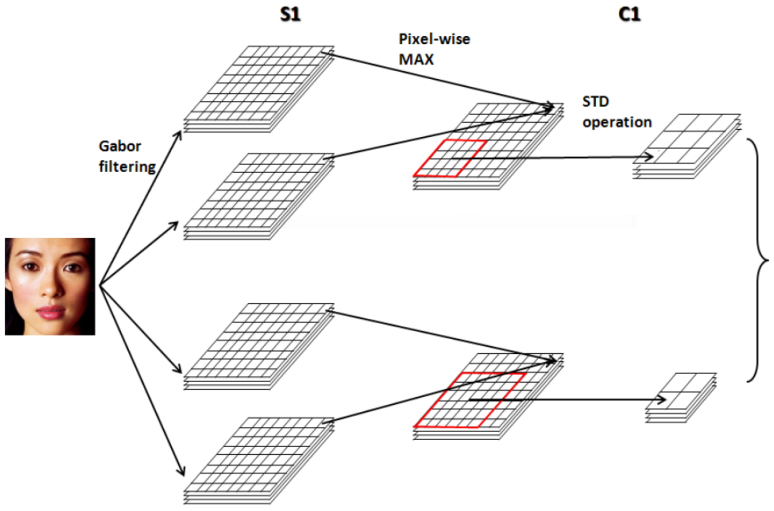


Figure 2.3: The biologically-inspired feature extraction for ethnicity estimation. [Guo & Mu 2010]

S, which measures the distances between samples:

$$S_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{t}) \tag{2.3}$$

Then, a diagonal matrix D and a laplacian matrix L can be further computed:

$$D_{ii} = \sum_j S_{ij}, \qquad L = D - S \tag{2.4}$$

At last, the optimal projection p is define as:

$$p = \arg \min_{p^T XDXp=1} \sum_{i=1}^{n} \sum_{j=1}^{n} (p^T x_i - p^T x_j)^2 S_{ij} \tag{2.5}$$

Finally a linear SVM classifier was employed to distinguish different ethnic groups using projected features. [Guo & Mu 2010] carried out their experiments on MORPH-II database, which contains more than 55000 facial images. The experimental results showed that their method was effective for discriminating Black (98.3%) and White (97.1%), but powerless for recognizing Hispanic (74.2%), Asian (59.5%) and Indian (6.9%). Besides, in [Guo & Mu 2010], they also studied the effect of gender and age variations on ethnicity estimation.

More recently, [Guo & Mu 2014] proposed a framework for joint estimation of age, gender, and ethnicity, which introduced Gabor filters for feature extraction and employed canonical correlation analysis (CCA) as classifier to discriminate soft biometric traits. The experiments are conducted on a very large database containing more than 55,000 face images

Moreover, some other approaches using Gabor based feature have also been proposed. [Hosoi *et al.* 2004] combined Gabor wavelet transform and retina sampling to extract features for ethnicity classification. A high degree of accuracy was achieved by SVM classifier: Asian (96.3%), European: (93.1%), African: (94.3%) on a data-set of 1991 facial images. Authors in [Lian *et al.* 2005] employed a Min-Max Modular Support Vector Machine (M^3-SVM) to classify features extracted by Gabor filters for gender recognition problem. The experimental results indicate that their new method with M^3-SVM have better performance than traditional SVMs. Their another work [Luo & Lu 2006] further verified the performance of this method by using a equal clustering based task decomposition method. The approaches developed in [Lu & Lin 2007] compared different features (Gabor, Haar-like, PCA, Independent Component Analysis (ICA)) respectively with two classifiers (SVM and Adaboost). The experimental results show that their proposed approach (combination of ellipse face images, Gabor wavelets and Adaboost+SVM classifier) achieved better performance: 90% for gender classification on FERET database.

Another system using Gabor filters along with $(2D)^2$PCA was introduced in [Rai & Khanna 2014]. The proposed approach achieved 98.4% classification rate for non-occluded face images and correct rates ranging from 10% to 60% for occluded face images on FERET database.

b. LBP-like feature based approaches

Local binary patterns (LBP) is a popular visual descriptor for classification in computer vision, first proposed in [Ojala *et al.* 1994]. A basic LBP operator simply thresholds a 3×3 neighborhood by the value of the central pixel, and the sign of thresholded neighboring values can form a binary number, which is then transformed into a decimal number. This decimal number is treated as the label of the central pixel. The histogram of the labels within a region is often used as a

texture descriptor.

Formally, given a pixel at (x_c, y_c), the derived LBP decimal value is:

$$LBP(x_c, y_c) = \sum_{n-0}^{8} s(i_n - i_c)2^n \tag{2.6}$$

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{2.7}$$

where n covers the eight neighbors of the central pixel, and i_c and i_n are the gray level values of the central pixel and its surrounding ones respectively.

Due to the simplicity and effectiveness of LBP, many methods based on LBP-like features have been proposed. More specifically, [Yang & Ai 2007] proposed a method using LBPH features and Adaboost classifier. The Chi square distance was employed as a measure of similarities between local patches. The experiments on gender and ethnicity classification demonstrate its effectiveness, which achieved respectively error rate of 6.7% for gender classification on FERET database and 6.8% for ethnicity classification on PIE database. The method developed in [Li *et al.* 2012a] extracted LBPH features on five facial components: forehead, eyes, nose, mouth and chin. Furthermore, given features of each component, a set of SVM classifiers were trained. Finally, they also discussed different strategies of fusion. The system proposed in [Lyle *et al.* 2010] extracted LBP features only from periocular region images with high resolution. For 4232 periocular images of 404 subjects, they obtained a baseline gender and ethnicity classification accuracy of 93% and 91% respectively. [Shan 2012] investigated gender recognition on real-life faces. Local Binary Patterns (LBP) was employed to describe faces, and Adaboost was used to select the discriminative LBP features. They obtained the performance of 94.81% by applying Support Vector Machine (SVM) with the boosted LBP features. Authors in [Shih 2013] proposed a robust gender classification system, which firstly located landmarks using AAM algorithm, then extracted LBPH features in the patches around landmarks, and finally used a Bayesian model to predict gender of face images. The best recognition rate of proposed method is 86.5% on a dataset

of 1862 male and 1503 female.

The basic LBP employs a binary quantization which only retains the sign of pixel level differences. Thus, [Patel *et al.* 2015] proposed a new quantization schema being able to encode both the sign and magnitude information for LBP.

A comparative analysis about the performances of LBP-like features has been done by [Hadid *et al.* 2015]. In their work, they investigated totally 13 variants of LBP features for gender classification, such as classic LBP, Local Phase Quantization (LPQ), Binarized Statistical Image Feature (BSIF), etc. The experiments were carried out on a data-set of 20,127 face samples collected from Flickr and taken in uncontrolled environment. Each face image was normalized based on the eye coordinates, resulting in images of 64×64 pixels. These images are divided into 6×6 blocks with an overlap of 4 pixels. The extracted features from each block are concatenated into a feature vector which is fed to the SVM classifier. The best results are obtained with BSIF but at the cost of higher computational time compared to basic LBP.

[Jia & Cristianini 2015] implemented Haar-like features based face detection algorithm [Viola & Jones 2004] to locate faces regions on a wild database. Then multi-scale LBPs were used to encode global information on the whole face and local traits on a set of patches around facial landmarks. More specifically, every face sample was cropped to 90×120, and divided into 90 blocks of 10×12. Images of facial landmarks were resized to 40×40 and the block size is 10×10. Hence, there are 90 and 16 blocks for faces and local features, respectively. Multi-scale LBP features, $LBP^{u2}_{8,1}, LBP^{u2}_{8,2}, \ldots, LBP^{u2}_{8,8}$ were extracted so totally we have $59 \times 90 \times 8 + 59 \times 16 \times 8 \times 3 = 65,136$ dimensions for each sample. Two on-line classifiers the Pegasos and the C-Pegasos [Shalev-Shwartz *et al.* 2011] were further trained on the four million images. Finally, A high prediction rate (96.86%) was obtained by integrating 400 weak classifiers on LFW database.

[Zheng & Lu 2011] propose a support vector machine with automatic confidence (SVMAC) [Ji *et al.* 2008] for gender classification. Multiple features including LGBP, MLBP, LBP, Gabor and Gray-scale, were introduced to demonstrate the effectiveness of their method.

c. Other features based approaches

More recently, [Mery & Bowyer 2014] a new general approach called Adaptive Sparse Representation of Random Patches (ASR+) to recognize facial expression, gender, ethnic and disguise. The pipeline of their method was as shown in Fig. 2.4:

- For each class of the gallery, several random small patches are extracted from their images.
- Both intensity and location information were considered to describe small patches.
- The descriptions of patches in training sets were employed to build representative dictionaries. However, only those patches that are not filtered out by the stop list are considered.
- In the test stage, each test patch was classified in accordance with the Sparse Representation Classification (SRC) methodology [Wright *et al.* 2009]
- Finally, the query image is classified by voting for the selected patches.

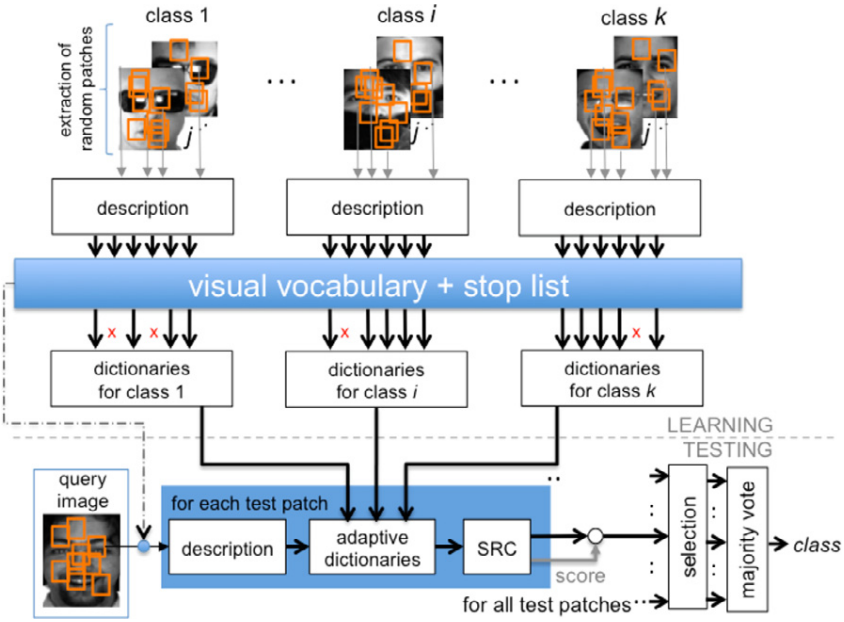


Figure 2.4: Overview of the proposed method. There are two stages: learning and testing. The stop list is used to filter out patches that are not discriminating for these classes. The stopped patches are not considered in the dictionaries of each class and in the testing stage. [Mery & Bowyer 2014]

More specifically, the description of patches was introduced as follows: In the

training stage, a set of n face images of k classes is available, where I_j^i denotes image j of class i , $i = 1, \ldots, k$ and $j = 1, \ldots, n$. In each image I_j^i, m patches are randomly extracted. In this work, the description of a patch $\mathcal{P}$ is defined as vector:

$$\mathbf{y} = f(\mathcal{P}) = [\mathbf{z}; \alpha x; \alpha y] \in \mathbb{R}^{d+2} \tag{2.8}$$

where $\mathbf{z} = g(\mathcal{P}) \in \mathbb{R}^d$ is a descriptor of patch $\mathcal{P}$; (x, y) are the image coordinates of the center of patch $\mathcal{P}$; and α is a relative weighting factor between description and location. Using Eq. (2.8), all extracted patches are described as $\mathbf{y}_{jp}^i = f(\mathcal{P}_{jp}^i) = [\mathbf{z}_{jp}^i; \alpha x_{jp}^i; \alpha y_{jp}^i]$, for $p = 1, \ldots, m$.

The description $\mathbf{Y}^i$ of class i is defined as an array with the description of all non stopped patches $\mathbf{y}_{jp}^i$. A k-means algorithm was used to clustering $\mathbf{Y}^i$ into Q centers, which was referred to as ***parent clusters***:

$$\mathbf{c}_q^i = k\text{-}means(\mathbf{Y}^i, Q) \tag{2.9}$$

for $q = 1, \ldots, Q$, where $\mathbf{c}_q^i \in \mathbb{R}^{d+2}$ is the centroid of parent cluster q of class i. $\mathbf{Y}_q^i$ was defined as the array with all samples $\mathbf{y}_{jp}^i$ that belong to the parent cluster with centroid $\mathbf{c}_q^i$. In the same way, each ***parent clusters*** was clustered again in R ***child clusters***:

$$\mathbf{c}_{qr}^i = k\text{-}means(\mathbf{Y}_q^i, R) \tag{2.10}$$

for $r = 1, \ldots, R$, where $\mathbf{c}_{qr}^i \in \mathbb{R}^{d+2}$ is the centroid of child cluster r of parent cluster q of class i. All centroids of child clusters of class i are arranged in an array $\mathbf{D}^i$, and specifically for parent cluster q are arranged in a matrix:

$$\hat{\mathbf{A}}_q = [\mathbf{c}_{q1}^i, \ldots, \mathbf{c}_{qr}^i, \ldots, \mathbf{c}_{qR}^i]^T \in \mathbb{R}^{(d+2)\times R} \tag{2.11}$$

In the test stage, given the set of all centroids of parent clusters $\hat{\mathbf{A}}_q$ and the dictionaries of all centroids of child clusters $\mathbf{D}$, the best representative $\mathbf{A}(Y)$ patches could be found by measuring their distance to centroids. Only the patches having a minimal distance less than a threshold θ would be choose, as shown in Fig. 2.5.

Finally, a sparse representation for patch $\mathbf{y}$ was obtained by Sparse representation

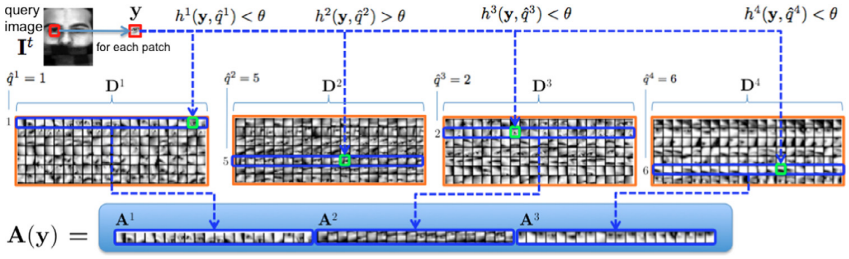


Figure 2.5: Adaptive dictionary $\mathbf{A}$ of patch $\mathbf{y}$. In this example there are k = 4 classes in the gallery. [Mery & Bowyer 2014]

classification (SRC) using the l_1-minimization approach:

$$\hat{\mathbf{x}} = \arg\min \|\mathbf{x}\|_1 \qquad \text{subject to} \qquad \mathbf{A}\mathbf{x} = \mathbf{y} \tag{2.12}$$

[Mery & Bowyer 2015] extended their previous work in [Mery & Bowyer 2014]. The experiments were carried out on eight face databases and the results showed their method outperformed various representative methods in the literature in a complex scenarios.

2.1.2 3D face shape based approaches

Similar as 2D textured based approaches, on 3D modality, researchers are also interested in finding characteristics to highlight differences between gender and ethnic groups. On 3D modality, the most common features used always take into geometric properties, such as gradient, curvature, normal, etc.

The first work using 3D shape data for gender classification is [O'toole *et al.* 1997]. In this work, they compared the performance of 3D head structure data with that of gray-scale images for sex classification. The experimental results showed that their performance were comparative and their information were relevant somewhere.

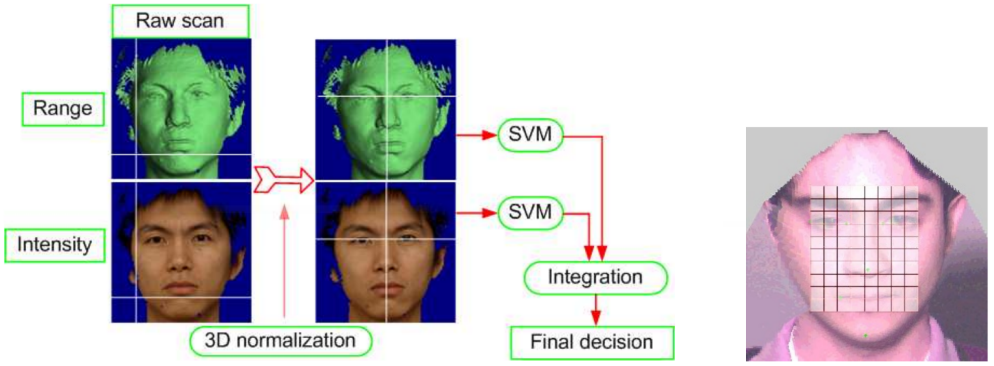
[Han *et al.* 2009] proposed a method, which computed volumes and areas of different facial sub-regions located by manual landmarks as the feature and employed SVM as the classifier. The average error rate of 17.44% was achieved on GavabDB

database.

For gender classification, [Wu *et al.* 2009] adapted facial needle-maps as the facial representation, and introduced the weighted principal geodesic analysis (PGA) [Fletcher *et al.* 2004], which is a generalization of PCA, to construct the subspace of measuring similarities. [Wu *et al.* 2010] extended this work. An iterative SFS method [Smith & Hancock 2008] was employed for the recovery of the facial needle-maps. Besides, they also investigated the performance of intensity image + SFS for gender classification. The experimental result shown the recognition rates of needle-maps and intensity images were comparative. Furthermore, [Wu *et al.* 2011] gave more details about their methods, and newly proposed a learning strategy to find the optimal weight PGA map. Finally, their method using supervised weighted PGA achieved 97.00% on a dataset of 200 facial needle-maps.

[Hu *et al.* 2010] proposed an approach, which integrated the results respectively from five different facial regions for gender classification. More specifically, in the first step, face landmarks were extracted from 3D face shape based on profiles and curvature. Then, the whole face was divided into several sub-regions according to landmarks. For each sub-region, the representation of the geometrical property, Shape Index [Dorai & Jain 1997], was computed as the feature. Finally, SVM classifiers were employed for each sub-region, and in the testing stage, the outputs of SVMs were further integrated to determine the gender. The highest correct classification rate of their method was 94.3% on a dataset of 945 facial scans.

[Toderici *et al.* 2010] studied an interesting problem: whether a face recognition system could be leveraged to identify gender or ethnicity, and how effective it could be. Considering this motivation, they introduced a 3D face recognition system in their previous work [Kakadiaris *et al.* 2007]. The same similarity function, which was previously used for distinguish identifications, was employed again for measuring relations of gender and ethnicity features. Four different types of classification experiments: KNN, kernelized kNN, learning based on the face-similarity space (MDS) and learning based on wavelet coefficient, were carried out on a dataset of 3676 facial meshes. The experimental results achieved on average 93.5% for gender classification and 99.5% for ethnicity classification, showing the effectiveness of 3D



(a) (b)

Figure 2.6: (*a*) System Diagram for gender and ethnicity identification. (*b*) Cropping face areas for construction of feature vectors. A 10×8 grid is overlaid on the facial scan for demonstration. [Lu *et al.* 2006]

face for soft biometric recognition.

2.1.3 Multi-modal approaches

Motivated by the fact that 2D texture and 3D shape are complementary somewhere, more and more researchers are interested in developing multi-modal approaches for soft biometric recognition.

One of the most popular approaches was proposed by [Lu *et al.* 2006], which proposed an integration scheme for ethnicity and gender identifications by combining the registered range and intensity images, as shown in Fig. 2.6(a). The construction of feature vectors was a little simple, just using mean values of cells in a 10×8 grid overlaid on the cropped face for both modalities, Fig. 2.6(b).

Instead of matching scores, the posterior probabilities are extracted from the SVMs [Platt *et al.* 1999]. And the sum rule [Kittler *et al.* 1998] was employed as combination strategies of fusion at the decision level, formulated as:

$$\begin{aligned} p(male|s) &= \frac{p(male|s_{range})+p(male|s_{intensity})}{2} \\ p(female|s) &= \frac{p(female|s_{range})+p(female|s_{intensity})}{2} \end{aligned} \tag{2.13}$$

The experiments were conducted on a database containing 1240 facial scans of 376 subjects. It was demonstrated that the range modality provides competi-

tive discriminative power on ethnicity and gender identifications to the intensity modality.

[Alexandre 2010] did similar work as [Lu *et al.* 2006], but with more complex features. Reviewing that LBP have shown its strong power on 2D intensity images, they employed LBP as the texture feature. Meanwhile, they also introduced histograms of edge direction inspired by [Dalal & Triggs 2005] as the features of 3D shape. A majority vote decision was set as the fusion strategy. The proposed method achieved the accuracy of 99.07% on FERET database and 91.19% on UND database for gender recognition.

Another multimodal framework was proposed by [Zhang & Wang 2009]. LBP histograms are extracted from multi-scale, multi-ratio rectangular regions over both texture and range images, and Adaboost is utilized to construct a strong classifier from a large amount of weak classifiers built by the extracted LBP histograms. Decision level fusion is performed to get the final decision. Above 99.5% classification accuracy was obtained on the FRGC v2.0 database.

2.1.4 Some interesting work

Most of proposed approaches for soft-biometric classification is based on static images. There are few work focusing on estimating soft-biometric using videos except [Hadid & Pietikäinen 2013]. More specifically, the volume LBP (VLBP) [Zhao & Pietikainen 2007] was employed as the feature, and manifold learning was used to exploit the correlation between the face images. The experiments on the gender and age classification problems showed that the proposed method outperformed traditional static image based methods

Another interesting work is proposed by [Bekios-Calfa *et al.* 2014]. In this work, the dependencies among gender, age and pose facial attributes were studied when building the classifier. The experimental results confirmed the existence of dependencies among gender, age and pose facial attributes and proved that the performance and robustness of gender classifiers can be further improved by exploiting these dependencies.

Recently, some low-cost depth sensors such as Microsoft Kinect allow extracting

directly 3D information together with RGB color images, providing new opportunities for computer vision and face analysis research. [Boutellaa *et al.* 2015] explored the usefulness of RGB-D data provided by Kinect for identity, gender and ethnicity recognition. Four local features (LBP, LPQ, HOG and BSIF) were investigated for both face texture and shape description. The experimental results were promising, beyond the expectations based on the human perception.

2.1.5 Summary

In the previous sections, we have reviewed and presented the proposed approaches in the literature. Many of them have been demonstrated their effectiveness for gender or ethnicity classification. Some experimental comparisons of the proposed approaches have been done by [Mäkinen & Raisamo 2008, Andreu *et al.* 2014]. [Mäkinen & Raisamo 2008] compared a set of methods (Neural Network, SVM, Threshold Adaboost, LUT Adaboost [Wu *et al.* 2003], Mean Adaboost [Freund & Schapire 1995]), which took histogram equalized image pixels as the input. Meanwhile, [Andreu *et al.* 2014] carried out a comprehensive comparison of two representation approaches (global and local), three types of features (grey levels, PCA and LBP) and three classifiers (1-NN, PCA+LDA and SVM).

However, it is unfair and impossible to compare the performances of all existed systems because of various databases, various configurations of experiments and various criteria of evaluation. Thus, we list the representative approaches and their configurations in the Table 2.1.

As mentioned in the previous chapter, for 2D texture, we have a wide range of facial database with good qualities for face analysis. For 3D modality, the choice of databases is relatively limited. Thus we list the available databases for gender and ethnicity classification here in Table 2.2.

2.2 Facial expression recognition

Facial expression analysis/recognition1 has interested many researchers due to its various purposes and applications. To our best knowledge, Facial Action Coding

Table 2.1: Overview of methods for gender and ethnicity classification in the literature.

Approaches	Features-Classifiers	Data-set	Performance	
			Gender	Ethnicity
[Moghaddam & Yang 2000]	thumbnail faces + SVM	FERET(1755 images)	96.62%	
[Hosoi *et al.* 2004]	Gabor Wavelet + SVM	1991 images		94%
[Lu & Jain 2004]	Raw image + LDA	2630 images		96.3%
[Lian *et al.* 2005]	Gabor + M^3-SVM	2055 images	91.53%	
[Lu *et al.* 2006]	Pixel + SVM	376 scans	91%	98%
[Lu & Lin 2007]	Gabor+Adaboost+SVM	FERET(818 images)	90%	
[Yang & Ai 2007]	LBPH+Adaboost	FERET,PIE,snapshots	93.3%	93.2%
[Han *et al.* 2009]	Volume,Area+SVM	GavabDB	82.56%	
[Zhang & Wang 2009]	MM-LBP+Adaboost	FRGC(1917 scans)		99.58%
[Alexandre 2010]	multiple features+SVM	FERET(411 images) UND(487 images)	99.07% 91.19%	
[Guo & Mu 2010]	BIF+OLPP+SVM	MORPH-II(55000 images)		92% ∼ 94%
[Hu *et al.* 2010]	Shape-Index + SVM	945 scans	94.3%	
[Toderici *et al.* 2010]	statistical model	FRGC(4007 meshes)	93.5%	99.6%
[Lyle *et al.* 2010]	LBP + SVM	FRGC	93%	91%
[Wu *et al.* 2011]	needle-maps+PGA	Max-Planck(200 scans) UND(944 scans)	97% 96.9%	
[Shan 2012]	LBP + Adaboost	LFW	94.81%	
[Hadid & Pietikäinen 2013]	LBP+Manifold	2000 sequences	97.2%	
[Shih 2013]	PPH + PGC	3365 images	86.5%	
[Guo & Mu 2014]	BIF+CCA	three sets / 5265 images	above 98%	99%
[Rai & Khanna 2014]	Gabor+$(2D)^2$PCA	FERET(700 images) LFW(13010 images)	98.4% 89.1%	
[Mery & Bowyer 2015]	Patches + SRC	FERET(1040 images) UND FRGC 2.0 GROUPS(1978 images)	94.1% 92.5%	 87.1% 93.3%
[Jia & Cristianini 2015]	LBP + C-Pegasos	4 million images	96.86%	

Table 2.2: Overview of database for gender and ethnicity classification in the literature.

Database	of People	Total sample	Intensity	Mesh(Range)	Gender	Ethnicity
FERET	1199	14126	Yes		Yes	Yes
PIE	68	41368xs	Yes		Yes	Yes
UND	591	3378	Yes		Yes	
MORPH-II		55,000	Yes		Yes	Yes
LFW	5749	13,233	Yes		Yes	
GavabDB	61	437	Yes	Yes	Yes	
FRGC V2	466	4007	Yes	Yes	Yes	Yes

System (FACS) is still the most powerful tool for systematically analyzing expressions. FACS measures all visible expressions in terms of several decomposition components, so called Action Units. Therefore, each expression can be recognized by detecting the combination of appearance of Action Units.

Reviewing the techniques proposed in the literature, a wide range of systems proposed focus on recognizing expressions from 2D facial expression data (static and dynamics), static 3D facial expression data, and more recently dynamic 3D (4D) facial expression data. Although 2D FER systems have achieved remarkable performance, 3D FER are still considered to be more potential in expression analysis, because 3D face is more robust to illumination and pose variations and 3D geometric traits are more rational to present motions of facial muscles.

Effective expression analysis normally depends on accurate representation of locations and motions of facial features. Traditionally, a static FER method consists of two main stages: feature extraction, and selection and classification of features, while approaches of dynamic FER usually employ temporal modeling techniques to highlight changes in sequences. Meanwhile, considering that expressions are caused by motions of facial muscles and single-view2D analysis is unable to fully exploit the information displayed by the face, a number of statistical models are also used to provide the prior knowledge to accurately measure the deformation of facial features or locate and track the salient points.

Since the techniques of expression analysis have been focused many years, some comprehensive surveys in this area have been done by [Fasel & Luettin 2003, Zeng *et al.* 2009, Fang *et al.* 2011, Bettadapura 2012, Sandbach *et al.* 2012b,

Danelakis *et al.* 2015]. More specifically, [Fasel & Luettin 2003] concluded approaches of FER in 2003, while [Bettadapura 2012] surveyed some published work in the field of FER from 2001 to 2012. [Zeng *et al.* 2009] provided exhaustive coverage of efforts in the field of automatic recognition of human affect proposed before 2009. In 2011, a comprehensive survey of 3D FER was carried out by [Fang *et al.* 2011]. Furthermore, [Sandbach *et al.* 2012b] analyzed 3D static and dynamic FER systems. More recently, [Danelakis *et al.* 2015] gave a detail introduction of FER systems in 3D video sequences.

Similar as surveys mentioned above, in this section, we present the state of the art in the field of FER. Firstly, we give a comparative introduction of several well-known statistical models respectively for 2D and 3D faces. Then, a set of features available for FER and their correspond classifiers are listed. Some temporal modeling techniques are discussed in the following part. Finally, we summarize the approaches in the literature.

2.2.1 Statistical models for FER

Many techniques have been proposed for tracking, aligning, and modeling faces in FER systems. In this section, we present some of the most representative statistical models.

2.2.1.1 Point Distribution Models

Active Shape Model (ASM) is a specific point distribution model firstly proposed in [Cootes *et al.* 1995]. Formally, for a 2D image, a set of n landmark points, $\{(x_i, y_i)\}$, can be presented in a single $2n$ element vector, $\mathbf{x}$, where $\mathbf{x} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^T$. In the training stage, a Principal Component Analysis (PCA) is then applied to the space constructed by landmarks data, we can then approximate any of the training set, $\mathbf{x}$ using

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P_s}\mathbf{b_s} \tag{2.14}$$

where $\bar{\mathbf{x}}$ is the mean of $\mathbf{x}$, $\mathbf{P_s} = (p_{s_1}|p_{s_2}|\ldots|p_{s_t})$ contains t eigenvectors of the

covariance matrix and $\mathbf{b_s}$ is a t dimensional vector given by

$$\mathbf{b_s} = \mathbf{P_s}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{2.15}$$

[Edwards *et al.* 1998] extended the basic idea of ASM and proposed a new model, Active Appearance Model (AAM). Unlike ASM only applies the PCA to shape vectors, AAM applies the PCA projection to both vectors of shape and texture information. More specifically, the texture modeling can be described as

$$\mathbf{g} \approx \bar{\mathbf{g}} + \mathbf{P_g}\mathbf{b_g} \tag{2.16}$$

where $\mathbf{g}$ indicates the gray level information extracted from shape-normalised face patches.

The biggest advantage of ASM/ASM is that it can be fast fitted and has the robustness to noise, but it only provide spatial correspondences and only utilize information restricted to salient points.

In many cases for FER, ASM/AAM model was employed to local facial landmarks [Wang & Yin 2007, Shbib & Zhou 2015, Tsalakanidou & Malassiotis 2009]. Some of FER systems also utilized the fitting results of ASM/AAM as facial features to distinguish expressions. One of the first such woks is [Lanitis *et al.* 1997], in which an ASM was employed for representing both shape and gray-level appearance. Authors in [Abboud *et al.* 2004] utilized the standard AAM representation to perform automatic FER. [Zalewski & Gong 2004] extended the basic AAM to implicitly incorporate parameters for rotation, scale and large pose variations. [Shbib & Zhou 2015] employed coordinates of salient point fitted by ASM as the features.

2.2.1.2 Bilinear Expression Models

In [Mpiperis *et al.* 2008], authors proposed bilinear models able to handle joint identity and expression contribution to facial appearance. The pipeline of the method is as follows. Firstly, an elastically deformable 3D model was employed for establish-

ing point correspondence among faces. Modeling consists in finding the base-mesh vertices that minimize the deformation energy. Meanwhile, Some boundaries should be detected first as a prior knowledge to constraint parameters. Once the dense correspondences between vertex is established, two types of bilinear models, the symmetric and asymmetric bilinear model, are further proposed for identity and expression recognition, respectively. The symmetric model symmetrically weighs the coefficients which control identity and expression. This is accomplished by letting the interaction coefficients vary with the expression control parameters w_{ijk} vary with the expression control parameters a_i [Tenenbaum & Freeman 2000],

$$a_{kj}^{x} = \sum_{i=1}^{I} w_{ijk} a_i^x \tag{2.17}$$

and the vector representation of the face is now given by

$$v_k^{xp} = \sum_{j=1}^{J} a_{kj}^x b_j^p \tag{2.18}$$

$$\mathbf{v}^{xp} = \mathbf{A}^x \mathbf{b}^p \tag{2.19}$$

where $\mathbf{A}^x$ controls expression, and $\mathbf{b}^p$ controls the identity of faces. Furthermore, the asymmetric model is fitted by minimizing the total square error [Freeman & Tenenbaum 1997]:

$$E_a = \sum_{i=1}^{T} \sum_{x=1}^{T_x} \sum_{p=1}^{T_p} h_{xp}(t)(\mathbf{v}(t) - \mathbf{A}^x \mathbf{b}^p)^2 \tag{2.20}$$

The fitting results were shown in Fig. 2.7. The performance of bilinear models was evaluated on BU-3DFE database [Yin *et al.* 2006b]. This evaluation was repeated ten times ensuring that every subject is included in one test. An average recognition rate of 90.5% was achieved, which proved the effectiveness of their method.

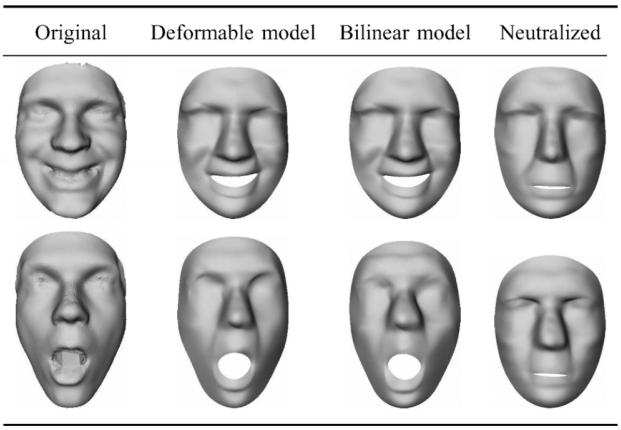


Figure 2.7: Expression manipulation. The first column shows the original 3-D face scans of the same subject displaying a smiling (first row) and a surprising (bottom row) expression. The second column shows the elastically deformable model fitted to the original surfaces, while the third column shows reconstructions of the surfaces using bilinear model coefficients. Neutralization of expressions shown in the fourth column is achieved by modulating the expression control parameters [Mpiperis *et al.* 2008].

2.2.1.3 Morphable Expression Models

3D morphable models (3DMMs) [Blanz & Vetter 1999] are widely used for face analysis because the intrinsic properties of 3D faces provide a representation that is immune to intra-personal variations such as pose and illumination. Considering the effectiveness of 3DMMs on facial recognition [Blanz & Vetter 2003], several variants of 3DMM developed for FER have been also proposed.

One of the first works was proposed by [Ramanathan *et al.* 2006], in which a new model, Morphable Expression Model (MEM), was able to model a range of different expressions for a particular individual. The MEM was described as follows. Given m expressions of a person, $\mathbf{C} = \{C_1, \ldots, C_m\}$ presented the differences of each expression and the neutral face. The mean expression was defined as $\bar{C} = \frac{1}{m}\sum_{i=1}^{m} C_i$. The $(m+1)C_i'$ (including the neutral) were further defined relative to the base face, i.e., $C_i' = (C_i - \bar{C})$. Then a PCA was performed on the covariance matrix MM^T, where $M = [C_1', \ldots, C_{m+1}']$, to generate the m + 1 eigen-expressions that characterize the variations among the various expressions. $\{V_1, \ldots, V_m\}$ denoted the leading m eigen-expressions being able to effectively differentiate the various

expressions. the MEM is defined as

$$S = \bar{C} + \sum_{i=1}^{m} \zeta_i V_i \tag{2.21}$$

[Amberg *et al.* 2008] extended their previous work [Blanz & Vetter 1999], and proposed a new model, which was able to project faces into two independent subspaces of identity and expressions. The model is defined as follows:

$$f(\alpha_n, \alpha_e) = \mu + M_n\alpha_n + M_e\alpha_e = \mu + M\alpha \tag{2.22}$$

$$M = [M_n \,|\, M_e] \qquad \alpha = \begin{bmatrix} \alpha_n \\ \alpha_e \end{bmatrix} \tag{2.23}$$

where M_n is identity matrix and α_n is identity coefficients, while M_e is expression matrix, and α_e is expression coefficients. The fitting of 3DMM is solved by minimizing the cost function in the following:

$$f(R, t, \alpha) = \sum_i \|R(\mu_i + M_i\alpha) + t - u_i\|^2 + \lambda\|\alpha\|^2 \tag{2.24}$$

where R and t denote respectively matrix of rotation and translation. $\mathbf{u} = [u_1, \ldots, u_n]$ represent the dense correspondence points found by non-rigid algorithm [Amberg *et al.* 2007].

The main advantage of this method is that it is robust against scans with artifacts, noise, and holes, as shown in Fig. 2.8(a).

Furthermore, instead of the way to apply PCA on the whole face in [Amberg *et al.* 2008], [Cheng *et al.* 2015] employed the PCA on each sub-regions of faces segmented by an annotated model of the face (AFM) [Passalis *et al.* 2005], as shown in Fig. 2.8(b). Using a dynamic subdivision framework made this method possible to accurately capture the subtle details in a high-resolution facial scans.

The experimental results obtained on BU4DFE database, demonstrated that the proposed algorithm largely outperforms state-of-the-art algorithms for 3D face fitting and alignment especially when it came to the description of the mouth region.

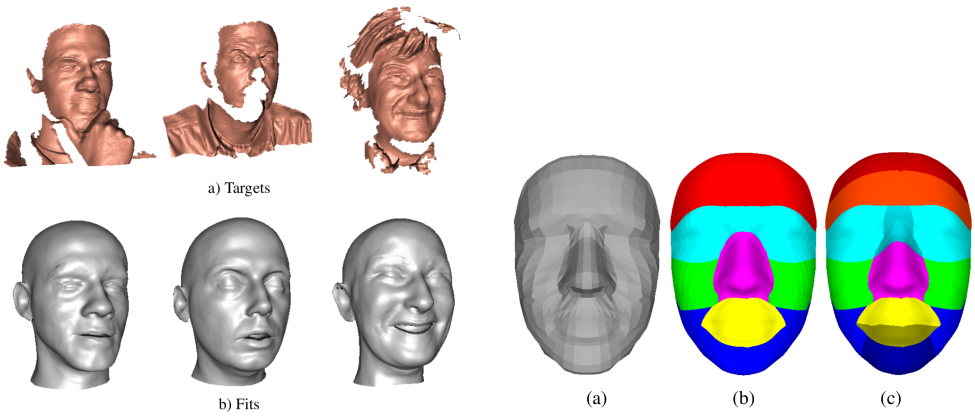


Figure 2.8: Left: The reconstruction (b) is robust against scans (a) with artifacts, noise, and holes [Amberg *et al.* 2008]. Right: Annotated areas for different levels of AFM. (a) The starting level, without subdivision and segmentation; (b) First subdivision level, with 6 annotated face parts (cheeks are combined into one part); (c) Second subdivision level, with 13 individual parts. [Cheng *et al.* 2015]

Another work making use of 3DMM was proposed by [Chu *et al.* 2014], in which a complex system employing an extended 3DMM to neutralize and transfer expressions was presented. Specifically, the 3DMM is fitted to an input 2D face image to compute the identity coefficients, the expression coefficients and the pose parameters. The framework of the method is shown in Fig. 2.9

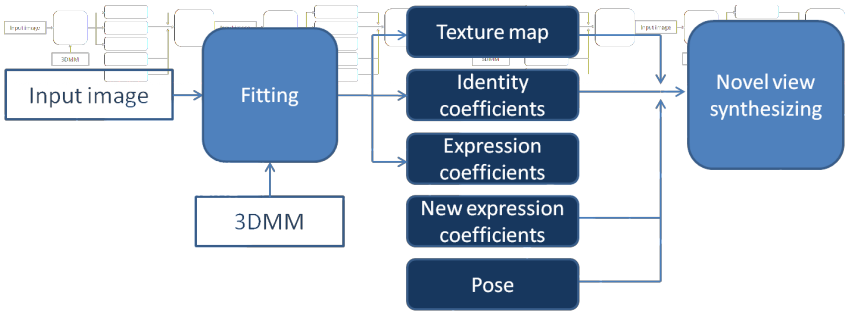


Figure 2.9: The framework for synthesizing a novel view of the system in [Chu *et al.* 2014].

Concluded by [Sandbach *et al.* 2012b], fitting process of 3DMM is robust to noise in the raw input and dense correspondence achieved. But variations allowed by model are restricted by range of data used to create it.

2.2.1.4 Conformal mapping

In 2003, [Gu *et al.* 2004] first proposed an approach able to compute the spherical conformal mapping for 3D surface analysis. The basic idea of this mapping is minimizing the harmonic energy through a nonlinear heat diffusion process. Then in 2008, [Gu & Yau 2008] extended their previous work and proposed a new method for calculating Riemann conformal mapping. More recently, [Yueh *et al.* 2015] employed the conformal mapping for morphing facial expressions. The pipeline of their method is as follows:

1. *Surface meshing.* In the first step, we assume that the correspondence of the boundary points of two different human faces S_a and S_b are known. Meanwhile m landmarks have been already selected for each facial expression. The matching process starts by projecting each surface to the unit disk $\mathbb{D}$. Then the matching function combining the optimal Möbius transformation and the global matching function (OMGMF) is solved by minimizing the least square error of landmarks on two unit disks, as shown in Fig. 2.10(c).
2. *Surface morphing.* Inspired by the mesh warping technique [Wolberg 1998], which partitions the 2D image into several pieces and interpolate them piece by piece, discrete surface geodesics that connect pre-selected feature points are further computed frame by frame. Using conformal mapping, surface geodesics can partition both facial surface and unit disk, as shown in Fig. 2.10(a) and 2.10(b).
3. *Surface registration.* In this stage, a one-to-one surface registration is first carried out on the unit disks. Furthermore, a piece-wise cubic spline homotopy of the mean curvatures and the conformal factors is used to generate the morphing sequence. The process is shown in Fig. 2.10(d)
4. *Surface reconstruction.* Finally, they utilize the mean curvatures and the conformal factors to reconstruct surfaces by solving the Laplace-Beltrami equations.

Fig. 2.10(e) and Fig. 2.10(f) showed the morphing sequence between two different facial expression, which demonstrated the feasibility of the proposed morphing

method.

Overall, the advantage of conformal mapping is that it can provide accurate dense correspondence, but it need a prior knowledge of boundaries of surfaces, and is computationally expensive.

Some other works using conformal mapping were [Rosato *et al.* 2008] and [Zeng *et al.* 2013]. Typically, [Rosato *et al.* 2008] utilized conformal mapping to obtain dense correspondences between facial meshes. In the framework of [Zeng *et al.* 2013], facial features (Mean curvature and conformal factors) were directly extracted from the unit disks projected by conformal mapping.

2.2.2 Feature extraction and representation for FER

According to [Sariyanidi *et al.* 2015], facial features for FER can be recognized spatial or spatio-temporal. Spatial features only consider still images, while spatio-temporal features additionally take into a neighborhood of frames.

However, from another perspective of encoded information similar as other pattern recognition, those features for FER can also be concluded by two main streams: texture and shape. Appearance based features highlight distinguishing between facial expressions using global or local texture information, whereas shape based features pay more attentions to geometrical surface measurements of faces, such as curvature, geodesic distance, etc. Moreover, in many cases for 2D or 3D FER, the parameters of employed statistical models are also considered as features of expressions. In the following parts, a detailed presentation is proposed in a systematic manner.

2.2.2.1 Spatial features

Gabor based feature is one of the most popular feature descriptors for facial biometric system. Many researchers employed Gabor based features for describing both 2D texture and 3D shape information. In the domain of FER, several approaches based on Gabor features have been proposed in [Lyons *et al.* 1998]. More specifically, [Lyons *et al.* 1998] utilize a multi-orientation, multi-resolution set of Gabor

filter to code facial expressions. Authors in [Wan & Aggarwal 2014] also employed Gabor representation as the feature of appearance.

LBP-like feature is another facial descriptor which has been widely employed in many biometric systems because of its remarkable effectiveness and robustness. Some methods using LBP-like features to discriminate expressions have been proposed. Precisely, [Shan *et al.* 2006] and [Shan *et al.* 2009] proposed the boost-LBP feature in their system, which utilized Adaboost to select the most discriminative LBP features. [Zhao *et al.* 2010] extracted multi-scale LBPs from faces fitted by the SFAM model. In [Happy & Routray 2015], LBP feature were extracted from salient facial patches. More recently, an expression-specific local binary pattern (es-LBP) was firstly introduced for FER in [Chao *et al.* 2015].

SIFT descriptor was also employed to extract discriminating information from morphological regions located by landmarks on both intensity and range images [Berretti *et al.* 2010]. Using SVM classifier, an average recognition rate of 77.5% was obtained on the BU-3DFE database.

Distance based features were widely employed since facial landmarks were more and more easier to be accurately located. One of the first works was [Soyel & Demirel 2007, Soyel & Demirel 2008], in which 6 characteristic distances between landmarks were served as discriminating features. [Wang *et al.* 2007] calculated distances between specific fiducial landmarks as the features on 2D modality in their system. The method developed in [Tang & Huang 2008b, Tang & Huang 2008a] utilized a pool of features composed of normalized Euclidean distances between 83 facial feature points. [Tekguc *et al.* 2009] computed totally $C_{83}^2 = 3403$ normalized distances between 83 landmarks and introduced NSGA II algorithm to select the most discriminating distance feature.A series of facial measurement based distance between landmarks was presented in [Tsalakanidou & Malassiotis 2010]. Moreover, [Li *et al.* 2010] extended the features proposed in [Soyel & Demirel 2007], which did not only compute six characteristic distances, but also considered eye slopes and angles.

Geometric properties based features explicit shape deformations, which are normally considered to be more effective to discriminate expressions since expectations

are caused by movement of facial muscles. In general, coordinates, gradients, normal vectors and curvatures are the most common 3D geometric features for FER. The method developed in [Wang *et al.* 2006] and [Rosato *et al.* 2008] employed the principal curvatures λ_1, λ_2 as the primitive features to symbolize geometric surfaces. [Wang *et al.* 2007] utilized Gaussian and mean curvatures to calculate HK label [Trucco & Fisher 1995] at each face region. The spatial coordinates of the facial feature points was employed to find the residue matrix, which could be further converted to feature matrix for FER classification [Srivastava & Roy 2009]. In [Ocegueda *et al.* 2011a], three types of expressive maps (geometry, normal and local curvature) were computed and further selected by a feature selection framework [Ocegueda *et al.* 2011b] before modeling the Markov Random Field. [Sha *et al.* 2011] proposed the distribution of surface curvature features (SCF) to represent skin surface deformation in the 3D expressional faces. The mean curvature was introduced in both [Zeng *et al.* 2013] and [Lemaire *et al.* 2011]. Moreover, [Li *et al.* 2012b] projected three normal components, in X,Y, and Z-plane respectively, into 2D representation. Multi-scale LBP-like features were then extracted for expression prediction. An average recognition rate of 80.14% was achieved on BU-3DFE database, which was better than most of the state-of-the-art results. A set of features with multiple differential quantities, including coordinate, normal, and shape index values, were extracted to describe the geometry deformation of each segmented region in [Zhen *et al.* 2015]. The support vector machine and the hidden Markov model were then used to predict the expression label in 3D and 4D, respectively. The experiments were conducted on the BU-3DFE and BU-4DFE databases, and the results achieved clearly demonstrated the effectiveness of the proposed method. Similar as [Zhen *et al.* 2015], [Yang *et al.* 2015a] adopted geometric scattering representation using a set of maps of shape features in terms of multiple order differential quantities, i.e. the Normal Maps (NOM) and the Shape Index Maps (SIM).

Curve Shape Analysis is another effective method for 3D FER system. The technique firstly proposed in [Joshi *et al.* 2007] provided an efficient representation for studying shapes of closed curves. Their framework allowed performing curve

shape analysis trough the computation of the distance between any two curves, that is the length of the geodesic path separating them in the space of closed curves. [Maalej *et al.* 2010] evaluated the effectiveness of this curves-based representation calculated from multiple selected patches on the 3D face model for 3D FER. The approach achieved promising results in a binary classification scheme.

As mentioned in the previous section, several method employed the parameters of statistical modeling as the features for discriminating expressions, such as ASM/AAM ([Lanitis *et al.* 1997, Abboud *et al.* 2004]), Bilinear model ([Mpiperis *et al.* 2008]), 3DMM ([Sandbach *et al.* 2012b]), and Conformal Mapping ([Yueh *et al.* 2015]). Here, some other modeling based approaches are presented. Specifically, the method developed in [Wang & Yin 2007] employed a topographic modeling technique to recognize and analyze facial expression from single static images. The experimental results showed that Topographic Context (TC) is a good feature representation for recognizing basic prototypic expressions. Another modeling technique was the Basic Facial Shape Component (BFSC) model [Gong *et al.* 2009]. Considering the nature BFSC was capable of modeling only neutral faces, therefore the subtraction of the depth map of BFSC from the depth map of the original aligned mesh provided the Expression Shape Component (ESC), that contained the expression information. The classification using expression components achieved a recognition rate of 76.2% on the BU-3DFE database.

2.2.2.2 Spatio-temporal features

Spatial features were normally designed to describe single or multiple static images for 3D FER. However, facial expressions caused by motions of facial muscles. Thus, some spatio-temporal features have been proposed to encode temporal information among a continue range of frames for identifying expressions.

One of the first work was proposed in [Yin *et al.* 2006a]. The pipeline of their method was as follows: Firstly, a 3D expression model's label distribution, called facial expression label map (FELM), was proposed for tracking face. After the facial model tracking and adaptation, the 3D motion trajectories were estimated by vectors from the tracked points of the current frame to the corresponding points

of the first frame with a neutral expression. Furthermore, a spatial-temporal facial expression descriptor was proposed as a combination of motion trajectories and label distributions tracked by the FELM. Finally, the feasibility of the method was evaluated on a dataset of 50 subjects.

[Sandbach *et al.* 2011, Sandbach *et al.* 2012a] proposed a Free-Form Deformations (FFDs) based motion-features. Firstly the 3D FFDs was employed for face tracking. Based on tracking result, we obtained the vector fields of displacement, which were further projected into 2D representations. A Quad-tree decomposition was then introduced to encode temporal information and construct features on the sequence of 2D projected images. At the stage of classification, authors utilized Gentle Boost (GB) classifiers [Friedman *et al.* 2000] for feature selection and HMMs for temporal modeling of sequences. The convincing experiment results suggested the use of the 3D information for FER.

A level curve based approach was proposed by [Le *et al.* 2011] to compare the shape of 3D facial models. The level curves were extracted directly from depth images using an arclength parameterized function. The Chamfer distance of these level curves was applied to measure the distances between the corresponding normalized segments. Combined with the proposed feature extraction method, an HMM-based classification algorithm yielded a high overall recognition accuracy of 92.22% on the BU-4DFE database.

More recently, a new automatic approach for 4D FER was proposed in [Amor *et al.* 2014], which extended their previous work [Maalej *et al.* 2010]. The pipeline of proposed method was as follows: firstly, the 3D deformation is captured based on a Dense Scalar Field (DSF) that represents the 3D deformation between two frames. LDA was then used to transform derived feature space to an optimal compact space to better separate different expressions. Finally, the expression classification was performed in two ways: (1) using the HMM models for temporal evolution; and (2) using mean deformation along a window with Random Forest classifier. Experimental results showed that the proposed approaches were capable of improving the state of art performance on the BU-4DFE database.

2.2.3 Feature selection and classification for FER

In the previous sections, we have roughly reviewed modeling techniques and features for FER. Here, we give a brief presentation of algorithms of feature selection, classification and temporal modeling.

Due to the fact that extracted features in proposed methods for FER were in high dimensions, some features selection techniques were usually employed to find the most discriminating characteristics and reduce the redundancy. As one of the most popular dimensionality reduction technique, PCA is also widely used in the procedure of feature extraction [Le *et al.* 2011, Soyel & Demirel 2008, Soyel & Demirel 2010]. Meanwhile, in [Yin *et al.* 2006a, Wang *et al.* 2006, Rosato *et al.* 2008], LDA is also commonly employed to find the directions which are most effective for discrimination by maximizing the ratio between the between-class and within-class scatters. Moreover, [Tekguc *et al.* 2009] made use of Non-dominated Sorted Genetic Algorithm II (NSGA II) [Deb *et al.* 2002] to select most discriminant features. Meanwhile, authors in [Sha *et al.* 2011] developed an algorithm, named normalized cut-based filter (NCBF), to reduce the graph's redundancy. A class-regularized locality preserving projection (cr-LPP) method was proposed by [Chao *et al.* 2015] for dimensionality reduction and subspace projection. However, Adaboost or Gentleboost is able to select features and predict expressions in a same framework, which also attracted many researchers [Tang & Huang 2008b, Shan *et al.* 2009, Maalej *et al.* 2010].

For static FER, SVM and its variants, i.e, multi-class SVM, is one of the most attractive classifier due to its simplicity and effectiveness [Shan *et al.* 2009, Tang & Huang 2008b, Li *et al.* 2012b]. Meanwhile, Probabilistic Neural Network (PNN) is another widely employed classifier in this field [Tekguc *et al.* 2009, Soyel & Demirel 2008, Li *et al.* 2010]. Moreover, Metric Learning [Wan & Aggarwal 2014], Bayesian Belief Net (BBN) [Zhao *et al.* 2010], Markov Random Field [Ocegueda *et al.* 2011a], and Random Forests [Amor *et al.* 2014] also demonstrated their power in FER systems

For dynamic FER, there are few temporal modeling technique except Hidden

Markov Models (HMMs) [Sandbach *et al.* 2012a, Le *et al.* 2011, Amor *et al.* 2014]. Specifically, [Sun & Yin 2008] investigated and compared the performances of a pool of HMMs: Temporal-HMM(T-HMM), Pseudo Spatio-Temporal HMM (P2D-HMM) and real 2D HMM (R2D-HMM). The experimental results showed that real 2D HMM (R2D-HMM) outperformed other two variants of HMMs for sequences classification.

2.2.4 New directions of FER

In this section, we present and discuss two new research directions respectively for FER: Deep Network based approaches, and micro-expression recognition.

2.2.4.1 Deep Network based approaches

Recently, deep learning techniques have shown the remarkable performances on many field of pattern recognition, such as digital identification, facial verification, object detection, etc. Normally, a large number of samples are required for training a deep network. Thus, the development of deep learning based FER system is restricted by the task of collecting enough data. However, more recently, some databases have been proposed available for deep learning, such as FER2013 [Goodfellow *et al.* 2013] and SFEW 2.0 [Dhall *et al.* 2011].

A number of deep networks, such as Multi-Layer Perceptron (MLP) [Rumelhart *et al.* 1988], Deep Belief Networks (DBN) [Hinton & Salakhutdinov 2006], Deep Convolutional Neutral Network (DCNN), etc., have been already introduced to identify facial expression. Thus, we introduce some of most representative work based on deep networks in the following. Authors in [Liu *et al.* 2015a] proposed a novel deep architecture, AU-inspired Deep Networks (AUDN), inspired by the psychological theory that expressions can be decomposed into multiple facial Action Units (AUs). The method proposed in [Jung *et al.* 2015] integrated two deep networks: the deep temporal appearance network (DTAN) and the deep temporal geometry network (DTGN). The DTAN, which was based on CNN, was used to extract the temporal

appearance feature necessary for FER. The DTGN, which was based on DNN, catched geometrically moving information from the facial landmark points. Moreover, a classification module with the ensemble of multiple deep convolutional neural networks (CNN) was proposed in [Yu & Zhang 2015]. Each CNN model was initialized randomly and pre-trained on a larger dataset. The proposed method generated state-of-the-art result on the FER dataset and also achieved 55.96% and 61.29% respectively on the validation and test set of SFEW 2.0.

However, Recurrent Neural Networks(RNNs), particularly, LSTM, have demonstrated their effectiveness of processing and predicting sequences. [Graves *et al.* 2008] made a baseline which use a recurrent network to extract the temporal dependencies in the image sequences. Authors in [Khorrami *et al.* 2016] presented a system that performed emotion recognition, which combined CNNs with RNNs.

However, although a set of deep network based approaches for FER have been proposed. But the performance of those method for recognizing facial expression in a wild environment still need improvements. To fix this problem, a workshop, Emotion Recognition in the Wild Challenge (EmotiW), has been held once a year since 2013.

2.2.4.2 Micro-expression recognition

Facial micro-expression represents genuine emotions that people try to conceal. Generally, the duration of micro-expression is shorter than 0.5s. Thus, the recognition of such short-lived subtle expressions is a challenging task. Another reason of little attentions paid to micro-expression recognition in the past is lacking of well-established databases. However, due to a range wide of applications, such as lie detection, more and more researchers have been interested in developing algorithms for micro-expression recognition.

More recently, along with the publication of several available micro-expression databases, such as CASME [Yan *et al.* 2013], CASME2 [Yan *et al.* 2014a], and SMIC [Li *et al.* 2013], some benchmarks have been proposed. Specifically, in [Wang *et al.* 2014a], authors treated a gray facial image as a sec-

ond order tensor and introduced the discriminant tensor subspace analysis (DTSA) [Wang *et al.* 2011] to reduce dimensionality.

LBP-TOP has demonstrated its simplicity and efficiency for facial expression recognition, thus, LBP-TOP and its variants also have been widely in micro-expression analysis [Pfister *et al.* 2011, Yan *et al.* 2014b, Wang *et al.* 2015]. [Wang *et al.* 2015] employed LBP-TOP to extract features on a Tensor Independent Color Space. Authors in [Huang *et al.* 2015] utilized an integral projection method to obtain horizontal and vertical projection and employed LBP operators to extract the appearance and motion features on those projections. Meanwhile, a LBP with Six Intersection Points (LBP-SIP) volumetric descriptor was proposed in [Wang *et al.* 2014b]. Moreover, [Huang *et al.* 2016] modified the way of quantization of LBP, including information of sign, magnitude and orientation components to improve the performance.

More recently, [Liu *et al.* 2015b] applied a robust optical flow method on micro-expression video clips and proposed a ROI-based, normalized statistic feature (MDMO) that considered both local statistic motion information and its spatial location. Experimental results showed that the MDMO can achieve better performance than two state-of-the-art baseline features, i.e., LBP-TOP and HOOF.

Overall, although several methods have been proposed for micro-expression recognition, there still need some improvements. Trying existing method for FER or proposing novel approaches are both possible ways. Considering the fact that 3D can provide more effective features than 2D face, to our best knowledge, there has no method still now developed on 3D face for micro-expression recognition. The biggest reason is lacking of real-time capturing techniques for 3D scans, because the duration of micro-expression is always shorter than 0.5s.

2.2.5 Summary

In the previous sections, we have reviewed all aspects of FER techniques in the literature. Firstly, several most representative modeling techniques are introduced and their advantages and disadvantages are also analyzed. Secondly, we categorize existing facial features for facial expression recognition in the literature. Some works

with specific features are also listed and presented. Moreover, we make a brief review of algorithms for features selection and classification. Finally, some new research directions in the field of facial expression recognition are further discussed.

As mentioned previously, there already have some comprehensive surveys in this area by [Fasel & Luettin 2003, Zeng *et al.* 2009, Fang *et al.* 2011, Bettadapura 2012, Sandbach *et al.* 2012b, Danelakis *et al.* 2015].

In fact, the development of facial expression recognition technique always depends on available databases. Thus, we list popular databases for analysis of facial expression and micro-expression, respectively, in Table. 2.3 and 2.4.

Table 2.3: Overview of databases for facial expression classification in the literature. S/D: Static or dynamic data. Size:Number of subjects. [Sandbach *et al.* 2012b]

Database	2D/3D	S/D	Size	Content
PIE	2D	S	68 subjects	Neutral and 4 exps: smile, blink, talking, without Glasses
Cohn-Kanade	2D	D	97 adult	full AUs
MMI	2D	D	75 adults	full AUs
FER2013	2D	S	35887 image	6 basic expressions
SFEW	2D	S	700 images	6 basic expressions
AFEW	2D	D	957 video clips	6 basic expressions
BU-3DFE	3D	S	100 adults	6 basic expressions at 4 intensity levels 83
BU-4DFE	3D	D	101 adults	6 basic expressions
BP4D-Spontanous	3D	D	41 adults	8 tasks
Bosphorus	3D	S	105 adults	24 AUs, neutral, 6 basic exps, occlusions 24
D3DFACS	3D	D	10 adults	Up to 38 AUs per subject
CASIA	3D	S	123 adults	Neutral and 5 exps: smile, laugh, A, Su, eyes closed
Gavdb	3D	S	61 adults	3 exps: open/closed smiling and random

Table 2.4: Overview of databases for micro expression classification in the literature. P/S: Posed expression or spontaneous expression

Database	P/S	Content
USF-HD	P	100 micro-expressions
YorkDDT	S	18 micro-expressions
SMIC	S	77 micro-expressions
CASME	S	195 micro-expressions
CASME II	S	247 micro-expressions

2.3 Conclusion

In this chapter, we review the state of the art for two issues: facial gender and ethnicity classification, as well as facial expression recognition.

For the former, we firstly review the existing methods from three perspectives: 2D texture, 3D shape,and multimodalites. Some famous feature descriptors have been introduced in details. Moreover, we also list available databases for facial gender and ethnicity classification in the last part. From this survey, we could observe that the existing methods have achieved some adorable results and local feature descriptors show their potential advantages for this issue.

For the latter, we also survey the state of the art from three perspectives: 2D texture, 3D shape, and multimodalites. Meanwhile, some well-known modeling techniques for facial expression or facial surface deformation are presented in great detail. In the same way, the available databases for facial expression analysis are further listed. Reviewing the literature, we could find that the majority of existing work are single modality based, which still could be improved using multimodalites based techniques.

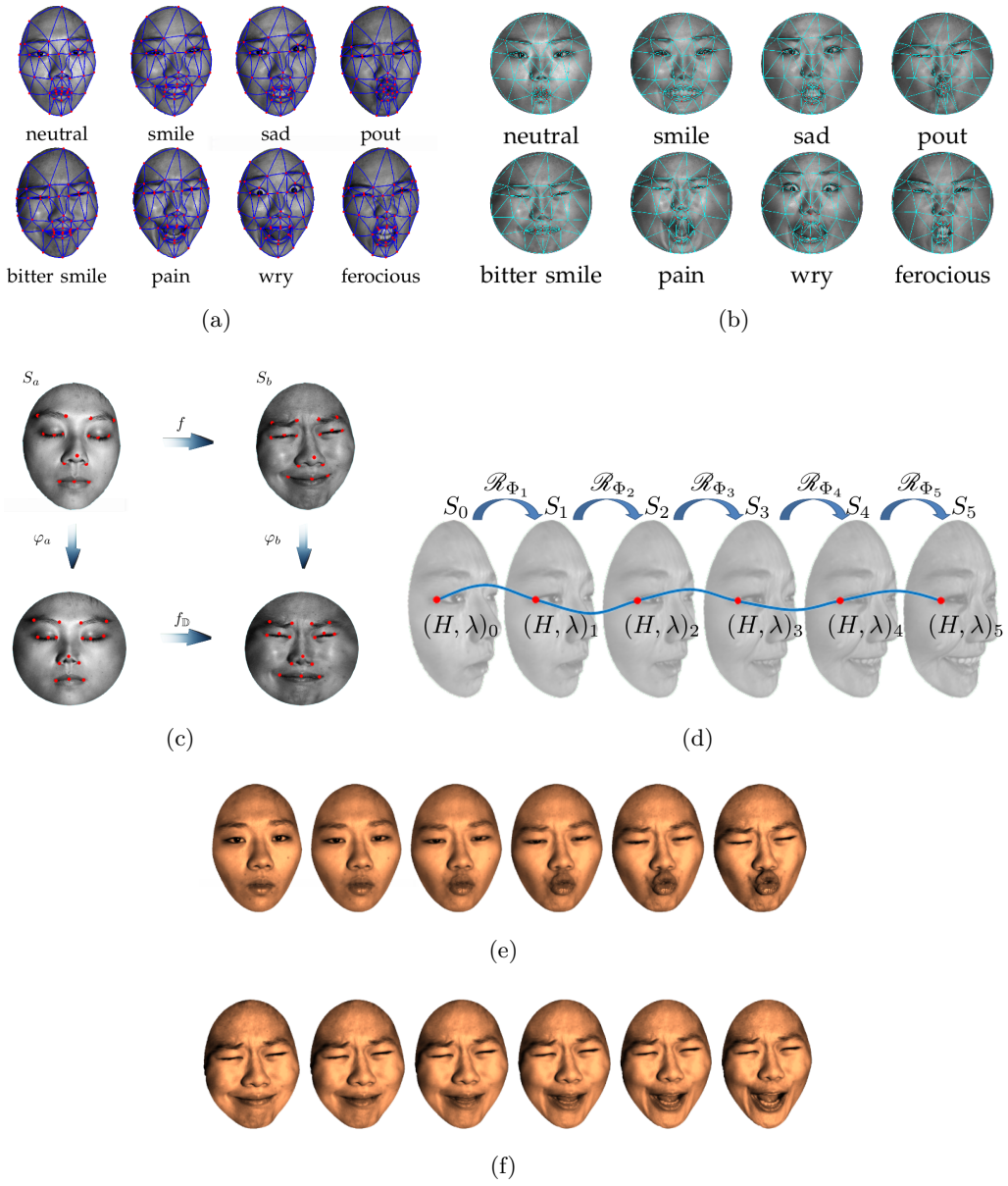


Figure 2.10: (a) The geodesics on a human face with different facial expressions. (b) The partition mesh of the unit disk for different facial expressions. (c) The idea of surface matching. (d) The cubic spline homotopy of the mean curvature and conformal factor of a vertex. (e) The morphing sequence of eye blinking. (f) The morphing sequence of mouth opening. [Yueh *et al.* 2015]

Chapter 3

Facial Soft biometric estimation: gender and ethnicity

3.1 Introduction

As we discussed in the previous chapters, during the last decade, research on face-based gender and ethnicity classification has grown up rapidly since it emerged. Actually, most efforts in the literature have been made within the 2D domain using texture information, as people from different genders and ethnicities commonly have a diversity of face textures.

However, due to the development of 3D imaging technologies, 3D shape information of human faces can be easily captured, which facilitates the advance in 3D shape-based approaches. 2.5D and 3D facial scans have regarded as a major solution to deal with these unsolved issues in 2D intensity image based face recognition, i.e. illumination and pose changes. Meanwhile, according to the anatomical studies, 3D geometrical information of faces of human beings also reflects distinctions among ethnics and genders, and is thereby essential for gender and ethnicity classification as well. For example, faces of white people and male are commonly craggier than that of Asian people and female. Caucasian brow bones are always deeper, with eyes more sunken than Asian ones; while Asian noses tend to possess lower bridges; Caucasian noses extend slightly upward. Thus, several 3D shape information based methods have also been further proposed for solving this problem.

More currently, since most of the current 3D imaging systems deliver 3D face

models along with their aligned texture counterparts, a major trend in the literature of face recognition is to adopt both the 3D shape and 2D texture based modalities, arguing that the joint use of these two clues can generally achieve more accurate and robust accuracy than using only either of the single modality [Huang *et al.* 2011b]. Many researchers believe that fusion of 2D and 3D data will improve the classification accuracy in the classification of gender and ethnicity; nevertheless, very limited research has investigated this topic using multiple-modalities. [Lu *et al.* 2006] can be regarded as the pioneer for this attempt where they integrated similarity measurements of texture and shape (i.e. intensity and depth value of the central part of human faces), showing that the combination of multiple modalities leads to an improvement in both the accuracies of gender and ethnicity classification. This work is somewhat intuitive, and it has several downsides. For example, the direct use of pixel values of facial intensity and range images cannot sufficiently discriminate the difference between male and female or between various ethnic groups, and it also tends to incur the sensitivity to illumination for the texture modality. In addition, they treat the entire face area equally, which is actually inappropriate.

Similarly, in this work, our basic assumption, as the one behind multi-modal face analysis, is that the result of single modality (i.e. only 2D texture or 3D shape) based techniques can be ameliorated by combining various clues from different modalities. Based on this idea, in this chapter, we propose a novel method especially for facial ethnicity classification, and another novel method for facial gender and ethnicity classification. Both of them make use of a novel facial representation for this issue respectively, along with a boosted framework for feature selection on two modalities: 2D texture and 3D shape.

Our main contributions can be summarized as follows.

(1). Inspired by the fact that local feature-based method has the potential advantage of being robust to facial expression, pose and lighting changes and even to partial occlusions, an ethnicity classification method that uses a novel local feature, Oriented Gradient Maps (OGMs), to describe ethnicity related local texture and shape characteristics is firstly presented.

The OGMs is a biological vision based representation originally proposed

for 3D face recognition [Huang *et al.* 2012b], which achieved the state-of-the-art performance and proved insensitive to affine illumination and geometrical transformations. The proposed facial representation is inspired by the study of [Edelman *et al.* 1997], who proposed a representation concept of complex neurons in primary visual cortex. These complex neurons respond to a gradient at a particular orientation and spatial frequency, but the location of gradient is allowed to shift over a small receptive field rather than being precisely localized. For this task, OGMs is used to extract local details at pre-defined quantized orientations to highlight the distinctiveness in both modalities.

(2). Motivated by the simplicity, effectiveness and low computational complexity of LBP on 2D texture based facial gender and ethnicity classification, a novel variant of LBP, Local Circular Pattern (LCP), is presented for this issue.

Inspired by some recent studies on local feature based face recognition [Lei *et al.* 2014, Vu & Caplier 2012], in contrast to the original pixel values of facial texture and range images holistically used in [Lu *et al.* 2006], we investigate the way to represent the information of texture and shape in a local feature space with more discriminative power, aiming to minimize within-class variations and maximize between-class similarities. Due to its tolerance to monotonic lighting changes as well as computational simplicity, LBP is regarded as one of the most effective and successful local descriptors in many fields including texture analysis, facial image analysis, image and video retrieval, environment modeling, visual inspection, motion analysis, biomedical image analysis, aerial image analysis, and remote sensing [Huang *et al.* 2011a]. However, LBP still has several main limitations, such as the insufficiency in discriminative power and the insensitivity to noise. In this work, rather than explicitly quantizing the sign and magnitude components of local patterns as adopted in LBP and its variants, we propose to quantize local patterns through clustering, and the descriptor is called local circular patterns (LCP for simplicity). Compared with the binary quantization scheme, clustering based quantization can generate better approximation with less distortion, therefore LCP possesses a greater ability in discrimination and is less sensitive against noise. Moreover, the quantization accuracy can be manageable through modifying

the number of clusters. Additionally, in clustering based quantization, the parameters are tuned using training data, and it thus can deal with various local pattern distributions.

(3). To select most discriminate feature for distinguishing different gender and ethnic groups and integrate the information extracted respectively from 2D texture and 3D shape, a novel boosting based framework is presented.

Among various feature selection techniques proposed in the community of machine learning [Guyon & Elisseeff 2003], we make use of the wide-spread Adaboost algorithm, to select a compact subset of facial features from the entire multi-modal feature set. The reason to employ Adaboost is that it is capable of obtaining a strong classifier through combining several weak ones while selecting features, as a result there is no need to retrain a classifier for gender or ethnicity label prediction. On the one hand, the features extracted from various facial regions (such as the ones of eyes, nose, forehead) represented as textures or shapes, highly related and discriminative to the task of facial gender and ethnicity classification, can be determined and assigned with different weights according to their importance to final performance. The relevance of the features is thereby largely decreased, and the combination of these selected ones tends to improve the classification accuracy. On the other hand, the entire feature set generally contains redundant information, and utilizing all the features is also time and memory expensive which probably gives rise to the problem of curse of dimensionality. After feature selection, the dimensionality of the feature can be reduced and the efficiency of classification can be increased.

Besides, some minor contributions have also been made for further analyze the issue of facial gender and ethnicity classification. For example, we evaluate the contribution of each organ of human faces to the accuracy of ethnicity classification and analyze the impact of the percentage between training and testing samples to final performance.

The remainder of this chapter is organized as follows. Section 3.2 show the framework of the proposed system. The proposed two facial descriptions, OGMs and LCP, are shown in section 3.3, and section 3.4 presents the boosting based

local feature selection algorithm. Experimental results of both OGMs and LCPs based methods are described and analyzed in section 3.5. Section 3.6 concludes this chapter.

3.2 Overview of the proposed system

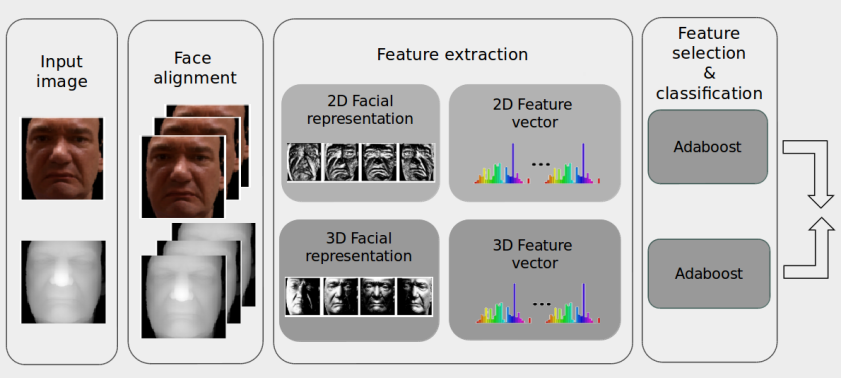


Figure 3.1: The overview of proposed system

In fact, in this work, we propose two novel multi-modalites based methods for facial gender and ethnicity classification, both of which involve the same framework of system, but use different facial representations. As shown in Figure 3.1, the pipeline of proposed methods is defined as follows:

1. Firstly, given 3D face scans and their correspondent textures images, a standard preprocessing is carried out to remove spikes and fill holes. Then, an algorithm of Iterative Closest Points (ICP) is further employed for face alignment. Furthermore, we project 3D face models to depth images and texture images using Zbuffer algorithm, which are regarded as the inputs of feature extraction.

2. Secondly, two facial representations (OGMs and LCPs) are respectively extracted to represent local shape and texture variations from 2D texture images and 3D range images. Then, histograms of facial representations are further computed as facial features hierarchically.

3. Thirdly, Adaboost is used to select the most discriminative features from the high dimensional ones, while boosting weak classifiers into a strong one. Decision level fusion is made for final decision.

3.3 Facial representations

Actually, an effective method for gender and ethnicity classification generally addresses two important problems: extraction and selection of gender and ethnicity related features. In this section, we present the processing of feature extraction in details. Considering the simplicity and effectiveness of LBP for this issue in the state of the art, two facial representations, which extend the idea of LBP, are presented here.

3.3.1 Oriented Gradient Maps

To highlight the discriminative power of local texture and geometry clues of human faces for ethnicity analysis, we introduce OGMs, a biological vision based representation method originally proposed for 3D face recognition [Huang *et al.* 2012b], in which it achieved the state of the art performance and proved insensitive to affine illumination and geometrical transformations. Since the OGMs simulate the response of complex neurons to gradient information within a given neighborhood, they are able to describe local texture changes of 2D facial maps and local shape changes of 3D facial maps at the same time.

The pipeline of extraction of OGMs feature is defined as follows: Firstly, as introduced previously, given a textured 3D face model, through the preprocessing pipeline including removing spikes and filling holes, we can extract a facial range image and its texture counterpart for the following steps. Then, given a range or texture image I, for each pre-defined quantized orientation o, a certain number of gradient maps $G_1, G_2, \ldots, G_o$, which describes gradient norm of input image at direction o, are first computed as in Eq. 3.1.

$$G_o = (\frac{\partial I}{\partial o})^+ \tag{3.1}$$

We then convolve gradient maps with a Gaussian kernel G for avoiding abrupt changes. The standard deviation of the Gaussian kernel G is proportional to the radius of the given neighborhood area R, as in Eq. 3.2.

$$\rho_o^R = G_R \times G_o \tag{3.2}$$

Thirdly, according to Eq. 3.3, the response vector $\rho^R(x, y)$ of pixel location (x, y) can be built by collecting all the values of the convolved gradient maps at that location. The response vector is further normalized to a unit norm vector denoted by ρ^R .

$$\rho^R(x, y) = [\rho_1^R(x, y), \ldots, \rho_o^R(x, y)]^t \tag{3.3}$$

Finally, an OGM J_o is generated as ρ^R at orientation o, as shown in Fig. 3.2.

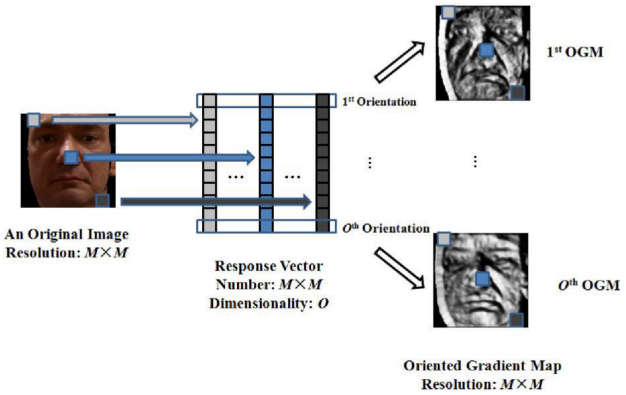


Figure 3.2: Illustration of the oriented gradient maps; each is for a quantized orientations o [Huang *et al.* 2012b].

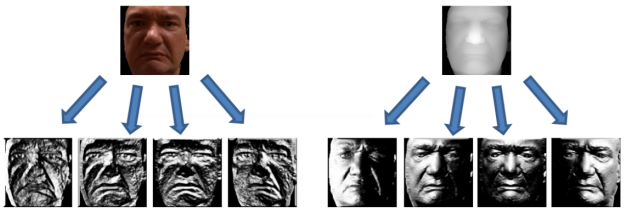


Figure 3.3: Example of oriented gradient maps (at 4 quantized orientations) of a facial range image and its corresponding texture.

Fig. 3.2 presents such a process, and from Fig. 3.3, we can see that local shape

and texture changes of human faces are both highlighted by OGMs. Compared with face recognition that requires more discriminative features to tolerate much smaller inter-class changes caused by similar facial appearances, while in the given task, inter-class variations tend to be larger, and we thus generate four OGMs for each facial range or texture images instead of eight in [Huang *et al.* 2012b] to increase system efficiency. Specifically, for each of input images, we consider its gradient maps at following orientations respectively: 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, $\frac{3\pi}{4}$, which conserve both positive and negative polarity.

3.3.2 Local Circular Patterns

Since local circular patterns (LCP) can be regarded as a variant of local binary patterns (LBP), we will first recall the basic concept of the original LBP descriptor, and then introduce the proposed LCP and LCP based facial representation.

A. The LBP methodology

A basic LBP operator simply thresholds a 3×3 neighborhood by the value of the central pixel, and the sign of thresholded neighboring values can form a binary number, which is then transformed into a decimal number. This decimal number is treated as the label of the central pixel (Fig. 3.4 (a)). We call this quantization scheme as binary quantization in the following. The histogram of the labels within a region is often used as a texture descriptor.

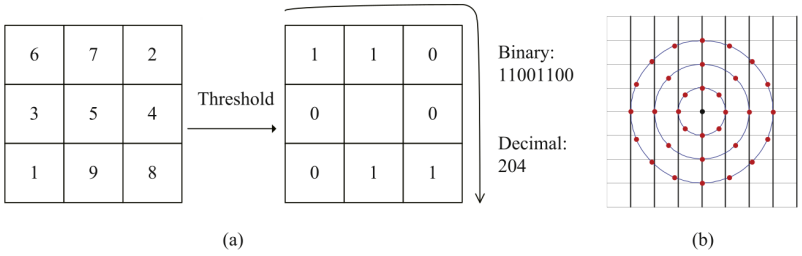


Figure 3.4: LBP Operators: (a) Basic LBP; (b) Multi-resolution LBP.

The basic LBP was later extended to multi-resolution and "uniform" [Ojala *et al.* 2002]. Multi-resolution denotes that LBP can operate on any radius R and any number of pixels P within the neighborhood, as shown in Fig. 3.4

(b). The uniform pattern is defined as a local binary pattern which contains at most two transitions between 0 and 1. The extended LBP is notated as $LBP^{u2}_{P,R}$, indicating that the operator works in a (P, R) neighborhood, and employs only uniform patterns and labels all the remaining patterns with a single bin. The authors of [Ojala *et al.* 2002] pointed out that uniform patterns are fundamental patterns providing the vast majority of all 3×3 patterns present in the observed textures.

However, LBP still has some limitations. One of the most critical limitations is that it only extracts the sign between neighboring pixels, while ignoring the magnitude, leading to the deficiency in discrimination. To better describe local micro patterns, various improvements have been explored [Huang *et al.* 2006, Tan & Triggs 2007, Guo *et al.* 2010b, Guo *et al.* 2010a]. In spite of the performance which is better than that of the LBP descriptor, these aforementioned variants are mostly artificially designed, and to comprehensively encode useful information, more bits are required which tends to incur a large increase in the memory and computational expense. Another limitation of LBP lies in that its binary coding scheme is very sensitive to noise, and once a single bit of the code alters, the resulting decimal number changes seriously. In order to improve the robustness to noise, several variants of LBP have been presented [Yang & Wang 2007, Liao *et al.* 2009]. Most of them intrinsically inherit the binary coding scheme of the original LBP, and cannot completely solve this problem. Besides, the distribution of local patterns within images taken under different scenarios varies greatly, for example in scene images and face images, and using the same quantization parameters (e.g. the same threshold) is therefore unsatisfied.

B. Local circular patterns

According to the analysis on the properties of LBP as well as its variants, we find out that the two vital limitations above are mainly caused by the binary quantization scheme. In this study, rather than explicitly quantizing the sign or/and magnitude components of local patterns, we propose to make the quantization of local patterns through clustering, aiming to generate better approximation with less distortion and thus leading to improvement in discriminative power and robustness to noise.

Specifically, as illustrated in Fig. 3.4 (b), for each pixel i_c whose gray value is t with its P neighboring pixels i_n, $n = \{1, 2, \ldots, P\}$ (gray values are denoted as $\{t_1, t_2, \ldots, t_P\}$) located on the circular neighborhood at the radius of R, the corresponding code p of this local circular pattern is defined as $p(LCP_{P,R}) = (t_1 - t, t_2 - t, \ldots, t_P - t)^T$. Given N training local circular patterns p_i, $i = 1, 2, \ldots, N$, the K-means clustering algorithm is performed to find a partition $C = \{c_1, c_2, \ldots, c_k\}$ by minimizing the following function:

$$J(C) = \sum_{i=1}^{k} \sum_{p_i \in c_i} D(p_j, \mu_i) \tag{3.4}$$

where $D()$ represents the distance function, and μ_i is the center of c_i. Then a new local circular pattern p' can be quantized into the nearest cluster center.

$$l(p') = \arg\min_i D(p', \mu_i) \tag{3.5}$$

K-means is a greedy algorithm which can only converge to a local minimum, but the recent study has shown that K-means can converge to the global optimum with a large probability when clusters are well separated [Meilă 2006].

Two main issues associated with K-means clustering are the number of clusters as well as the distance metric. The number of clusters k controls the balance between descriptive power and sensitivity to noise. Reducing the number of clusters increases the distances among them, and little vibration of the pattern does not change its quantization, but the low number of clusters also reduces the descriptive power. K-means can be performed with various distances $D()$, and two of them are introduced in this study, namely the Euclidean distance ($L2$) and city block ($L1$) distance. Given a local circular pattern $p' = (p'_1, p'_2, \ldots, p'_P)$, the Euclidean distance between p' and a cluster center $\mu = (\mu_1, \mu_2, \ldots, \mu_P)$ is defined as:

$$D_{L2}(p', \mu) = \sqrt{\sum_{i}^{P} (p'_i - \mu_i)^2} \tag{3.6}$$

and their city block distance is defined as:

$$D_{L1}(p', \mu) = \sum_{i=1}^{P} |p'_i - \mu_i| \tag{3.7}$$

Euclidean distance is usually used for computing the distance between points and cluster centers. The clusters found by K-means with Euclidean distance are spherical or ball-shaped. K-means with city block distance was proposed in [Kashima *et al.* 2008]. Compared with L2 distance, L1 distance is computationally more efficient. Each cluster center in the L1 distance case is calculated as the component-wise median of the points in that cluster. According to the clustering process, cluster centers obtained with L1 distance are all integers, while the ones obtained with Euclidean distance may be with decimals. In the following, in order to simplify the description, we call the LCP descriptor quantized by K-means with L2 and L1 distances LCP-L2 and LCP-L1 respectively.

C. LCP based facial representation

As we know, when LBP operates on the images formed by light reflection, i.e. 2D images, it can be used as a texture descriptor. Each of the LBP codes can be regarded as a micro-texton. Local primitives codified by the bins include different types of curved edges, spots, flat areas, etc. Meanwhile, as LBP works on range images which are based on depth information, it can also describe local shape structures [Huang *et al.* 2012a], such as flat, concave and convex. Similar to LBP, to comprehensively represent facial texture and shape images, we follow the scheme proposed by [Ahonen *et al.* 2004] for 2D face recognition. The basic idea lies in that a face image can be considered as a composition of the micro-patterns described by an LBP-like descriptor. One can build an LBP-like histogram computed over the entire facial image. However, such a representation only encodes the occurrences of micro-patterns without any indication about their locations. In addition, to consider the configuration information of faces, face images can be divided into a certain number of local regions, from which local LBP-like histograms can be extracted. These histograms are then concatenated into a single, spatially enhanced feature vector. The resulting histogram encodes both the local texture and global

(a) (b)

Figure 3.5: Region division scheme.

shape of face images. In our case, as shown in Fig. 3.5, both 2D texture and 3D range images are aligned based on eye outer corners, and cropped by an average mask. They are then divided into m (to be fixed experimentally) rectangular regions. For each region, histograms of clustering quantization based LCP are extracted which are further concatenated into a single histogram as gender and ethnicity related features in both the texture and shape modalities.

D. Multi-scale extension

Some LBP histogram-based applications change the neighborhood of the LBP operator for improved performance. By varying the value of radius R, the LBP codes of different resolutions are obtained. The multi-scale strategy was originally used for texture classification [Ojala *et al.* 2002], and it was also introduced to 2D face recognition [Yan *et al.* 2007, Chan *et al.* 2007]. In [Shan & Gritti 2008], Shan and Gritti studied MS-LBP for facial expression recognition by firstly extracting MS-LBP histogram-based facial features and then using the AdaBoost algorithm to learn the most discriminative bins. They reported that the boosted classifiers of MS-LBP consistently outperform those based on single-scale LBP, and the selected LBP bins distribute at all scales. MS-LBP can thus be regarded as an efficient method for facial representation. When considering it in multi-modal facial gender and ethnicity classification, this multi-scale technique can be applied to enhance the descriptive power of LCP as well.

3.4 Feature selection and decision level fusion

OGMs and LCP histogram based features extracted from various sub-regions of facial texture and range images have different discriminative abilities to distinguish between genders and ethnicities, and thus present non-equal contributions to the final classification accuracy. Moreover, the manner for facial representation by using the division scheme tends to incur a very high dimensional feature space leading to expensive time and memory cost, and may even give rise to the problem of curse of dimensionality. To overcome these shortcomings, the step of feature selection is necessary. In this study, the Adaboost algorithm is exploited to select a compact subset of features from the whole feature set. The reason to use Adaboost for feature selection is that it is able to train a strong classifier while selecting features, and there is thus no need to retrain a classifier in the process of label prediction. For histogram based features, we can either treat each bin in the histograms or an individual histogram as a single feature. Considering that the differences between genders or ethnic groups lie in discrimination of certain local patterns rather than all of them, therefore, we apply Adaboost to choose a set of discriminative bins as in [Shan & Gritti 2008, Zhao & Pietikainen 2008] that also employ it to select LBP-like features.

Adaboost, originally proposed by [Freund & Schapire 1995], iteratively selects a small number of weak classifiers whose performances are just better than random guess, and boosts them into a strong classifier. Viola et al. [Viola & Jones 2001] employed a variant of Adaboost to do face detection, and proposed the first real-time face detection algorithm. A distribution on the training samples is maintained, and in each iteration, weak classifiers are trained based on each feature according to the distribution. The classifier with the lowest weighted error is selected, so the corresponding feature is chosen in this iteration. We make use of Viola's variant of Adaboost to select a subset of histogram bins for gender and ethnicity classification. Details of the Adaboost can be found in [Viola & Jones 2001], but in order to maintain consistence of this work, the algorithm is posted in Algo. 1.

Algorithm 1 The Adaboost algorithm for feature selection.

INPUT:

Given example images $(x_1, y_1), \ldots, (x_n, y_n)$ where x_i stands for a sample, and $y_i = 0, 1$ for negative and positive examples respectively

OUTPUT:

1: Initialize weights $\omega_{1,i} = \frac{1}{2m}; \frac{1}{2m}$ for $y_i = 0, 1$ respectively, where m and l are the number of negatives and positives respectively;

2: **for** $t = 1, \ldots, T$ **do**

3: Normalize the weight,

$$\omega_{t,i} = \frac{\omega_{t,i}}{\sum_{j=1}^{n} \omega_{t,j}}$$

so that ω_t is a probability distribution.

4: For each feature, j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to $\omega_i, \varepsilon_j = \sum_i \omega_i \|h_j(x_i) - y_i\|$

5: Choose the classifier, h_t, with the lowest error ε_t.

6: Update the weights:

$$\omega_{t+1,i} = \omega_{t,i} \beta_t^{1-\varepsilon_i}$$

where $\varepsilon_i = 0$ if example x_i is classified correctly, $\varepsilon_i = 1$ otherwise, and $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$

7: **end for**

8: **return** The final strong classifier:

$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha_t = \log \frac{1}{\beta_t}$.

The weak classifier $h_j(x)$ is defined as

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) > p_j \theta_j \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

where the parity p_j controls the direction of the inequality between feature f_j and the threshold θ_j. The threshold θ_j is calculated as the average of weighted centers of positive and negative samples' features.

Adaboost is performed for both texture and range image features, and results in two strong classifiers $h(x)$ and $h'(x)$, one for each modality. As shown in Algo. 1,

these two classifiers are defined as below:

$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

$$h'(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha'_t h'_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha'_t \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

During testing, decision level fusion is performed. With the output of $h(x)$ and $h'(x)$, the final decision is made according to

$$H(x) = \begin{cases} 1 & \sum_{t=1}^{T} (\alpha_t h_t(x) + \alpha'_t h'_t(x)) \geq \frac{1}{2} \sum_{t=1}^{T} (\alpha_t + \alpha'_t) \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

Even though, Adaboost was proposed to solve the two-class problem as the one of gender classification, i.e. distinguishing male from female, it can also deal with the multi-class problem, e.g. ethnicity classification, by training a strong classifier for every two classes. In the test phase, the probe is decided by each of the strong classifiers learnt in the training phase, and is predicted with the label of class which has the proximal similarity measurement. Recently, some variants of Adaboost have been investigated to classify multiple classes [Zhu *et al.* 2009, Kim *et al.* 2011], and they can be exploited as well.

3.5 Experiments

In order to evaluate the effectiveness of the proposed approaches in the task of multi-modal facial gender and ethnicity classification, experiments are carried out on the FRGC v2.0 and BU-3DFE databases.

Gender classification is a typical binary classification problem, but the one for ethnicity classification is generally not the case. However, since the distribution of 3D face samples in current public databases (including FRGC v2.0) with ethnicity labels available is generally unbalanced, we also treat the task of ethnicity classifi-

cation as a binary classification problem, in the same way as the previous studies do [Lu *et al.* 2006, Toderici *et al.* 2010].

OGMs feature-based approach is our early work, which is only designed to classify ethnicities, while LCP feature-based approach is able to recognize both gender and ethnicities. However, for both of them, we mainly evaluate their performance on FRGC V2.0 database and verify their ability of generalization on BU3D-FE database.

In the following sections, we introduce the datasets and corresponding results subsequently.

3.5.1 Databases

FRGC v2.0 [Phillips *et al.* 2005] is one of the most comprehensive datasets publicly available for 3D face analysis. It contains 4007 textured 3D face models of 466 subjects, and each face model is made up of a 3D point-cloud and its 2D texture counterpart. Among the subjects, 22% are Asian, 68% are white, and 10% are others (a detailed ethnicity distribution is shown in table 3.1.). While 57% are male and 43% are female, with the age distribution: 65% 18–22 years old, 18% 23–27 and 17% 28 years or over.The database was collected during the 2003–2004 academic year, and contains time and illumination variations. Expressions such as "Neutral", "Happiness", "Surprise", "Disgust", "Sadness", and "Other" are included in the database as well.

Table 3.1: The ethnic distribution in the complete FRGC V 2.0 database

	Scans	Subjects
Asian	1121	99
White	2554	319
Hispanic	113	13
Asian-Middle-Eastern	16	1
Asian-Southern	78	2
Black-or-African-American	28	16
Unknown	97	16
Total	4007	466

BU-3DFE [Wang *et al.* 2006] is also one of the most popular databases in 3D face analysis, especially for 3D facial expression recognition. It contains 100 subjects

among which 56 are female and 44 are male, ranging from 18 to 70 years old. Among those subjects. there are 51 Whites, 24 East-Asians and others (more details are shown in Table 3.2). All individuals are asked to perform six prototypic expressions. Each includes four levels of intensities, and there are hence 25 instant 3D expression models for each subject (plus one model with a neutral expression), leading to 2500 models in total.

Table 3.2: The ethnic distribution in the complete BU-3DFE database

	Scans	Subjects
East-Asian	600	24
White	1275	51
Black	225	9
Middle-East Asian	50	2
Indian	150	6
Hispanic-Latino	200	8
Total	2500	100

3.5.2 Experiments: OGMs based approach

As introduced previously, to evaluate the proposed OGMs based method, we mainly design and carry out the experiments on the FRGC V2.0 database. Then, we verify the ability of generalization of the proposed method on the BU3D-FE database. Thus, we introduce the experiments on two databases subsequently.

3.5.2.1 OGMs based approach on FRGC

Recalling that FRGC V2.0 contains 99 Asians (1121 facial images) and 367 non-Asians (2886 facial images), as most of the tasks in the literature, due to lacking labels of minority, we consider ethnicity classification as a binary classification problem, classifying ethnics into two-class: i.e. Asian and Non-Asian.

In the preprocessing, we firstly cropped the face area of each model according to the indicator matrix showing whether there is a valid point in that position. While median filter was adopted to remove spikes and cubic interpolation was employed to fill holes. The face samples for experiments possess expression, illumination, and slight pose variations. Several examples of facial range and texture images are

demonstrated in Fig. 3.6. All the images are resized to 96 × 120 pixels.

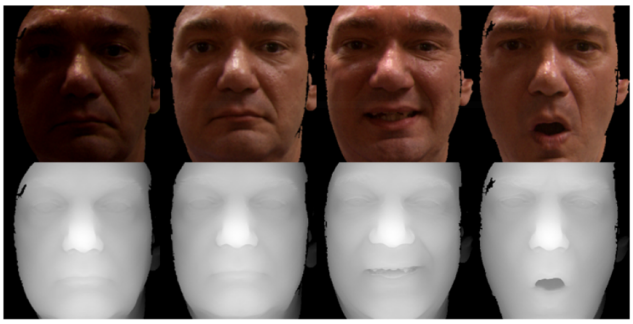


Figure 3.6: An example of range and intensity images in FRGC v2.0

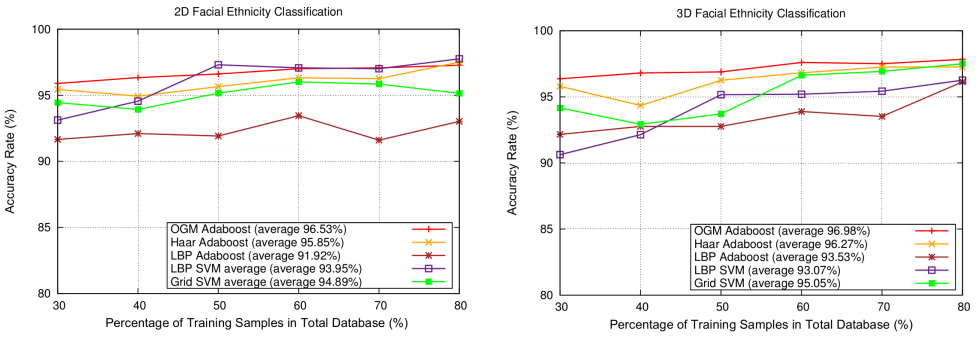
We designed four experiments as follows: the first one is to test performance of the proposed method for texture based facial ethnicity classification; while the second is to evaluate its accuracy for shape based ethnicity classification. The third one combines both 2D and 3D modalities at matching score level for result improvement. In the last one, we investigate the importance of certain face regions, such as eyes and nose, in ethnicity classification.

In each experiment, we set the size of the gallery samples using the percentage of the whole dataset, varying from 30% to 80% with a step of 10%, and the remaining samples were treated as probes. The 3D facial scans from the same subject are grouped into the same set to ensure that the results are not biased by the similarity between the testing and the training data in terms of the identity. For each experiment, we repeated 10 times and calculated the average performance.

A. Texture based Ethnicity Classification (2D Modality)

In this experiment, we evaluated performance of the proposed ethnicity classification approach on facial texture images. In order to highlight its effectiveness, we implemented several techniques in the literature for comparison including Grid+SVM [Lu *et al.* 2006], Haar+Adaboost [Yang & Ai 2007], LBP+Adaboost [Zhang & Wang 2009], LBP+SVM [Lyle *et al.* 2010].

From Fig. 3.7(a), we can see that even though in a few settings (50% and 80% samples are exploited in the training phase), the results of the proposed approach are slightly inferior to those of the state of the art ones (but still comparable),



(a) (b)

Figure 3.7: Performances of different methods for ethnicity classification on the FRGC v2.0 dataset. (a) 2D modality and (b) 3D modality.

and the performance generally remains stable as the size of the gallery set changes. The average classification rate is 96.5% which outperforms that of the other tasks, demonstrating the effectiveness of OGMs to describe texture information.

B. Shape based Ethnicity Classification (3D Modality)

At this step, in order to be consistent with the experiment using facial texture images, we adopted the same experimental configurations.

Fig.3.7(b) clearly illustrates that in facial range image based ethnicity classification, our approach achieves better results than any of the others do. Even if the percentage of training samples is only 30%, its classification accuracy is still above 96%. Those experimental accuracies demonstrate that OGMs outperform other well-known local features in discriminative power for shape information representation.

C. Multi-modal Score Fusion

Table 3.3 displays the classification results when combing both the 2D and 3D modalities. We can see that the joint use of the texture and geometry clues leads to better performance than either of the single modality. From Fig. 3.8, we can see the similar phenomenon as in the last experiment only using shape information. The proposed approach surpasses the other ones and keeps robust when the gallery size varies.

D. Importance Analysis of Facial Regions

Recall that different ethnic groups usually have local discriminations in facial

Table 3.3: Performances of the proposed approaches based on 2D, 3D and both modalities for ethnicity classification on FRGC V 2.0.

Percentage	Texture	Shape	Fusion
30%	95.56%	96.37%	97.22%
40%	95.89%	96.80%	97.62%
50%	96.34%	96.88%	97.63%
60%	97.00%	97.60%	98.06%
70%	97.07%	97.50%	98.17%
80%	97.27%	97.84%	98.26%

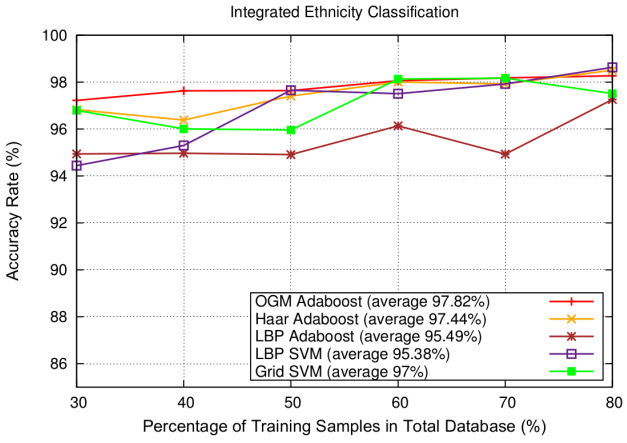
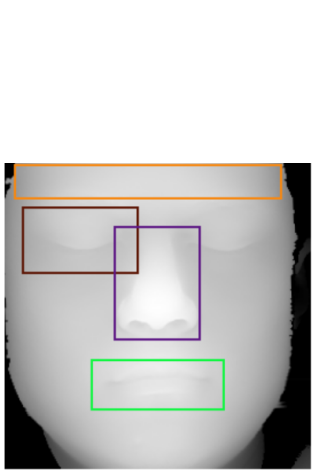
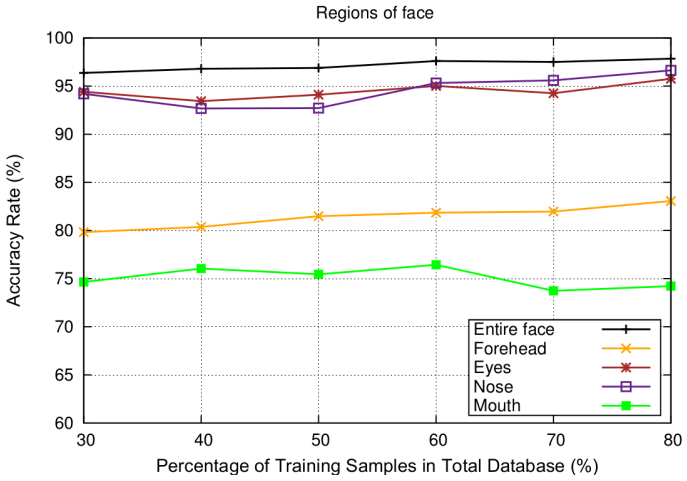


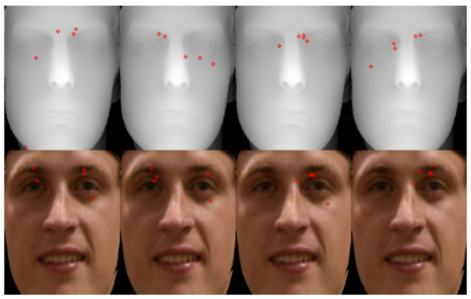
Figure 3.8: Integrated Performances of different approaches for facial ethnicity classification on FRGC v2.0



(a)



(b)



(c)

Figure 3.9: (a) Four regions of face intercepted to estimate their contributions for ethnicity classification. (b) Performances of different face regions using OGM+Adaboost on 3D facial ethnicity classification. (c) The first five selected features for each OGM of both modalities.

regions. To investigate the importance of face areas, in this experiment, since all images are resized to a pre-defined scale, according to the positions in the facial images and utilizing some fixed rectangle boxes, we roughly divide facial images into four areas, each of which possesses a distinctive characteristic: eyes, nose, forehead and mouth, as shown in Fig. 3.9(a). Then, we exploit the proposed approach to estimate the contribution of each region.

Fig.3.9(b) demonstrates that among the four selected areas, the eyes and nose are more discriminative than mouth and forehead in ethnicity classification. When combining the four facial regions, performance is improved, suggesting that each area has its own impact.

On the other hand, we also analyze the importance of each facial organ in the viewpoint of machine learning and observe the set of OGM features selected by Adaboost. As depicted in Fig. 3.9(c), the first five chosen features of each OGM of both modalities are mainly located in the eye region and the nose region, highlighting the importance of the two areas. Additionally, we can see that the conclusion in this analysis accords with the previous one.

3.5.2.2 OGMs based approach on BU-3DFE

Additionally, in order to examine the applicability of the proposed approach, we also validated our method on the BU-3DFE database which contains 100 subjects: 51 Whites, 24 East-Asians and others (more details are shown in Table 3.2). Because of the imbalance of ethnic groups, we employed 24 samples randomly selected from Whites and all samples of East-Asians as a binary classification. The settings of probe and gallery set are the same as those of FRGC v2.0.

Table 3.4 demonstrates the results of the proposed method for the binary classification (Whites vs. Asians) on the BU-3DFE database. Since samples employed from BU-3DFE in the experiment is less than those of FRGC v2.0, the accuracy is slightly lower in the same experimental setting. However, we can still see that the proposed method achieves more than 90% accuracies in both the single modalities, i.e. texture and geometry as the number of training samples increases. When we finally combine these similarity measurements of the two modalities, the result is further improved, indicating that both clues contribute to the classification result.

Table 3.4: Performances of the proposed approaches based on 2D, 3D and both modalities for ethnicity classification on BU3D-FE

Percentage	Texture	Shape	Fusion
30%	92.92%	92.92%	94.84%
40%	95.24%	94.44%	96.63%
50%	95.00%	96.38%	97.21%
60%	94.78%	97.48%	97.56%
70%	95.35%	97.50%	97.70%
80%	96.72%	97.32%	97.88%

3.5.3 Experiments: LCP based approach

Similar to experiments of OGMs based approach, we also evaluate LCP based approach respectively on FRGC and BU3D-FE databases.

3.5.3.1 LCP based approach on FRGC

On the FRGC database, asian and white subjects are chosen for both gender and ethnicity classification, and in totality there are 3676 face samples belonging to 319 white and 99 Asian people. Even though all face samples in the FRGC v2.0 dataset are nearly frontal, Iteratively Closet Point (ICP) [Zhang 1994] is adopted to align the face model to the reference that is pre-defined, in order to control the error caused by slight pose changes. We then extract a pair of registered facial range and texture image from the aligned face model and all of the facial texture and range images are normalized so that the outer corners of two eyes have a fixed distance of 100 pixels. An average mask is further used to eliminate non-face regions and segment face out, and finally images are normalized to the size of 140 $\times$ 140 pixels. Examples of normalized face images are shown in Fig. 3.5.

We design four experiments: the first is to test the performance of the Euclidean distance (L2) and the city block (L1) distance in clustering based quantization; the second is to analyze the robustness of the LCP descriptor to the LBP based one; the third is to evaluate the effectiveness of the proposed approach to gender and ethnicity classification in the modality of 2D, 3D and their combination; and the last one is to make the comparison with the state of the art techniques.

A. Performance of L1 and L2 Distance in clustering based quantization

In order to evaluate the performance of L1 and L2 distance in clustering based quantization adopted in LCP, we randomly select 20 textured 3D face models from the FRGC database, and extract the features of local circular patterns LCP 2,8 from both the texture and shape modality as plotted in Fig. 3.10(a) and 3.10(d). The clustering results using L2 distance and L1 distance with the same number of clusters (59, identical to the number of bins in the LBP descriptor with 8 neighboring pixels) are shown in Fig. 3.10. In Fig. 3.10, the x-axis denotes the index of each

element in an LCP code starting from the left-top position as shown in Fig. 3.4 (a), and the y-axis displays the exact value of each element. In this case, the number of pixels around a central pixel is set at 8, thereby leading to 8 elements in an LCP code, and their values vary in the range of $[-255, 255]$.

We observe the difference between distributions of texture and shape data, and find out that the texture data are more concentrated than the shape data, therefore using the same quantization parameter like the traditional LBP to deal with these different distributions is obviously unsatisfied. In contrast, both the clustering results of L2 and L1 reflect the distribution difference much better. Meanwhile, the results obtained with L2 distance and L1 distance are different: the results of L2 distance are more evenly spaced than the ones of L1 distance, while L1 distance results focus on the concentrations of the data and pay less attention to the outlier of the distribution, likely leading to better performance in the step of gender and ethnicity classification (we show these accuracies subsequently).

Furthermore, we compare the computational cost for each distance metric, i.e. L1 and L2. Experiments are carried out on a PC with Intel Core i3 CPU using Matlab implementation. With 314,960 training local circular patterns, the computational time taken by K-means clustering with L2 distance repeated 10 times is $12,233$ s, while that for L1 distance is only 600 s. L1 distance is thus computationally more efficient than L2 distance as well.

B. Analysis on robustness to noise of LCP

The performance under noise influence of clustering based quantization in LCP vs. binary quantization in LBP is also evaluated. Gaussian random noises with deviation of 2 are added on both facial texture and range images. Then local circular patterns $LCP_{2,8}$ are extracted before and after adding noise, and quantized using K-means clustering with L1 and L2 distances respectively. The number of clusters is set to 59 so as to achieve a fair comparison with the $LBP^{u2}_{2,8}$ operator of the same dimensionality. Histograms of quantization labels are constructed for the entire image. Fig. 3.11 shows an example, from which we can see that the influence of noise is more serious to the LBP based histograms than the LCP based one.

In order to quantitatively measure the difference between histograms, the dis-

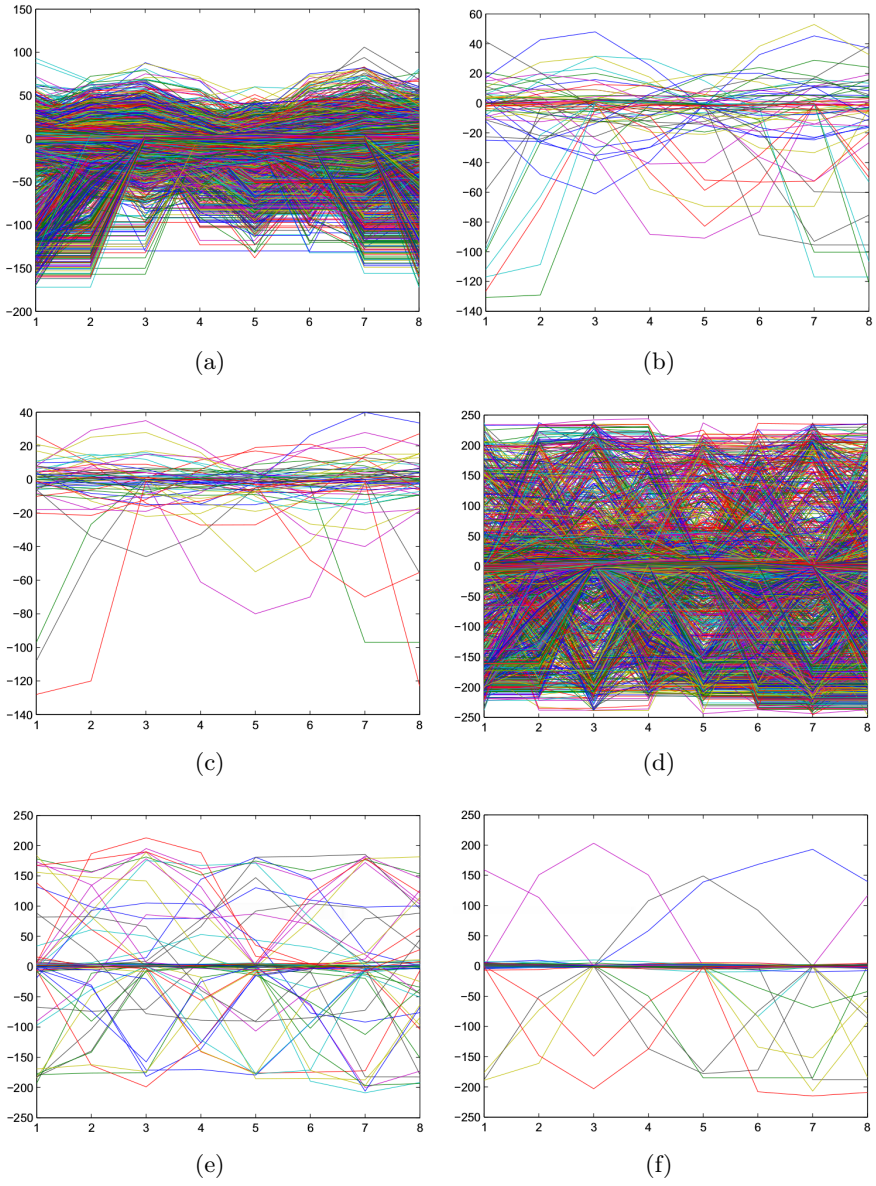


Figure 3.10: The clustering results on the FRGC v2.0 database in the texture and shape modality respectively: (a) training texture data for clustering, (b) clustering result of texture data using L2 distance, (c) clustering result of texture data using L1 distance, (d) training shape data for clustering, (e) clustering result of shape data using L2 distance, and (f) clustering result of shape data using L1 distance.

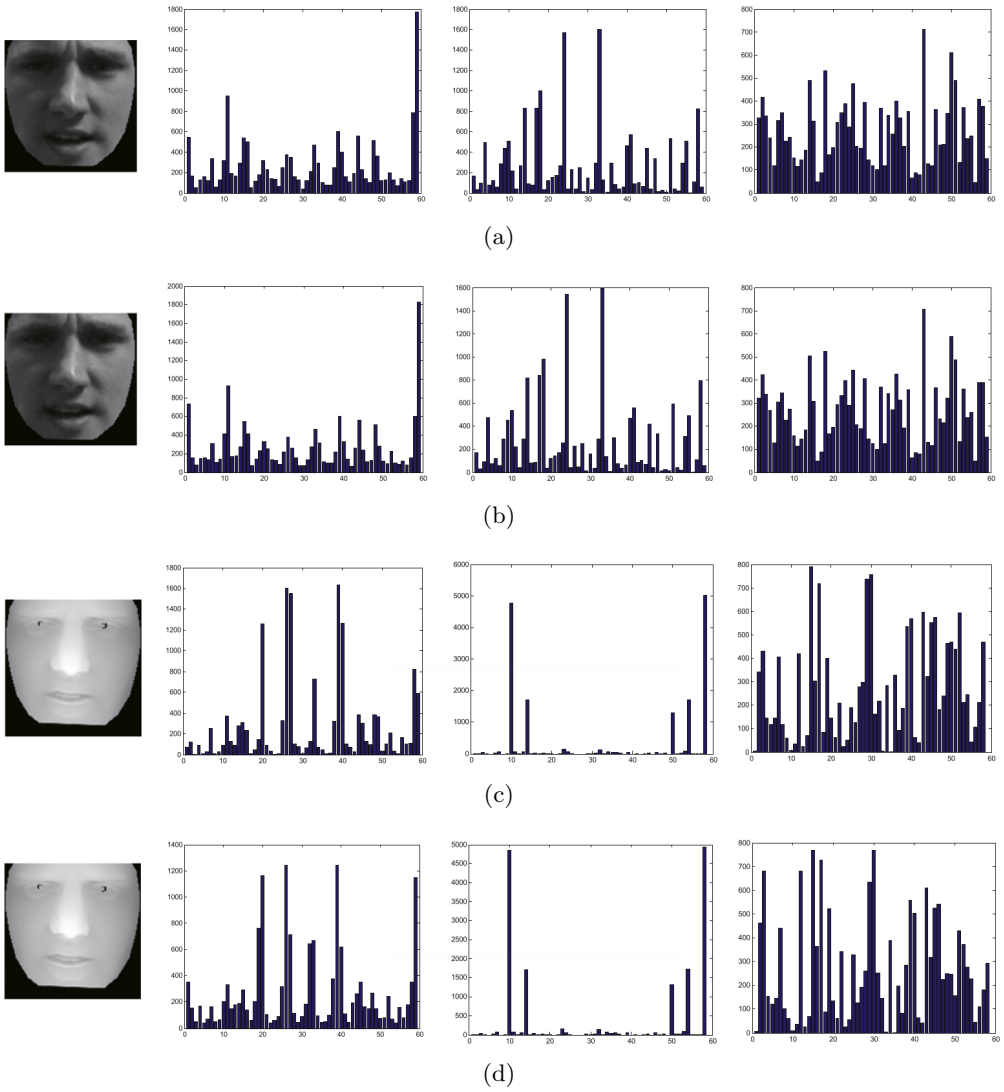
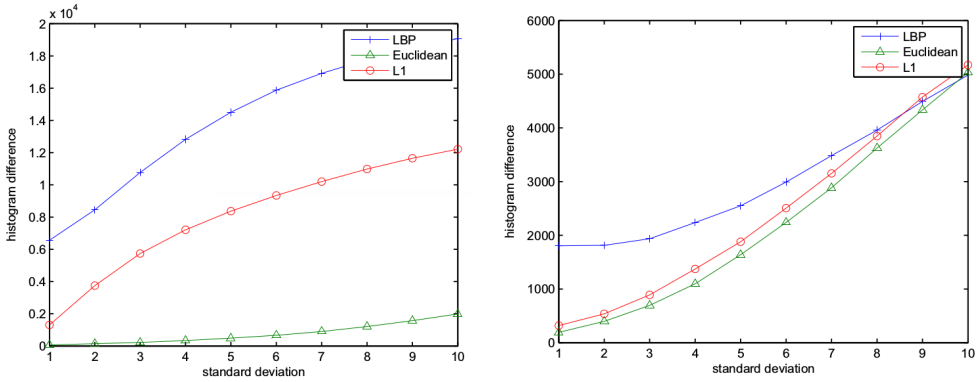


Figure 3.11: Histograms extracted before ((a) and (c)) and after ((b) and (d)) noise adding using the samples from FRGC v2.0. The second column shows LBP histograms, in the third column are histograms extracted with L2 distance, histograms extracted with L1 distance are plotted in the fourth column.



(a) (b)

Figure 3.12: The difference between histograms extracted before and after noise adding using the samples from FRGC v2.0: (a) for the shape modality and (b) for the texture modality.

tance $Diff(B, A)$ of histograms extracted before (B) and and after (A) noise is added is calculated as below.

$$Diff(B, A) = \sum_i |B_i - A_i| \tag{3.12}$$

where B and A are the two histograms to be compared, and B_i and A_i are the ith bin value of B and A. 20 samples are randomly selected from the FRGC v2.0 database as in the previous experiment, and Gaussian random noises with deviations vary from 1 to 10 and are added on range and texture images. Fig. 3.12 shows the average difference of histograms for LBP, LCP-L2, and LCP-L1. From the comparison, we can conclude that in both modalities, the clustering based quantization (LCP-L2, LCP-L1) outperforms binary quantization (LBP) under noise influence. Clustering with L2 distance performs better than L1 distance.

C. Gender and ethnicity classification with LCP

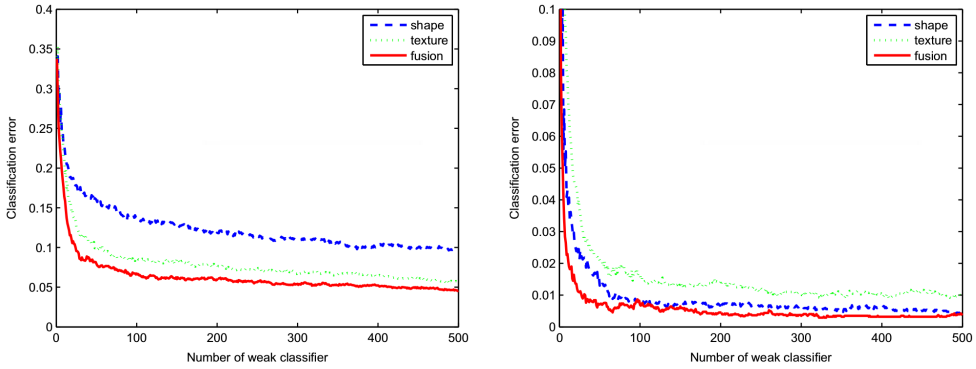
This experiment tests the performance of K-means clustering quantization based local circular patterns (LCP) in both tasks of gender and ethnicity classification. For multi-modal facial representation, $LCP_{8,1}$, $LCP_{8,2}$ and $LCP_{8,3}$ are used to extract features from the facial texture and range images, and quantized using K-means clustering. L1 distance is utilized in clustering due to its efficiency. All 2D

texture and 3D range images are divided into 6×6 rectangular facial regions as shown in Fig. 3.5. As a result, for each modality of a face, three LCP histograms are extracted, and they are concatenated again to construct the final description of this modality. Through extracting LCP based features using multi-resolution filters and calculating histograms hierarchically, we can extensively find those distinctive features to represent gender and ethnicity related texture and shape variations.

As we defined in the previous Section, N is the number of local circular patterns used to compute cluster centers. Actually, in our experiments, N corresponds to the number of 2D or 3D facial images randomly selected, since we have to ensure that these patterns are in an even distribution for face analysis. Therefore once a facial image is chosen, these local circular patterns of all pixels are used in the K-means clustering. We further vary the number of facial images for training and observe its impact on classification performance on the FRGC v2.0 dataset. In gender classification, when this number reaches about 50 and 60 for 2D and 3D modality respectively, their accuracies remain stable; while in ethnicity classification, this number is around 30 and 40. Meanwhile, we cannot set this number too large in order to avoid overfitting. As a result, in the following experiments, we set it at 60 so that it fits the two modalities in both tasks.

A 10-fold cross validation is adopted to evaluate the performance of the proposed approach, in which the database is randomly partitioned into 10 folds. Experiments are carried out 10 times, and each time 9 folds are exploited as the training set, and the remaining 1 fold as the testing set. Thus, each fold is tested once. We ensure that each subject is only assigned to one fold, so that the classification is person independent.

Fig. 3.13 shows the results of gender and ethnicity classification. As we can see, for gender classification texture features outperform shape features, while for ethnicity classification shape features perform better than texture features. In both tasks of gender and ethnicity classification, performance in either of the single modality is enhanced by combining shape and texture modalities. The classification errors achieved by fusion of these two modalities in the experiments of gender and ethnicity classification are 4.55% and 0.37% respectively. The confusion matrixes for



(a) (b)

Figure 3.13: Classification results of LCP achieved on the FRGC v2.0 database: (a) gender classification and (b) ethnicity classification

gender and ethnicity classification are displayed in Tables 3.5 and 3.6.

Table 3.5: Confusion matrix of gender classification using LCP-L1 on the FRGC v2.0 dataset, and each item is depicted in the form of (average, standard deviation).

Modality		Male	Female
Shape	Male	(0.9097, 0.0478)	(0.0903, 0.0478)
	Female	(0.1017, 0.0326)	(0.8983, 0.0326)
Texture	Male	(0.9476, 0.0345)	(0.0524, 0.0345)
	Female	(0.0579, 0.0512)	(0.9421, 0.0512)
Fusion	Male	(0.9596, 0.0324)	(0.0404, 0.0324)
	Female	(0.0509, 0.0473)	(0.9491, 0.0473)

Table 3.6: Confusion matrix of ethnicity classification using LCP-L1 on the FRGC v2.0 dataset, and each item is depicted in the form of (average, standard deviation).

Modality		White	Asian
Shape	White	(0.9987, 0.0024)	(0.0013, 0.0024)
	Asian	(0.0133, 0.0260)	(0.9867, 0.0260)
Texture	White	(0.9941, 0.0081)	(0.0059, 0.0081)
	Asian	(0.0198, 0.0254)	(0.9802, 0.0254)
Fusion	White	(0.9990, 0.0041)	(0.0010, 0.0041)
	Asian	(0.0087, 0.0220)	(0.9913, 0.0220)

We then analyze the results of L1 and L2 distance based K-means clustering quantization by comparing their accuracies in both classification tasks, i.e. gender and ethnicity. Fig. 3.14 shows the curves of classification errors vs. the number of

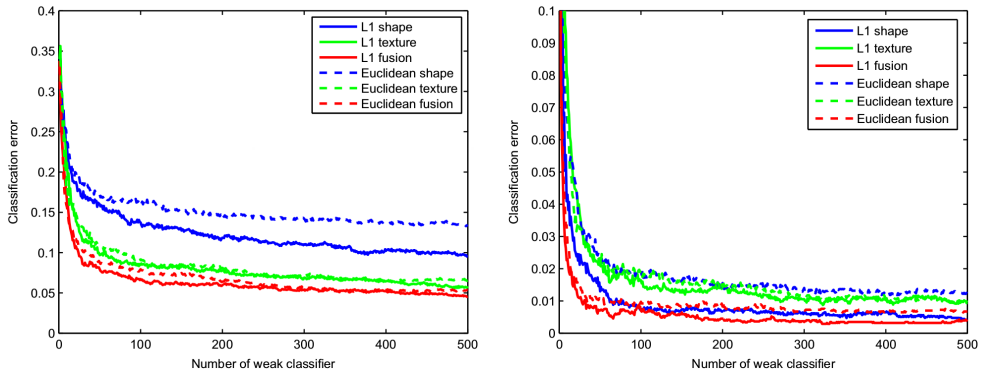


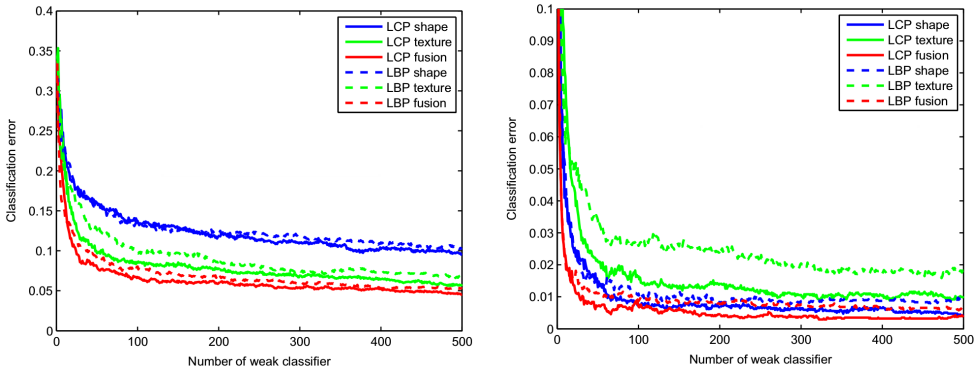
Figure 3.14: Classification results of LCP achieved on the FRGC v2.0 database: (a) gender classification and (b) ethnicity classification

weak classifier, and Table 3.7 shows the comparison of classification errors between L1 and L2 distances. From these results we can see that L1 distance generally performs better than L2 distance, especially for the shape modality, while for texture modality the two distance metrics perform similarly.

Table 3.7: Performance comparison between L1 distance and L2 distance of the LCP descriptor on the FRGC v2.0 database, and each item is depicted in the form of (average, standard deviation).

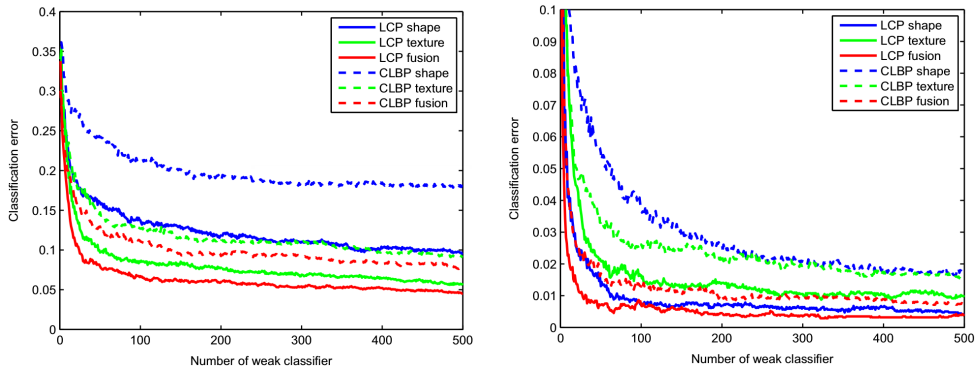
Task		Shape	Texture	Fusion
Ethnicity	L1	(0.0042, 0.0058)	(0.0096, 0.0081)	(0.0037, 0.0051)
	L2	(0.0122, 0.0099)	(0.0094, 0.0092)	(0.0067, 0.0083)
Gender	L1	(0.0949, 0.0330)	(0.0557, 0.0262)	(0.0455, 0.0272)
	L2	(0.1344, 0.0380)	(0.0654, 0.0327)	(0.0536, 0.0279)

D. Comparison with state of art In this experiment, we compare the performance of LCP with related local descriptors, i.e. LBP and one of its best variants, namely Complete LBP (CLBP) [Guo *et al.* 2010a]. Using the same parameters in neighborhood setting, i.e. combining the histograms extracted in the neighborhood of $(8, 1)$, $(8, 2)$, and $(8, 3)$, the uniform LBP results in 59 different values while CLBP provides 200 different values for each histogram. Fig. 3.15and Fig.3.16 compare the classification error with respect to the number of weak classifier curves among these three methods, and Table 3.8 compares the classification error achieved by these methods. We can see from the results that the proposed method (LCP) consis-



(a) (b)

Figure 3.15: Comparison of classification results between LCP and LBP on the FRGC v2.0 database: in (a) gender classification and (b) ethnicity classification.



(a) (b)

Figure 3.16: Comparison of classification results between LCP and CLBP on the FRGC v2.0 database: in (a) gender classification and (b) ethnicity classification

tently outperforms the LBP and Complete LBP methods, which demonstrate the superiority of clustering-based quantization of local circular patterns over binary quantization scheme used by LBP and CLBP.

Meanwhile, we compare the performance of the proposed approach with the ones of the state of the art techniques, which also concentrate on classifying gender and ethnicity using both the modality of 2D tex- ture and 3D shape of human faces. Table 3.9 summarizes the comparison to highly related tasks. Although we carry out experiments with significantly more scans and more subjects, the accuracies of both gender and ethnicity classification are higher than

Table 3.8: Performance comparisons among LCP, LBP, and Complete LBP on the FRGC v2.0 dataset, and each item is depicted in the form of (average, standard deviation).

Task		Shape	Texture	Fusion
Ethnicity	LCP	(0.0042, 0.0058)	(0.0096, 0.0081)	(0.0037, 0.0051)
	LBP	(0.0088, 0.0081)	(0.0182, 0.0105)	(0.0063, 0.0082)
	CLBP	(0.0177, 0.0114)	(0.0153, 0.0078)	(0.0074, 0.0086)
Gender	LCP	(0.0949, 0.0330)	(0.0557, 0.0262)	(0.0455, 0.0272)
	LBP	(0.1034, 0.0346)	(0.0686, 0.0342)	(0.0524, 0.0308)
	CLBP	(0.1807, 0.0481)	(0.0909, 0.0367)	(0.0791, 0.0403)

the ones in [Lu *et al.* 2006, Wu *et al.* 2010]. The performance of gender classification is slightly lower than that of the work [Huynh *et al.* 2012], while it should be noted that Huynh et al. [Huynh *et al.* 2012] make use of uniform LBP features and Gradient-LBP features (a special case of CLBP) extracted from facial texture and range images respectively, and these features prove inferior to the proposed LCP features in our experiments. Furthermore, their result is based on 1149 pairs of facial range and gray images of 105 subjects, and the experiment is performed only once with half samples for training and half for testing. In our work, we carry out the experiment using 10-fold cross validation, where 3676 textured 3D face models of 418 subjects are involved.

Table 3.9: Performance comparison with those of the state of the art techniques of multi-modal facial gender and ethnicity classification on the FRGC v2.0 database (* indicates an exception that only makes use of the 3D modality, and the figures in bold are the best ones in individual tasks).

Approach	Sub. num.	Protocol	Gender	Ethnicity
[Lu *et al.* 2006]	376 Sub. 1240 scans	10-fold C.-V.	91.00% ± 0.03	98.00% ± 0.16
[Wu *et al.* 2010]	260 (200 vs. 60) scans	6 Times	93.60% ± 0.04	-
[Huynh *et al.* 2012]	105 Sub. 1149 scans	1 Time	96.70%	-
[Toderici *et al.* 2010]*	418 Sub. 3676 scans	10-fold C.-V.	≈ 93.50%	≈ 99.50%
Our method	418 Sub. 3676 scans	10-fold C.-V.	**95.50% ± 0.03**	**99.60% ± 0.01**

Furthermore, we also list the work [Toderici *et al.* 2010] that only makes use of the 3D modality in Table 3.9 since it exploits the same experimental protocol as we do. If regarding the shape information, LCP achieves comparable results as [Toderici *et al.* 2010] does in ethnicity classification and it does not perform as

good as [Toderici *et al.* 2010] in gender classification, but when our system combines the clues of texture and shape, the accuracies in both the tasks are improved, which surpass the ones in [Toderici *et al.* 2010]. Such a fact highlights the advantage of multimodal facial gender and ethnicity classification over the single modality based one. Additionally, [Toderici *et al.* 2010] employs the 3D face recognition system (URxD) to measure the similarity of faces, holding a pipeline of deformed model based 3D surface registration and Haar wavelet decomposition as well as steerable pyramid transform based feature extraction, which is computationally expensive. In contrast, our system tends to be more efficient.

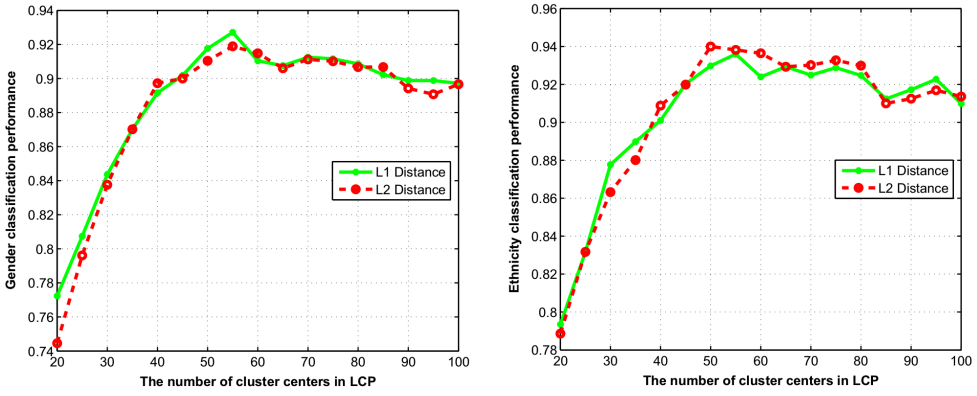
3.5.3.2 LCP based approach on BU-3DFE

As introduced previously, we also verify the ability of generalization of the proposed LCP based method on BU3D-FE database.

In the preprocessing, median filter is firstly utilized to remove spikes and cubic interpolation is adopted to fill holes. We employ ICP to align the face model to the pre-selected reference to correct possible pose variations. The registered facial range and texture image are then generated from each aligned face model and all of the facial texture and range images are normalized to the size of 140 $\times$ 140 pixels as in FRGC v2.0.

For gender classification, all 2500 3D face models belonging to these 100 persons are used in our experiments. While for ethnicity classification, due to the imbalance distribution of different ethnic groups (51 Whites, 24 East-Asians, 9 Blacks, 8 Hispanic–Latinos, 6 Indians, and 2 Middle-East Asians), we exploit 1875 face models of Whites and East-Asians, as a binary classification problem. The settings of probe and gallery set in both tasks are the same as those of FRGC v2.0 stated previously. For each experiment, we make use of 10-fold cross validation and calculate the average performance.

A. Discussing the number of cluster centers in LCP As we mentioned in the previous Section, one of the key issues associated with K-means clustering is the number of clusters. This number controls the balance between descriptive power and sensitivity to noise. We experimentally evaluate this factor in the task of gender



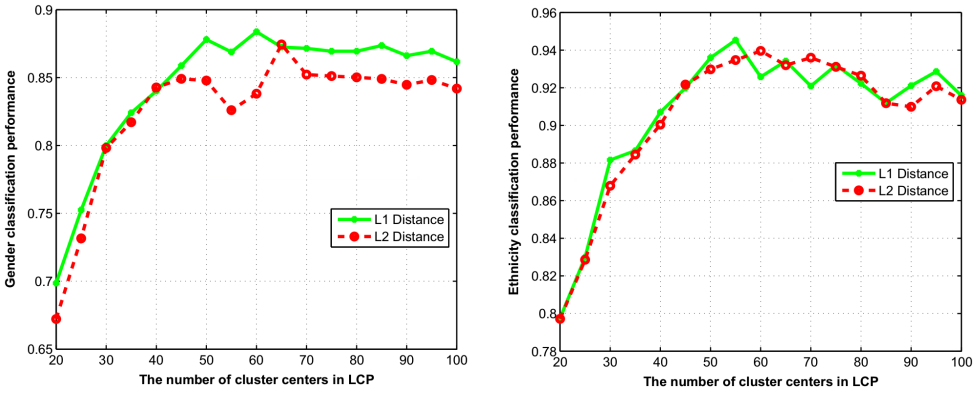
(a) (b)

Figure 3.17: Performance based on texture with regard to the number of cluster centers in $LCP_{1,8}$ in (a) gender classification and (b) ethnicity classification on the BU-3DFE dataset.

and ethnicity classification subsequently.

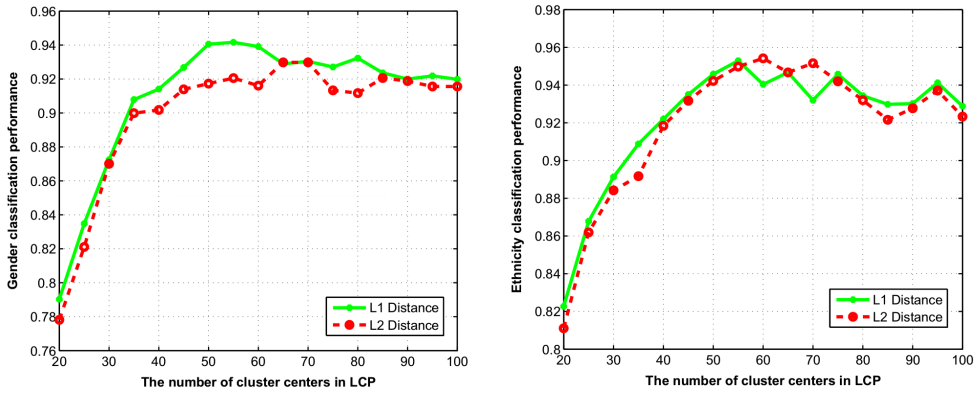
Taking the $LCP_{1,8}$ operator as an example, we increase the number of cluster centers from 20 to 100 at an interval of 5, and discover that the best performance of texture, shape or their combination is achieved in the range of [50, 70] for L1 and L2 distances in both tasks (as depicted in Fig. 3.17, Fig. 3.18, and Fig. 3.19), which coincidentally accords with the number (59) previously assigned for fair comparison with LBP.

B. Comparison with the approaches in the Literature In this experiment, except LBP and its two variants, namely LTP [Tan & Triggs 2010] and CLBP [Guo *et al.* 2010a], we also compare LCP with the features in [Lu *et al.* 2006, Varma & Zisserman 2009, Liu & Fieguth 2012]. [Lu *et al.* 2006] applies the holistic feature which is the raw pixels of a number of patch cropped from the facial image (denoted as "Grid" in Table 3.10). [Varma & Zisserman 2009] and [Liu & Fieguth 2012] both focus on texture classification, and we discuss them since they both make use of K-means clustering to learn the vocabulary of local pixel patterns. However, in LCP, we define the pattern as the gray value difference between the central pixel and its neighboring ones within the patch, rather than the original gray values [Varma & Zisserman 2009] or their Random Projection (RP) [Liu & Fieguth 2012] (denoted as "Pixel" and "RP" respectively in the



(a) (b)

Figure 3.18: Performance based on shape with regard to the number of cluster centers in $LCP_{1,8}$ in (a) gender classification and (b) ethnicity classification on the BU-3DFE dataset.



(a) (b)

Figure 3.19: Performance based on multi-modal combination with regard to the number of cluster centers in $LCP_{1,8}$ in (a) gender classification and (b) ethnicity classification on the BU-3DFE dataset.

Table 3.10: Comparison of approaches of the texture and shape modality as well as their combination using Adaboost in the task of gender and ethnicity classification on the BU-3DFE dataset. (The figures in bold are the best ones in individual tasks).

Task		Texture	Shape	Fusion
Gender	LBP	$93.53\% \pm 0.03$	$87.35\% \pm 0.06$	$94.58\% \pm 0.03$
	LTP	$93.20\% \pm 0.01$	$87.44\% \pm 0.05$	$94.18\% \pm 0.03$
	LCP-L1	$94.18\% \pm 0.03$	$\mathbf{90.76\% \pm 0.04}$	$\mathbf{95.60\% \pm 0.03}$
	LCP-L2	$94.00\% \pm 0.04$	$89.09\% \pm 0.04$	$95.56\% \pm 0.03$
	CLBP	$93.82\% \pm 0.02$	$88.12\% \pm 0.04$	$94.91\% \pm 0.03$
	Grid	$81.27\% \pm 0.07$	$78.70\% \pm 0.07$	$85.60\% \pm 0.08$
	Pixel-L1	$94.36\% \pm 0.05$	$81.64\% \pm 0.04$	$95.45\% \pm 0.02$
	Pixel-L2	$91.64\% \pm 0.04$	$80.97\% \pm 0.05$	$93.82\% \pm 0.05$
	RP-L1	$\mathbf{94.91\% \pm 0.04}$	$81.64\% \pm 0.03$	$95.27\% \pm 0.01$
	RP-L2	$94.18\% \pm 0.06$	$81.94\% \pm 0.01$	$94.42\% \pm 0.02$
Ethnicity	LBP	$95.07\% \pm 0.04$	$95.56\% \pm 0.04$	$96.89\% \pm 0.03$
	LTP	$94.91\% \pm 0.03$	$95.05\% \pm 0.04$	$96.62\% \pm 0.04$
	LCP-L1	$\mathbf{97.31\% \pm 0.04}$	$96.33\% \pm 0.03$	$\mathbf{97.42\% \pm 0.04}$
	LCP-L2	$97.13\% \pm 0.04$	$95.98\% \pm 0.03$	$97.22\% \pm 0.03$
	CLBP	$95.86\% \pm 0.04$	$\mathbf{96.62\% \pm 0.02}$	$97.13\% \pm 0.04$
	Grid	$93.39\% \pm 0.10$	$87.87\% \pm 0.10$	$94.07\% \pm 0.05$
	Pixel-L1	$91.26\% \pm 0.05$	$91.70\% \pm 0.05$	$92.89\% \pm 0.01$
	Pixel-L2	$90.52\% \pm 0.06$	$91.26\% \pm 0.04$	$92.15\% \pm 0.01$
	RP-L1	$90.07\% \pm 0.06$	$92.30\% \pm 0.06$	$93.19\% \pm 0.02$
	RP-L2	$91.56\% \pm 0.06$	$91.41\% \pm 0.04$	$94.37\% \pm 0.02$

following).

For LBP, LTP, LCP, and CLBP, as in FRGC v2.0, we combine the results of different neighborhood settings, i.e. $(1, 8)$, $(2, 8)$, and $(3, 8)$, and divide the facial texture and range images into 6×6 regions. For "Grid", we directly inherit the parameters set in [Lu *et al.* 2006] that the number of patches is 80 (8×10) and the size of each patch is 8×8 pixels. For "Pixel" and "RP", to make a fair comparison with the LBP family, we set the patch size at 7×7, approximately equivalent to the combination of three neighborhood sizes in LBP, LTP, LCP, and CLBP, and employ their division scheme, i.e. 6×6 blocks for each face image. Additionally, in RP, we project the patch in an 8-dimensional PCA subspace for clustering.

From Table 3.10, we can see that:

- The performance of LCP-L1 and LCP-L2 in both gender and ethnicity classification is better than that of its counterparts in LBP family, i.e. uniform LBP, LTP, CLBP, on either of the single texture or shape modality as well

as their combination, except the case in shape based ethnicity classification where the accuracy of LCP is only 0.29% below that of CLBP (still comparable). This fact clearly indicates that LCP is an effective improvement to the LBP methodology. Meanwhile, in LCP, LCP-L1 always performs LCP-L2, showing that the L1 distance is a better choice to learn LCP code than the L2 distance does.

- Regarding on Pixel-L1 and Pixel-L2 or RP-L1 and RP-L2 which apply the K-means clustering technique (using different distances) to learn local descriptors from original gray level values within a patch [Varma & Zisserman 2009] or from their random projection [Liu & Fieguth 2012], the LCP descriptor is competitive as well. Even though the results of LCP are slightly inferior to those of Pixel and RP (using L1 distance) based on texture clues in gender classification, the results of LCP in other tasks (including shape-, fusion-based gender classification as well as texture-, shape-, and fusion-based ethnicity classification), are significantly superior to the ones of Pixel and RP, especially in the 3D modality. The reason mainly lies in that the facial range image is generally smooth and thus lacks discrimination, while the differences between neighboring pixels better highlight its details that are critical in classification than the original pixels [Varma & Zisserman 2009] or their random projection [Liu & Fieguth 2012]. It demonstrates the effectiveness of LCP for such issues.

- For all these methods discussed in the table, their classification accuracies based on the fusion of texture and shape cues are better than the corresponding ones using either of the single modality, illustrating that combining information conveyed in two modalities improves the performance in facial gender and ethnicity classification.

3.5.3.3 LCP: Time complexity evaluation

The K-means clustering based quantization in LCP is time consuming. However, it should be noted that this stage is carried out offline, and these cluster centers need

to be generated only once during the training process. In online feature extraction, the only difference between LCP and LBP lies in that LCP calculates certain distance between a given local circular pattern and the pre-computed cluster centers and chooses the minimum one for quantization; while LBP makes use of binary quantization. The time cost additional to LBP is thus the distance calculation with all cluster centers. The more the cluster centers are, the higher the time cost is. In our experiments, there are 59 cluster centers, and based on C++ implementation, the average time cost of LBP, LCP-L1, and LCP-L2 is 3.33 ms, 9.76 ms, and 9.72 ms, respectively. Such computational complexity is generally under control in efficient face analysis applications.

3.6 Conclusion and future work

In this chapter, we present two effective and efficient approaches on face based gender and ethnicity classification by combining both boosted local texture and shape features extracted from 3D face models. The proposed methods are in contrast to the existing ones that only make use of either modality of 2D texture or 3D shape of faces.

To comprehensively represent the difference between different genders or ethnicities, two novel local descriptor, namely oriented gradient maps (OGMs) and local circular patterns (LCP) are introduced. OGMs make use of gradient information to highlight local geometry and texture variations of human faces, while LCP improves the widely investigated local binary patterns (LBP) as well as its variants by replacing the binary quantization with a clustering based one, thereby resulting in higher discriminative power and better robustness to noise.

Moreover, the Adaboost based feature selection process finds the most discriminative gender- and ethnic-related features and assigns them with different weights to highlight their importance in classification, which not only further raises the performance but reduces the time and memory cost as well.

Experiments to evaluate the OGMs based approaches are carried out on the FRGC v2.0 database, and the performance is up to 98.3% to distinguish Asian

from non-Asian people when 80% face samples are exploited in the training stage, demonstrating the effectiveness of the proposed method.

The experimental results of LCP based approach for gender and ethnicity classification achieved are up to 95.50% and 99.60% respectively on the FRGC v2.0 dataset, and 95.60% and 97.42% respectively on the BU-3DFE dataset, which clearly demonstrate the advantages of the proposed method.

Although our method have achieved remarkable performances for gender and ethnicity classification, there are still some improvements in our further work. Firstly, more and more large-size databases have been proposed recently, such as FLW, we need evaluate the effectiveness of our approaches on the big databases and make some reasonable improvements. Secondly, in our work, we carry out ethnicity classification as binary classification due to lacking of samples of other ethnic groups. After we have available databases, we should extend our framework to solve a multi-class problem.

Chapter 4

Facial expression recognition

4.1 Introduction

As introduced previously, in the past decades, a large number of FER approaches have been proposed. They can be categorized from three perspectives, namely the data modality, expression granularity, and temporal dynamics. From the first perspective, they are classified into 1) 2D FER (which uses 2D gray or color face images), 2) 3D FER (which uses 3D range images, point clouds, or meshes of faces), and 3) multimodal 2D + 3D FER (which uses both 2D and 3D facial data). From the second perspective, they are divided into 1) six basic facial expression (i.e., anger, disgust, fear, happiness, sadness, and surprise) recognition, 2) facial Action Unit (AU, e.g., brow raiser, lip tightener, and mouth stretch) detection and recognition. From the third perspective, they are categorized into static (still images) and dynamic (image sequences) FER. In this work, we focus on the problem of recognizing the six basic facial expressions using multimodal 2D + 3D static images.

Appearance-based 2D FER has been widely investigated since 1990s [Bettadapura 2012]. The main research topics lie in three aspects: face detection, expression related feature extraction and classification. Comprehensive surveys of 2D FER approaches are given in [Kumar 2009, Pantic & Rothkrantz 2000]. They are mainly classified into two categories, i.e., template-based and feature-based [Pantic & Rothkrantz 2000]. Template-based approaches usually fit a holistic face model to the input image or track it in the input image sequence. Active appearance model [Abboud *et al.* 2004], point distribution model [Lanitis *et al.* 1997], mixture of probabilistic PCA [Zalewski & Gong 2004], and topographic model-

ing [Wang & Yin 2007] are some typical examples. Feature-based approaches generally localize the features of an analytic face model in the input image or track them in the input sequence. Gabor wavelets [Lyons *et al.* 1998] and Local Binary Patterns (LBP) [Shan *et al.* 2009] based face representations are two popular representatives. Although considerable advancements have been achieved, 2D FER is still very challenging mainly due to its sensitivity to illumination, pose variations, and possible occlusions [Kumar 2009, Pantic & Rothkrantz 2000].

Recently, with the rapid development of 3D imaging and scanning technologies, it becomes more and more popular to capture 3D face scans. Comparing with 2D face images, 3D face scans contain precise geometric shape information of facial surfaces, which is robust to illumination and pose variations, but more sensitive to facial expression changes. Thus, shape-based 3D FER has attracted increasing attentions. Similar to 2D, 3D FER approaches can also be categorized into template-based and feature-based. Template-based approaches usually build a parametric deformable face model first, and then extract the model parameters as expression features for recognition. 3D morphable model [Ramanathan *et al.* 2006], bilinear deformable model [Mpiperis *et al.* 2008], shape deformation model [Gong *et al.* 2009], and statistical feature model [Zhao *et al.* 2010] are some famous examples. The main drawback of template-based approaches lies in that they require to establish one-to-one correspondence between 3D face scans, which is still a very challenging issue. Meanwhile, time consuming procedures like dense 3D face registration and model fitting are indispensable. Feature-based approaches generally extract 3D expression cues around facial landmarks using different facial surface geometric or differential quantities. For example, the distances between 3D facial landmarks are widely used in [Soyel & Demirel 2007, Soyel & Demirel 2008, Tang & Huang 2008b], and [Tang & Huang 2008a]. Moreover, 3D facial curves [Maalej *et al.* 2010], facial geometry images and normal maps [Ocegueda *et al.* 2011a, Li *et al.* 2012b], facial conformal images [Zeng *et al.* 2013], facial surface normal [Li *et al.* 2011b, Zhen *et al.* 2015] and curvatures [Li *et al.* 2011b, Zhen *et al.* 2015, Wang *et al.* 2006], and local depth-SIFT features [Berretti *et al.* 2010] are some popular expression features. Feature-

based approaches generally perform better than template-based ones. However, the bottleneck of feature-based approaches lies in accurate and robust 3D facial landmark localization, which is still a very difficult task [Song *et al.* 2014]. More detailed surveys of 3D facial expression recognition are given in [Fang *et al.* 2011, Sandbach *et al.* 2012b].

Although the effectiveness of multimodal 2D + 3D face recognition has been well presented as in [Chang *et al.* 2005, Mian *et al.* 2007], the investigation of multimodal 2D + 3D FER is very limited. [Wang *et al.* 2006] compared the FER accuracy of 3D primitive surface feature distribution based approach with 2D Gabor-wavelet and Topographic Context based ones on the BU–3DFE database, and found that 3D shape based approach is superior to 2D ones, especially for non-frontal faces. However, the effectiveness of combing 3D and 2D approaches was not discussed. [Zhao *et al.* 2010] used both 2D features (RGB values and LBP) and 3D features (3D coordinates and shape index values) in the 3D statistical feature model for prototypical expression recognition. But the results using only 2D features or 3D features were not reported, and thus the complementarity between 2D and 3D features was also not studied. In [Tsalakanidou & Malassiotis 2010], the authors used both 2D and 3D dynamic data for real-time facial action and expression recognition. More precisely, they first extended the active shape model to handle 3D data for facial feature tracking. Then, they extracted numerous geometric measurements (e.g., the distances between landmarks and the boundary shape of lips) and surface deformation measurements (e.g., image gradient and surface curvature descriptors). Finally, the Rule Classifier was used for recognizing a subset of 11 important AUs and 4 facial expressions (i.e., happy, sad, surprise, disgust) on a dataset consisting of 832 sequences of 52 participants. Their experimental results demonstrated that the proposed 2D+3D algorithm performed much better than the 2D appearance-based algorithm (i.e., 2D ASM + Gabor filters + LDA) for recognizing the four facial expressions. This is a very illuminating approach for 2D+3D multimodal FER. However, they did not report the performance of each modality under their own framework. The importance of each modality is still unclear. [Savran *et al.* 2010] utilized multimodal 2D + 3D face data for facial AU detection. They found that 3D

data generally perform better than 2D data, especially for lower AUs. Moreover, the fusion of two modalities can improve the detection rates from 93.5% (2D) and 95.4% (3D) to 97.1% (2D+3D). Except for facial AU detection and expression recognition, [Wang *et al.* 2007] quantified facial expression abnormality in Schizophrenia by combining 2D and 3D features. Their experimental results demonstrated that the combined features better characterized facial expressions than either individual 3D geometric or 2D texture features.

4.2 Contributions

The above studies have preliminarily proved the fact that the combination of 2D and 3D data is better than either of the single 2D or 3D modality for expression characterization and AU detection, but deep analysis of the superiority for multimodal 2D+3D FER is still missing. An advantage of using 2D data is that it can be used to accurately localize a large set of facial landmarks on 2D face images and further on their 3D face scans due to the 2D–3D correspondence, which is the first contribution of this work. More precisely, we propose to explore the incremental Parallel Cascade of Linear Regression (iPar–CLR) algorithm[Huang *et al.* 2014] to automatically localize 49 landmarks for each 2D face image and its corresponding 3D mesh scan. This large set of expression related landmarks are then used for extracting local texture and shape descriptors for expression classification. To the best of our knowledge, this is the first work which uses such large number of automatically detected landmarks for 2D and 3D multimodal FER. In contrast, the majority of existing feature-based 3D FER approaches reported their results on the BU–3DFE benchmark based on a large set of (typically 83) 3D facial landmarks manually localized by the database providers[Li *et al.* 2011b, Wang *et al.* 2006, Berretti *et al.* 2010]. Therefore, the proposed framework presents a promising way to these landmark-based approaches so that they can be made automatic using the iPar–CLR algorithm in 2D and 3D multimodal face space.

The second contribution of this work is that a novel second-order image gradient based local texture descriptor (HSOG), a novel first-order mesh gradient (i.e.,

surface normal) based local shape descriptor (meshHOG), as well as a second-order mesh gradient (i.e., surface curvature) based local shape descriptor (meshHOS) are adapted in FER to comprehensively encode the expression variations in both the 2D and 3D modalities. According to our previous work [Huang *et al.* 2014], most of existing popular local image descriptors, such as HOG, LBP, and SIFT, only employ the first-order gradient information related to the slope and the elasticity, i.e., length, area, etc. When the image is regarded as a surface, and thereby partially characterize its geometric properties. By contrast, HSOG captures the curvature related cues of the surface, i.e., cliffs, ridges, summits, valleys, basins, and so on. Thus, HSOG can be applied to describe facial expression deformations (e.g., mouth stretch, lip stretcher, brow raiser). Moreover, in that work, it was also demonstrated that HSOG outperformed the first-order gradient based local image descriptors (i.e., HOG, LBP, SIFT) when there were not severe scale variations, as in the applications of local image matching and scene classification. In this work, we give another evidence of the effectiveness and generalization ability of HSOG for FER. Similarly, as general local shape descriptors, meshHOG and meshHOS provide a compact description of the facial surface normal and curvature information, and they have proved very efficient for 3D face identification in our previous works [Li *et al.* 2011a, Li *et al.* 2015]. In this work, we interested in exploring their generalization abilities in 3D FER.

During the FER stage, both the early fusion (i.e., feature-level) and late fusion (i.e., score-level) strategies of 2D descriptors, 3D descriptors, as well as 2D and 3D descriptors are comprehensively demonstrated and their complementary characteristics are well revealed, which is our third contribution. The important findings behind the fusion results can be summarized as: 1) The second-order gradient based local texture or shape descriptor (HSOG or meshHOS) generally have stronger discriminative power than the first-order gradient based ones (SIFT or meshHOG). Moreover, different order 2D or 3D descriptors are complementary in encoding local texture or shape cues. 2) There exist large complementary characteristics between 2D and 3D descriptors of the same order (SIFT and meshHOG, HSOG and mesh-HOS), different order (SIFT and meshHOS, HSOG and mesh-HOG), as well as

multiple orders (all four 2D and 3D descriptors).

Overall, we present an efficient multimodal (2D and 3D) and multiple-order (first and second) feature-based fully automatic FER approach, and validate it trough comprehensive experiments on the BU–3DFE database. Considerable complementary characteristics between the features of different orders and different modalities are highlighted either by early fusion or late fusion of 2D, 3D, as well as 2D and 3D descriptors. The generalization capability of our approach is further evaluated on the Bosphorus database.

The remainder of this chapter is organized as follows. Section 4.3 show the framework of the proposed system. The landmark detection is described in Section 4.4. The construction of local texture descriptions and local shape descriptors are shown respectively in section 4.5 and section 4.6. Experimental results are described and analyzed in section 4.7. Section 4.8 concludes this chapter.

4.3 Overview of the proposed system

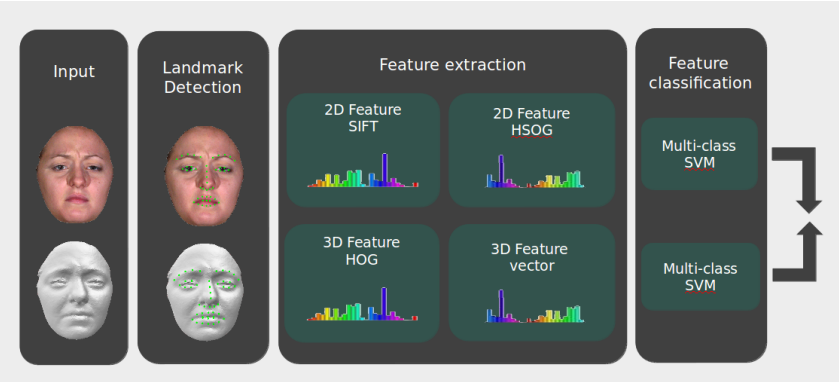


Figure 4.1: The overview of proposed system

Actually, in this work, we propose a novel multi-modalites based methods for facial expression recognition. As shown in Figure 4.1, the pipeline of proposed methods is defined as follows:

1. Firstly, given 3D face scans and their correspondent textures images, a standard preprocessing is carried out to remove spikes and fill holes. Then, an

algorithm of Iterative Closest Points (ICP) is further employed for face alignment.

2. Secondly, a joint 2D and 3D facial landmark localization algorithm is employed to automatically locate 68 landmarks on both 3D meshes and their correspondent texture images.

3. Thirdly, On both modalities, we extract first-order and second-order feature descriptors. From textures images, SIFT and HSOG are firstly computed, while from 3D meshes, two novel descriptors, HOG and HOS, are firstly introduced for highlighting the discrimination between different facial expressions.

4. Finally, a set of SVM classifiers is used to classify four kinds of feature vectors extracted above. The outputs of SVM for different feature on different modalities are further integrated for improving the final results.

In the following sections, we introduce in detail these steps of pipeline subsequently.

4.4 Joint 2D and 3D facial landmark localization

To extract expression related features, a set of key landmarks are required. In this work, we introduce the incremental Parallel Cascade of Linear Regression (iPar–CLR) [Asthana *et al.* 2014] for face landmarking in the 2D modality. iPar–CLR is an incremental and parallel version of the Sequential Cascade of Linear Regression (Seq–CLR) algorithm [Xiong & De la Torre 2013]. Given a set of training face images I_i associated with p 2D landmarks $x^i \in \mathbb{R}^{2p\times 1}$. f is a feature extraction function (e.g., SIFT) and $f(x^i) \in \mathbb{R}^{128p\times 1}$ in the case of extracting SIFT features. During training, one assumes that p corrected landmarks are known for each I_i , and denoted as x^i_*. To reproduce the testing scenario, one runs the face detector on the training images to provide an initial configuration of the p landmarks x^i_0, which corresponds to an average shape. In this setting, the Seq–CLR algorithm is

formulated as:

$$\arg\min_{W_k,b_k} \sum_{I_i} \sum_{x_k^i} \|x_*^i - x_k^i - W_k f(x_k^i) - b_k\| \tag{4.1}$$

In practice, W_0 and b_0 are first estimated using x_*^i, x_0^i, and $f(x_0^i)$. Then, a sequence of regressions are computed to update x_k^i and make it converge to x_*^i step by step. iPar–CLR improves Seq–CLR by introducing a parametric 3D shape model for the configuration of p landmarks, and solving Eq.(4.1) in the parameter space. By assuming that the distribution of the perturbations of shape parameters is Gaussian, iPar–CLR is well suited for the task of incremental update. That is, it can incrementally update the pre-trained shape model according to the newly added face images.

Figure 4.2: iPar–CLR based 2D landmark localization. 49 2D landmarks are localized on the projected 2D texture face images of the BU–3DFE database with different genders, ethnicities, ages, and expressions (from left to right, anger, disgust, fear, happiness, sadness and surprise).

When used for joint 2D and 3D facial landmark localization, the texture map is projected from each textured 3D face scan into a 2D regular grid domain using the

interpolation techniques. Then, we apply iPar–CLR to each projected 2D texture face image, outputting 49 2D landmarks (see Fig. 4.2). These 2D landmarks are then transferred to 3D texture face space by the inverse of the above projection. Note that since all these 2D landmarks are located at the frontal part of the projected 2D face texture, the one-to-one correspondence between 3D texture and 2D texture can be approximately preserved during the projection mapping. Finally, the corresponding 3D landmarks are directly determined by the one-to-one correspondence between 3D texture and 3D geometry of the 3D face model. We evaluate iPar–CLR on the whole BU–3DFE database and the expressive samples in Bosphorus, and find that it can precisely localize all the pre-defined 49 facial landmarks for all samples even with variations in expression, ethnicity, gender and age etc. (see Fig. 4.2 for some sampled results).

4.5 Construction of local 2D texture descriptors

4.5.1 First-order gradient based local texture descriptor: SIFT

We extract the SIFT [Lowe 2004] descriptor of each projected 2D texture face image at the locations of the detected 2D landmarks within 16×16 patches. The SIFT feature based facial representation of a 2D texture image is generated by concatenating all the SIFT features at the 49 landmarks according to the pre-defined order, resulting in a $128 \times 49 = 6,272$ dimensional feature vector. This vector is further normalized to the unit length for the following processing.

4.5.2 Second-order gradient based local texture descriptor: HSOG

The HSOG descriptor was originally proposed in [Huang *et al.* 2014] and proved very efficient for local image matching, object categorization, and scene classification. In this paper, we explore HSOG for 2D facial expression description. The construction of HSOG is composed of three steps:

(1) *Computation of the first order Oriented Gradient Maps (OGMs)*: The input of HSOG is a $R \times R$ image patch around each localized 2D facial landmark. For each image patch $I(x, y)$, it outputs a number of Oriented Gradient Maps (OGMs)

$\{J_o(x,y)\}_{o=1}^{L}$ by computing the Gaussian convolution of the positive orientation gradient maps, described as:

$$J_o(x,y) = G_\sigma \times \max(\frac{\partial I(x,y)}{\partial o}, 0), o = 1, 2, \ldots, L. \tag{4.2}$$

where o represents a quantized direction, and G_σ is a Gaussian kernel with standard deviation σ, which is proportional to the size of image patch R.

(2) *Computation of the second order gradients*: Once these first order OGMs of all quantized directions are generated, they are used as the inputs for computing the second order gradients. Precisely, for each OGM $J_o(x,y)$, we calculate its gradient magnitude $mag_o(x,y)$ and orientation $\theta_o(x,y)$ at every pixel location. The orientation value $\theta_o(x,y)$ is then re-scaled from the range of $[-\pi/2, \pi/2]$ to $[0, 2\pi]$, and quantized into L dominant orientations. After quantization, the entry n_o of each orientation θ_o is calculated as:

$$n_{\theta_o} = \mod(\lfloor\frac{\theta_o(x,y)}{2\pi/L} + \frac{1}{2}\rfloor, L), o = 1, 2, \ldots, L. \tag{4.3}$$

(3) *Spatial pooling*: Daisy-style spatial pooling strategy is used in HSOG as illustrated in Fig. 4.3. It is easy to find that there are four parameters that determine the HSOG descriptor, i.e., the size of the patch (R); the number of quantized orientations (L); the number of concentric rings (CR); the number of circles on each ring (C). The total number of the divided circles can be calculated as $T = CR \times C + 1$. Within each circle CIR_j, and for each OGM J_o, a second order gradient histogram is constructed by accumulating the gradient magnitudes mag_o of all the pixels with the same quantized orientation entry n_o.

$$h_{oj}(i) = \sum_{(x,y)\in CIR_j} \delta(n_{\theta_o}(x,y) == i) \times mag_o, \tag{4.4}$$

where $i = 0, 1, \ldots, L-1$; $o = 1, 2, \ldots, L$; $j = 1, 2, \ldots, T$, and δ is the characteristic function. Then, for each first order OGM J_o, its second order gradient histogram

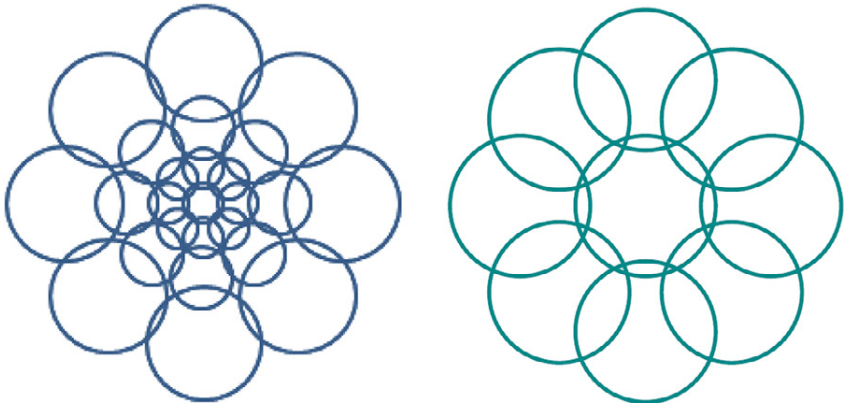


Figure 4.3: The daisy-style spatial pooling. Left: three concentric rings, and each with eight circles in HSOG. Right: one concentric ring with eight circles in meshHOG and meshHOS

h_o is generated by concatenating all the histograms from T circles:

$$h_o = [h_{o1}, h_{o2}, \ldots, h_{oT}]^T, \tag{4.5}$$

where $o = 1, 2, \ldots, L$. Finally, the HSOG descriptor is obtained by concatenating all L histograms of the second order gradients as in Eq. 4.6. Each histogram h_o is normalized to an unit norm vector $\hat{h_o}$ before concatenation.

$$HSOG = [\hat{h_1}, \hat{h_2}, \ldots, \hat{h_T}]^T, \tag{4.6}$$

Similar to SIFT, the HSOG feature based expression representation of a 2D texture image is generated by concatenating all HSOG features of the localized landmarks and then normalized to the unit length. In this paper, we set $R = 25$, $L = 8$, $CR = 3$, $C = 4$ as in [Huang *et al.* 2014]. Thus, the dimension of the final HSOG feature vector for a face image is $T \times L \times 8 \times 49 = 13 \times 8 \times 8 \times 49 = 40,768$.

4.6 Construction of local 3D shape descriptors

The meshHOG and meshHOS descriptors were originally proposed in [Li *et al.* 2011a, Li *et al.* 2015] and proved efficient in 3D face identification. In this work, we employ them in 3D FER. Similar to HSOG, meshHOG and

meshHOS are built by the following three steps:

(i) *Computation of facial surface normal and curvature*: Each 3D facial surface is represented by a triangular mesh $T = (F, V)$, where F and V are the face and vertex sets. We compute the unit normal vector of each face by the cross product of its two edge vectors. Then the unit normal of each vertex $n^v = [n_x^v, n_y^v, n_z^v]^T$ is achieved by averaging the normal vectors of its one-ring faces. The mesh gradient magnitude mag v and orientation θ_v at each vertex are calculated as:

$$mag_v = \sqrt{(n_x^v/n_z^v)^2 + (n_y^v/n_z^v)^2}, \qquad \theta_v = \arctan(n_y^v/n_x^v). \tag{4.7}$$

According to [Goldfeather & Interrante 2004], the principal curvatures k_{max} and k_{min} are computed by fitting a cubic-order surface:

$$f(x, y) = \frac{A}{2}x^2 + Bxy + \frac{C}{2}y^2 + Dx^3 + Ex^2y + Fxy^2 + Gy^3 \tag{4.8}$$

and its normal vectors $(f_x(x; y), f_y(x; y), -1)$ using both the 3D coordinates and the normal vectors of the associated local neighbor points (two-ring). Once we have two principle curvatures, the shape index values, which describe different shape classes by a single number ranging from 0 to 1, is calculated as:

$$SI = \frac{1}{2} - \frac{1}{\pi}\arctan(\frac{k_{max} + k_{min}}{k_{max} - k_{min}}). \tag{4.9}$$

Fig. 4.4 shows the shape index maps of sampled 3D faces with six prototypical expressions.

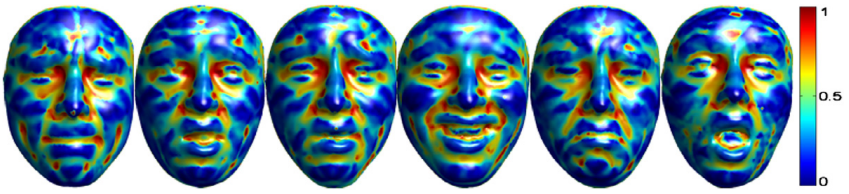


Figure 4.4: The shape index maps of sampled 3D faces with six prototypical expressions (from left to right, anger, disgust, fear, happiness, sadness, and surprise).

(ii) *Canonical orientation(s) assignment*: Similar to the SIFT feature, to achieve

rotation invariance, one or more local coordinate systems (i.e., canonical orientations) should be determined at each localized 3D landmark. This can be accomplished by the following three steps: First, we build an initial local coordinate system, where the landmark v and its normal n^v are the origin and the positive z axis, respectively. And two perpendicular vectors in the tangent plane of v are randomly chosen as the $x$axis and y axis, respectively. Then, the gradient magnitudes and orientations of the vertices around the landmark with a given geodesic distance r_0 are computed, Gaussian-weighted by their corresponding gradient magnitudes, and put in a histogram of 360 bins. Dominant gradient orientations, that is, peaks in the histogram, are used to assign one or more canonical orientations to the landmark. Finally, the initial local coordinate system is rotated in the local tangent plane, making each canonical orientation as new x axis, and the new y axis is computed by the cross product of the z axis (i.e., normal vector n^v) and the new x axis (i.e., canonical orientation). Once the canonical orientations are determined, all the neighbor vertices and their normal vectors are transformed to the new local coordinate system for the following processes.

(iii) *Spatial pooling*: Similar to the HSOG feature, a simplified daisy-style spatial pooling strategy is also used for meshHOG and meshHOS. However, the pooling strategy here is performed on the tangent plane of each 3D landmark and on the local coordinate system determined by the assigned canonical orientations. As illustrated in Fig. 4.3, for the 3D descriptors, there is only one concentric ring associated with eight circles, resulting in nine sequential circles. Within each circle CIR_j, a mesh gradient histogram and a shape index histogram are constructed respectively. The histogram of mesh gradient is constructed by accumulating the gradient magnitudes mag_v of all vertices with the same quantized orientation entry $n_\theta(v)$ as:

$$hog_j(i) = \sum_{v \in CIR_j} \delta(n_\theta(v) == i) \times mag_v, \tag{4.10}$$

where $i = 0, 1, \ldots, 7$; $j = 1, 2, \ldots, 9$; $n_\theta(v)$ is entry of the quantized gradient orientation computed the same as $n_{\theta_o}(x, y)$ in Eq. 4.3. The histogram of shape index is constructed by accumulating the Gaussian weights $G_\Sigma(v)$ of all vertices

with the same quantized shape index value $n_{SI}(v)$:

$$hos_j(i) = \sum_{v \in CIR_j} \delta(n_\theta(v) == i) \times G_\Sigma(v), \tag{4.11}$$

where $i = 0, 1, \ldots, 7$; $j = 1, 2, \ldots, 9$; $n_{SI}(v)$ is the quantized shape index values. Then, for each 3D landmark, its 3D descriptors are generated by concatenating all the histograms from nine circles in a clockwise direction,

$$HOG = [hog_1, hog_2, \ldots, hog_9]^T, \qquad HOS = [hos_1, hos_2, \ldots, hos_9]^T. \tag{4.12}$$

Each sub-histogram (hog_i or hos_i) is normalized to the unit length before concatenation to eliminate the influence of non-uniform mesh sampling. Note that, intuitively, HOG describes the point-level bending pattern of the local shape around a landmark while HOS indicates the distribution of different shape categories. The expression representation (based on meshHOG or meshHOS) of a 3D face surface is generated by concatenating all HOG or HOS features of the localized 3D landmarks and then normalized to the unit length. Following [Li *et al.* 2011a], the geodesic radius r_0 is set to 22.50 mm, the radius of each circle is set to 10 mm, and the distance between the center of the centric circle and the one of each rounding circle is set to 15 mm. As a result, the dimension of the final meshHOG or meshHOS feature is $9 \times 8 \times 49 = 3528$.

4.7 Experimental results

4.7.1 The BU–3DFE database

We make use of the widely used BU–3DFE database [Yin *et al.* 2006b] to evaluate the proposed multimodal 2D + 3D local feature-based FER approach. This database consists of 2,500 textured 3D face scans of 100 persons in different expression, gender, race, and age. Six prototypical facial expressions (anger, disgust, fear, happiness, sadness, and surprise) with four intensity levels plus a neutral expression are displayed for each person. Examples of some projected 2D texture face images

in BU–3DFE database are shown in Fig. 4.2

4.7.2 Experimental setup

To fairly conduct the identity–independent FER, we use the evaluation protocol in [Gong *et al.* 2009]. More precisely, we randomly select 60 persons, and keep the samples with the six prototypical facial expressions of two highest intensity levels. That is, $60\times6\times2 = 720$ samples are used for training and testing in total. Then, 648 samples of 54 persons (90%) and 72 of 6 persons (10%) are randomly divided for the training and testing data partition. To achieve stable recognition accuracy, this kind of 10-fold subject-independent cross-validation is conducted 100 rounds for all of our experiments. Based on these data partition strategies and the constructed 2D and 3D features, we utilize the SVM classifier with the Radial Basis Function (RBF) kernel for expression classification. The parameters for SVMs are tuned according to the 10-fold cross-validation in the training sets. To find the complementary characteristics between 2D descriptors, 3D descriptors, as well as 2D and 3D descriptors, we conduct both the early fusion (feature-level) and late fusion (score-level). For early fusion, the fused feature is generated by simply concatenating different descriptors. For late fusion, the mean of the recognition accuracies of different descriptors are used as the final accuracy.

4.7.2.1 Local 2D texture descriptors and their fusion

Table 4.1 shows the average expression recognition accuracies achieved using the single 2D descriptors and their fusion. From this table, we can see that: i) The average accuracies of the HSOG descriptor are much better than the ones of SIFT for anger and sadness, and comparable for the other expressions. ii) Early fusion largely improves the average accuracies of anger and sadness for SIFT, but also largely impairs the one of sadness for HSOG. iii) Late fusion generally performs better than early fusion, especially for the fear and sadness expressions. iv) Overall, the average accuracy of HSOG is 84.49%, which is better than SIFT (81.85%), and even slightly better than the ones of early fusion (82.85%) and late fusion

(84.29%). We can conclude that the second-order gradient based local texture descriptor (HSOG) has more powerful discriminative ability than the popular first-order gradient based one (SIFT) for local texture-based FER. Moreover, there also exists some complementarity between different order descriptors for some specific expressions (e.g., anger and fear).

Table 4.1: Average confusion matrices of SIFT, HSOG, and their early and late fusions on BU–3DFE database.

	SIFT (81.85)						HSOG (84.49)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**76.55**	6.53	2.83	0	14.09	0	**83.09**	4.24	3.21	0	9.37	0.08
DI	5.16	**84.51**	2.37	1.28	2.39	4.29	5.75	**85.00**	2.14	2.41	2.46	2.24
FE	3.64	6.41	**72.61**	5.73	8.85	2.76	1.06	5.63	**72.41**	9.48	8.35	3.07
HA	0	0.98	8.77	**89.37**	0	0	0.79	2.45	6.02	**89.82**	0.03	0.90
SA	20.25	1.43	7.29	0	**70.71**	0.32	12.82	3.42	3.57	0	**80.20**	0
SU	0.01	0.04	1.12	0.64	0.82	**97.38**	0	0.23	1.74	0.75	0.82	**96.47**
	Early fusion: SIFT + HSOG (**82.86**)						Late fusion: SIFT + HSOG (**84.29**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**83.91**	4.29	2.86	0	8.87	0.08	**82.16**	5.05	2.92	0	9.87	0
DI	6.12	**83.27**	2.87	1.68	1.66	4.42	5.81	**83.87**	2.56	1.64	2.22	3.9
FE	1.42	7.73	**70.82**	8.60	8.35	3.07	1.16	6.48	**74.40**	7.38	7.80	2.77
HA	0.03	2.59	8.28	**88.20**	0	0.90	0.02	1.70	7.52	**89.86**	0	0.90
SA	19.06	2.00	5.03	0	**73.91**	0	14.67	2.45	4.48	0	**78.42**	0
SU	0	0	2.13	0.01	0.82	**97.05**	0	0.07	1.33	0.72	0.82	**97.07**

4.7.2.2 Local 3D shape descriptors and their fusion

Table 4.2 shows the average expression recognition accuracies achieved using the single 3D descriptors and their fusion. From this table, we can find that: i) Except anger expression, meshHOS achieves better results than meshHOG, especially for happiness. ii) Early fusion and late fusion generally improve the accuracies of both 3D descriptors for all expressions except happiness with a slight drop in early fusion. iii) Overall, the average recognition accuracy of meshHOS is 80.55%, which is better than meshHOG (77.62%), and late fusion (82.70%) is superior to early fusion (81.23%). We can conclude that the second-order surface gradient-based local shape descriptor (meshHOS) has stronger discriminative capability than the first-order surface gradient-based one (meshHOG). Moreover, they also contain some com plementary information when classifying some specific expressions (e.g., sadness and surprise).

Table 4.2: Average confusion matrices of meshHOG, meshHOS, and their early and late fusions on BU–3DFE database.

	meshHOG (77.62)						meshHOS (80.55)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**79.75**	2.47	2.60	1.48	13.70	0	**77.96**	4.36	2.16	1.17	14.36	0
DI	4.11	**75.00**	8.89	4.92	2.17	4.91	3.34	**78.99**	6.78	3.83	3.31	3.74
FE	3.39	7.32	**66.23**	14.44	4.30	4.33	0.92	6.63	**69.50**	13.27	4.69	4.99
HA	0.77	0.59	15.52	**80.79**	0	2.33	0	0.37	9.85	**88.38**	1.40	0
SA	22.58	2.09	2.91	0	**72.32**	0.11	18.28	3.01	4.15	0	**74.52**	0.04
SU	0	1.01	7.38	0.01	0	**91.61**	0	2.08	3.98	0	0	**93.93**
	Early fusion: meshHOG + meshHOS (**81.23**)						Late fusion: meshHOG + meshHOS (**82.70**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**80.93**	2.56	2.56	1.45	12.50	0	**82.63**	2.48	2.56	1.48	10.85	0
DI	4.43	**80.12**	5.55	4.76	2.01	3.13	3.84	**80.97**	5.13	4.62	1.93	3.51
FE	1.29	6.18	**71.06**	13.08	3.29	5.10	1.67	5.72	**72.08**	12.22	3.44	4.87
HA	0	1.18	13.79	**85.03**	0	0	0	0.68	12.02	**87.21**	0.09	0
SA	19.13	1.38	2.82	0	**76.67**	0	15.78	1.57	3.91	0	**78.74**	0
SU	0	0.13	6.32	0.01	0	**93.54**	0	0.04	5.38	0	0	**94.57**

4.7.2.3 Local multimodal 2D + 3D descriptors and their fusion

In this section, we indicate that the local 2D texture and 3D shape descriptors contain strong complementary characteristics, and thus their fusion largely improves the expression recognition accuracies.

Table 4.3 lists the average expression recognition results of fusing the same order gradient-based 2D and 3D descriptors lead increase performance. Compared with the results in Table 4.1 and 4.2, we can see that both early fusion and late fusion of the same order gradient-based 2D and 3D descriptors are very efficient, especially for the case of late fusion, with an improvement up to 3% for SIFT, 7% for mesh-HOG, 2.3% for HSOG, and 6.3% for meshHOS in the average accuracy. Moreover, the improvement of sadness expression is up to 8% for meshHOG and 10% for SIFT. And the accuracies of happiness are improved about 5% for HSOG and 6% for meshHOS.

Table 4.4 shows the average expression recognition results of fusing different order gradient-based 2D and 3D descriptors. Compared with the results in Table 4.1and 4.2, we can find that the fusion of the different order gradient-based 2D and 3D descriptors is also very efficient except the case of early fusion of HSOG and meshHOG. Take the results of late fusion as an example, the average recognition accuracies are improved by 3.2% for SIFT, 5% for meshHOS, 1.3% for HSOG and 8.1% for meshHOG. In particular, the improvement of happiness expression is

Table 4.3: The effectiveness of fusing the same order gradient-based 2D and 3D descriptors on BU–3DFE database.

	Early fusion: SIFT + meshHOG (**83.68**)						Late fusion: SIFT + meshHOG (**84.91**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**82.09**	5.25	1.83	0.08	10.75	0	**83.14**	4.07	1.59	0.56	10.63	0
DI	5.67	**84.75**	4.00	1.33	0.08	4.17	6.38	**82.65**	3.47	2.04	0.75	4.68
FE	1.50	6.00	**71.33**	11.00	6.67	3.50	0.20	6.63	**74.02**	8.06	7.85	3.24
HA	0	1.08	8.67	**89.50**	0	0.75	0	0.87	6.98	**91.74**	0	0.41
SA	17.33	1.08	3.42	0	**78.17**	0	15.28	0.91	3.36	0	**80.45**	0
SU	0	0.58	2.83	0	0.33	**96.25**	0	0.04	1.75	0	0.77	**97.43**
	Early fusion: HSOG + meshHOS (**84.49**)						Late fusion: HSOG + meshHOS (**86.80**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**83.50**	5.75	2.17	0	8.58	0	**86.10**	3.23	1.64	0.33	8.70	0
DI	5.42	**85.67**	1.42	2.17	2.33	3.00	4.96	**85.45**	1.66	2.29	2.48	3.17
FE	0.25	6.92	**70.58**	8.08	9.75	4.42	0.05	4.31	**75.87**	9.33	6.77	3.67
HA	0	2.08	6.25	**91.67**	0	0	0.02	1.08	3.47	**94.85**	0.59	0
SA	14.00	2.50	4.42	0	**79.08**	0	13.47	0.75	4.40	0	**81.38**	0
SU	0	0.08	3.17	0	0.33	**96.42**	0	0.01	1.96	0.07	0.82	**97.15**

Table 4.4: The effectiveness of fusing different order gradient-based 2D and 3D descriptors on BU–3DFE database.

	Early fusion: SIFT + meshHOS (**85.15**)						Late fusion: SIFT + meshHOS (**85.07**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**85.42**	5.75	2.25	0	6.58	0	**80.67**	5.08	1.83	0.08	12.33	0
DI	5.00	**86.92**	0.92	1.75	1.50	3.92	4.67	**86.33**	2.17	0.83	2.50	3.50
FE	0	6.25	**73.67**	6.42	9.58	4.08	0	6.92	**75.50**	7.08	7.33	3.17
HA	0	1.00	7.33	**91.67**	0	0	0	1.00	5.08	**93.92**	0	0
SA	16.83	1.25	5.75	0	**76.17**	0	15.58	0.83	6.92	0	**76.67**	0
SU	0	0.25	1.92	0	0.75	**97.08**	0	0	1.83	0	0.83	**97.33**
	Early fusion: HSOG + meshHOG (**83.17**)						Late fusion: HSOG + meshHOG (**85.75**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**81.00**	4.33	1.83	1.00	11.83	0	**82.58**	3.92	1.92	1.50	10.08	0
DI	4.75	**85.83**	3.67	2.58	1.17	2.00	5.33	**86.67**	1.67	2.50	1.33	2.50
FE	1.17	5.92	**72.75**	10.25	6.25	3.67	0	6.58	**74.83**	8.25	7.00	3.33
HA	0.08	1.83	9.58	**87.83**	0	0.67	0	2.42	6.00	**90.17**	0	1.42
SA	19.17	1.08	3.50	0	**76.25**	0	11.42	1.17	4.17	0	**83.25**	0
SU	0	1.08	3.58	0	0	**95.33**	0	0	2.08	0	0.92	**97.00**

5.5% in the case of fusing SIFT and meshHOS. And the accuracy of the sadness expression is improved up to 11% when lately fusing HSOG and meshHOG.

As reported in Table 4.5, when considering the fusion of all the first-order and second-order gradient-based local 2D texture and 3D shape descriptors, our approach achieves an average recognition accuracy of 85.92% for early fusion and 86.32% for late fusion. These scores largely outperform the ones achieved by only fusing 2D descriptors (82.86% and 84.29%) in Table 4.1 or 3D descriptors (81.23% and 82.70%) in Table 4.2. More precisely, the confusion matrices of these scores indicate that the 2D descriptors and 3D descriptors have strong complementary characteristics for all the six prototypical facial expressions.

Table 4.5: The effectiveness of fusing all four gradient-based 2D and 3D descriptors on BU–3DFE database.

	Early fusion: all features (**85.92**)						Late fusion: all features (**86.32**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**86.33**	3.91	1.58	0.06	8.13	0	**85.56**	3.88	1.58	0.17	8.81	0
DI	5.78	**84.27**	2.19	2.32	0.98	4.47	5.17	**84.35**	2.00	2.50	2.03	3.95
FE	0.02	4.06	**75.03**	10.93	6.41	3.56	0	4.64	**75.77**	9.26	7.02	3.30
HA	0	0.99	7.15	**91.86**	0	0	0	0.92	5.62	**93.42**	0	0.05
SA	14.69	0.43	3.52	0	**81.36**	0	14.28	0.92	3.55	0	**81.26**	0
SU	0	0	2.52	0	0.82	**96.67**	0	0	1.63	0	0.82	**97.55**

Table 4.6: Performance comparison with the state-of-the-art methods on BU–3DFE database.

Method	Modality	Landmark	Classifier	Accuracy in protocol (%)		
				I	II	III
[Wang *et al.* 2006]	2D/3D	64 manual	LDA	83.60	61.79	-
[Soyel & Demirel 2007]	3D mesh	11 manual	NN	91.30	67.52	-
[Soyel & Demirel 2008]	3D mesh	83 manual	NN	93.72	-	-
[Tang & Huang 2008b]	3D mesh	83 manual	LDA	95.10	74.51	-
[Tang & Huang 2008a]	3D mesh	83 manual	SVM	87.10	-	-
[Mpiperis *et al.* 2008]	3D mesh	global registration	ML	90.50	-	-
[Gong *et al.* 2009]	3D depth	global registration	SVM	-	76.22	-
[Zhao *et al.* 2010]	2D+3D	19 automatic	BBN	82.30	-	-
[Berretti *et al.* 2010]	3D depth	27 manual	SVM	-	-	77.54
[Lemaire *et al.* 2011]	3D mesh	21 automatic	SVM	-	75.76	-
[Li *et al.* 2012b]	3D depth	global registration	MKL	-	-	80.14
[Zeng *et al.* 2013]	3D depth	3 automatic	SRC	-	-	70.93
[Zhen *et al.* 2015]	3D mesh	global registration	SVM	-	84.50	83.20
[Yang *et al.* 2015b]	3D mesh	global registration	SVM	-	84.80	82.73
Our method	2D+3D	49 automatic	SVM	-	86.32	-

4.7.2.4 Comparison with other methods

To validate the effectiveness of the proposed method in FER, we compare it with the state-of-the-art methods on the BU–3DFE dataset. To give a comprehensive

analysis, four aspects, including the data modality, facial landmark, expression classifier, and recognition accuracy are compared.

From Table 4.6, we find that all previous methods (except [Wang *et al.* 2006] and [Zhao *et al.* 2010]) reported their FER accuracies on BU–3DFE using only 3D modality data. As mentioned in Section previously, the results of 2D and 3D data are separately reported in [Wang *et al.* 2006] and jointly reported in [Zhao *et al.* 2010]. Complementarity analysis of 2D and 3D data in FER is missing. On facial landmark, early studies such as [Wang *et al.* 2006, Soyel & Demirel 2008, Tang & Huang 2008b, Tang & Huang 2008a] and [Berretti *et al.* 2010] rely on a large number of manual landmarks. Recent studies try to avoid this impractical framework by utilizing global registration algorithms (e.g., [Mpiperis *et al.* 2008, Gong *et al.* 2009, Li *et al.* 2012b, Zhen *et al.* 2015]), or building general face models (e.g., [Zhao *et al.* 2010, Lemaire *et al.* 2011]). Our method solves this problem by exploring the iPar–CLR algorithm to jointly detect a large number of 2D and 3D landmarks. For expression classification, SVM is the most popular classifier compared with the others such as Neutral Networks (NN), Sparse Representation-based Classifier (SRC), Bayesian Belief Net (BBN), and Multiple Kernel Learning (MKL).

In the literature, there are three FER protocols on BU–3DFE. Early tasks (e.g., [Zhao *et al.* 2010, Soyel & Demirel 2008, Wang *et al.* 2006]) chose 60 subjects and average the accuracies of one or two rounds of 10-fold cross-validation, totally with 10 or 20 times of train and test sessions (denoted by protocol I). This protocol has proved very sensitive to the identity variations of training and testing samples [Gong *et al.* 2009]. Gong et al. in [Gong *et al.* 2009] later suggested to choose 60 subjects and average the accuracies of 100 rounds of 10-fold cross-validation, resulting in 1000 times of train and test sessions in total (i.e., protocol II). A similar protocol (i.e., protocol III) [Berretti *et al.* 2010], randomly chose 60 subjects in each round of 10-fold cross-validation and average the accuracies of 100 rounds. From Table 4.6, we can find that the accuracies of the same methods [Soyel & Demirel 2007, Tang & Huang 2008b, Wang *et al.* 2006] dropped more

than 20% from protocol I to protocol II. Moreover, the accuracies of the same method achieved by protocol II and protocol III were close to each other as shown in [Zhen *et al.* 2015] and [Yang *et al.* 2015b]. Our proposed multimodal 2D+3D local feature-based approach reaches the highest average accuracy (86.32%) in protocol II.

4.7.2.5 Generalization capability on Bosphorus database

In this section, we study the generalization capability of our proposed approach on the Bosphorus database. This database contains 4666 textured 3D face models of 105 subjects in various facial expressions, action units, poses and occlusions. To fairly conduct the identity-independent facial expression recognition, we still use the experimental protocol in [Gong *et al.* 2009] (i.e., protocol II). That is, we randomly select 60 persons who display all the six prototypical facial expressions. Totally, there are 60×6 = 360 samples used for training and testing. And 324 samples of 54 persons (90%) and 36 of 6 persons (10%) are randomly divided for the training and testing data partition. This kind of 10-fold cross-validation is conducted 100 rounds to achieve stable recognition accuracies, and the results are listed in Table 4.7.

Table 4.7: The average accuracies and confusion matrices (in %) on Bosphorus database.

	SIFT (**82.89**)			HSOG (**80.31**)			meshHOG (**65.38**)			meshHOS (**74.94**)		
	late fusion of SIFT + meshHOS (**84.44**)						late fusion of HSOG + meshHOS (**83.56**)					
	early fusion of 2D+3D descriptors (**84.33**)						late fusion of 2D+3D descriptors (**84.72**)					
%	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
AN	**83.00**	5.83	0.17	0	11.00	0	**82.33**	6.67	0	0	11.00	0
DI	5.83	**82.50**	5.67	1.33	4.67	0	5.83	**82.83**	5.17	1.33	4.67	0.17
FE	1.17	1.67	**69.83**	1.50	4.00	21.83	0.17	3.17	**72.33**	2.17	3.67	18.50
HA	0	0.17	0	**99.83**	0	0	0	0	0	**100**	0	0
SA	3.83	8.00	0	0	**88.17**	0	4.33	6.67	0	0	**89.00**	0
SU	0	0	17.33	0	0	**82.67**	0	0	18.17	0	0	**81.83**

Compare the results in Table 4.7 with the ones achieved on the BU–3DFE dataset, we can see that: 1) except the SIFT descriptor, the accuracies of other 2D and 3D descriptors are decreased. For example, the performance of meshHOG is dropped from 77.62% on BU–3DFE to 65.39% on Bosphorus. 2) The fusion of 2D and 3D descriptors is still efficient such as the late fusion of SIFT and meshHOG, HSOG and meshHOS. 3) Comparable expression recognition accuracies (86.32%

vs. 84.72%) are achieved on the two datasets when fusing all the 2D and 3D local descriptors. 4) Compare with the results in Table 4.5, the accuracies for happiness and sadness are much better on Bosphorus, while the ones for surprise are much better on BU–3DFE. The possible reasons resulting in 1) and 4) come from the large expression variations of different persons when they displaying the same expression. Noted that all the persons in Bosphorus are professional actors or actress, while the subjects in BU–3DFE are ordinary people such as the university students. As shown in Fig. 4.5, sadness and anger look very similar for some people, and fear is always with month opening, which makes fear and surprise are largely confused with each other. Moreover, the disgust expression is very special and diversiform, which makes it confusing with sadness, anger and fear. It is probably the reason that most anger samples are misclassified into sadness, and most surprise samples are misclassified to fear and vice versa as shown in the average confusion matrices in Table 4.7.

Figure 4.5: Examples of expression pairs with similar expression configurations but different expression labels in the Bosphorus database. The expression labels of the bottom three pairs are: anger and disgust, fear and disgust, sadness and disgust.

4.7.2.6 Complementarity analysis between 2D and 3D descriptors

To illustrate the complementary characteristics between 2D and 3D multimodal descriptors, we perform the Gentle AdaBoost algorithm [Friedman *et al.* 2000] on the HSOG and meshHOS descriptors to select the most discriminative 2D and 3D facial landmarks (i.e., local regions used to compute HSOG or meshHOS) on BU–

3DEF. More precisely, in each iteration of the Gentle AdaBoost algorithm, each landmark associated descriptor is first fed into a logistic regression weak classier, and the one with the lowest error rate is chosen as the most discriminative one in current iteration. Then, the weights of all the samples (landmarks) are updated, making the algorithm pay more attention on the misclassified samples. Finally, the algorithm stops when the top N discriminative landmarks are selected. Fig. 4.6 shows the top 15 most discriminative landmarks automatically selected by this algorithm. From this figure, it is not difficult to find that the distributions of the top 15 most discriminative 2D and 3D facial landmarks are largely different from each other for all the six sampled facial expressions. This finding once again indicates that our proposed 2D and 3D multimodal local texture and shape descriptors indeed have strong complementary characteristics.

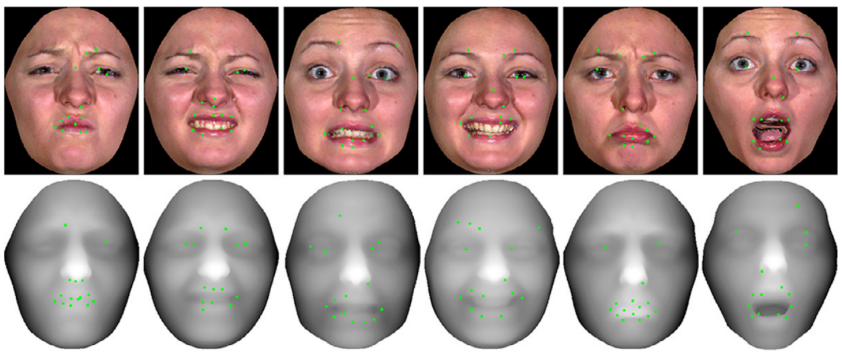


Figure 4.6: Illustration of the complementary characteristics between the 2D and 3D local descriptors. The top 15 most discriminative 2D and 3D landmarks are automatically selected by the Gentle AdaBoost algorithm from 2D and 3D face samples with different expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) on BU–3DFE. Note that depth images, instead of 3D meshes, are displayed for better visualization.

4.8 Conclusion and further work

In this work, we present an efficient multimodal 2D + 3D feature based approach for automatic FER. Based on the iPAR–CLR algorithm, we automatically localize 49 2D facial landmarks, and their corresponding 3D facial landmarks. Around each landmark, the HSOG based local texture descriptor and the SIFT descriptor

are integrated for local 2D facial texture description. Furthermore, two mesh-based local shape descriptors, which consider both the first-order (surface normal) and the second-order (curvatures) surface gradients, are introduced to describe local 3D facial shapes. Both early fusion and late fusion of 2D, 3D, as well as 2D and 3D descriptors are comprehensively evaluated on the BU–3DFE database. All the experimental results demonstrate the effectiveness of integrating the 2D and 3D descriptors for expression recognition. Furthermore, we also analyze the generalization capability of the proposed approach on Bosphorus, and illustrate the complementary characteristics between the 2D and 3D descriptors.

Considering the limitation of current approach, in the future, we will go deeply in the following directions: i) The iPar–CLR based joint 2D and 3D facial landmark localization algorithm may fail with large pose variations and data missing. To solve this problem, we will investigate more robust algorithms such as [Zulqarnain Gilani *et al.* 2015]. ii) In current work, we use the simplest early and late fusion schemes. To find more intrinsic complementary characteristics between 2D and 3D modalities, we are going to explore better strategies. iii) In this study, we focus on recognizing six basic expressions using multimodal 2D+3D static images. Following [Sun *et al.* 2008, Reale *et al.* 2013], this work will also be extended to the problem of 3D action unit recognition and to dynamic 3D face spaces.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

This research work mainly addresses the problem of multi-modal face analysis, including facial soft-biometric (Gender and Ethnicity) recognition and facial expression recognition.

The contributions in the thesis are discussed as follows.

5.1.1 Multi-modal facial soft biometric recognition

Considering that soft biometric has ability of providing additional information to make systems of facial recognition more robust, gender and ethnicity classification have attracted many researchers. However, most of the proposed method are based on the single modality, particularly, on 2D texture appearance. Meanwhile, there is a trend to adopt both the 3D shape and 2D texture based modalities, arguing that the joint use of these two clues can generally achieve more accurate and robust accuracy than using only either of the single modality. Our basic assumption for this work is that, the result of single modality (i.e. only 2D texture or 3D shape) based techniques can be ameliorated by combining various clues from different modalities. Thus, We believe that fusion of 2D and 3D data will improve the classification accuracy in the classification of gender and ethnicity. Two approaches have been proposed in this thesis for this purpose.

The first method is our principal contribution. In this method, a novel multi-resolution descriptor called local circular patterns (LCP) is proposed for multi-modal gender and ethnicity classification. The descriptor replaces the binary quantization scheme of LBP with a clustering strategy, which can generate better ap-

proximation with less distortion compared with the binary ones. Meanwhile, we are able to control the quantization accuracy by modifying the number of clusters. Moreover, the boosting algorithm is employed for reducing the dimensionality of the feature and selecting the most discriminative features extracted from various facial regions. The experimental results of gender and ethnicity classification achieved are up to 95.50% and 99.60% respectively on the FRGC v2.0 dataset, and 95.60% and 97.42% respectively on the BU-3DFE dataset, which clearly demonstrate the advantages of the proposed method.

The second work focus on ethnicity classification. We have introduced the Oriented Gradient Maps (OGM) originally proposed for the task of textured 3D face recognition, to extract local discriminative details among different ethnic groups on both 2D and 3D faces. Similarly, Adaboost is applied to select a set of most representative facial features and assign individual weights to them, for emphasizing the importances of different organs that are highly related to the ethnicity property, The proposed approach is evaluated on the FRGC v2.0 database, and the accuracy is up to 98.3% to distinguish Asian and non-Asian persons in the setting that 80% samples are used for training, outperforming most of the ones in the literature. Meanwhile, we also have discussed the impact of the percentage between training and testing samples to final performance for this case.

5.1.2 Multi-modal facial expression recognition

This work is based on the fact that the combination of 2D and 3D data is better than either of the single 2D or 3D modality for expression characterization and AU detection. Thus in this thesis, we have proposed a novel multi-modal system for facial expression recognition, which adapts a novel second-order image gradient based local texture descriptor (HSOG), a novel first-order mesh gradient (i.e., surface normal) based local shape descriptor (meshHOG), as well as a second-order mesh gradient (i.e., surface curvature) based local shape descriptor (meshHOS) to comprehensively encode the expression variations in both the 2D and 3D modalities. Furthermore, a new framework is also proposed to localize facial landmarks on both 2D face images and 3D face scans, using the iPar–CLR algorithm. Finally we

also have discussed the influences of different strategies of fusion in our system. All the experimental results demonstrate the effectiveness of integrating the 2D and 3D descriptors for expression recognition. Furthermore, we also have analyzed the generalization capability of the proposed approach on Bosphorus, and illustrate the complementary characteristics between the 2D and 3D descriptors.

5.2 Perspectives for future work

The extensions of this work that we envisage are presented in the following paragraphs.

5.2.1 Dynamic facial expression recognition

The approaches proposed in this thesis for facial expression recognition is based on static images. Our further work is extending our method or proposing new method for dynamic facial expression recognition, particularly, for the micro-expression recognition. As introduced in the first chapter, micro-expression is a challenging issue due to its short duration. Thus, an effective method for highlighting the deformations of faces caused by facial expressions in a short period is still required.

Meanwhile, for the temporal modeling, the traditional way is to use Hidden Markov models (HMMs). But we believe the recurrent neural networks (RNNs), especially Long short-term memory (LSTM), is another choice for dealing with data of sequences. The effectiveness of LSTM have been demonstrated in many cases, such as speech recognition and action recognition.

5.2.2 Deep learning based facial expression recognition

With the publication of large size databases, i.e, FER2013, just like other researchers, we also pay attention to deep-learning based facial expression recognition. Deep learning have shown its remarkable power in many applications of pattern recognition, e.g, facial verification, action recognition and object detection, but its development on facial expression recognition is relatively slow. Thus, we believe that deep learning based FER is promising research direction in the further.

Furthermore, considering that deep CNNs are effective for feature extraction from still images and RNNs are good at dealing with sequences, a complex framework using both CNNs and RNNs also interests us.

5.2.3 Dense 3D face tracking

For dynamic 3D facial expression recognition, finding correspondences is an inevitable step. Many existed methods are based spatial correspondences, but dense correspondences could certainly provide more information. Traditionally, in the literature, researchers obtain dense correspondences by registering frames of facial sequences to a specific statistical model or a reference face, e.g, Non-rigid ICP, 3DMM, Comformal maps, which definitely drop some temporal information between frames. Therefore, we plan to develop a dense tracking method to solve this issue.

Chapter 6

Publications

The results obtained during my PhD study have been the subject of one publication in an international conference and two in international journals.

International Conferences:

1. Huaxiong Ding, Di Huang, Yunhong Wang, Liming Chen: Facial ethnicity classification based on boosted local texture and shape descriptions. FG 2013: 1-6

International Journals:

1. Huibin Li, Huaxiong Ding, Di Huang, Yunhong Wang, Xi Zhao, Jean-Marie Morvan, Liming Chen:An efficient multimodal 2D + 3D feature-based approach to automatic facial expression recognition. Computer Vision and Image Understanding 140: 83-92 (2015)

2. Di Huang, Huaxiong Ding, Chen Wang, Yunhong Wang, Guangpeng Zhang, Liming Chen: Local circular patterns for multi-modal facial gender and ethnicity classification. Image Vision Comput. 32(12): 1181-1193 (2014)

Bibliography

[Abboud *et al.* 2004] Bouchra Abboud, Franck Davoine and Mo Dang. *Facial expression recognition and synthesis based on an appearance model.* Signal Processing: Image Communication, vol. 19, no. 8, pages 723–740, 2004. 30, 39, 89

[Ahonen *et al.* 2004] Timo Ahonen, Abdenour Hadid and Matti Pietikäinen. *Face recognition with local binary patterns.* In European conference on computer vision, pages 469–481. Springer, 2004. 59

[Alexandre 2010] Luís A Alexandre. *Gender recognition: A multiscale decision fusion approach.* Pattern Recognition Letters, vol. 31, no. 11, pages 1422–1427, 2010. 25, 27

[Amberg *et al.* 2007] Brian Amberg, Sami Romdhani and Thomas Vetter. *Optimal step nonrigid icp algorithms for surface registration.* In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007. 33

[Amberg *et al.* 2008] Brian Amberg, Reinhard Knothe and Thomas Vetter. *Expression invariant 3D face recognition with a morphable model.* In Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, pages 1–6. IEEE, 2008. viii, 33, 34

[Amor *et al.* 2014] Boulbaba Ben Amor, Hassen Drira, Stefano Berretti, Mohamed Daoudi and Anuj Srivastava. *4-D facial expression recognition by learning geometric deformations.* IEEE transactions on cybernetics, vol. 44, no. 12, pages 2443–2457, 2014. 40, 41, 42

[Andreu *et al.* 2014] Yasmina Andreu, Pedro García-Sevilla and Ramón A Mollineda. *Face gender classification: A statistical study when neutral and distorted faces are combined for training and testing purposes.* Image and Vision Computing, vol. 32, no. 1, pages 27–36, 2014. 26

[Asthana *et al.* 2014] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng and Maja Pantic. *Incremental face alignment in the wild.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1859–1866, 2014. 95

[Bekios-Calfa *et al.* 2014] Juan Bekios-Calfa, José M Buenaposada and Luis Baumela. *Robust gender recognition by exploiting facial attributes dependencies.* Pattern Recognition Letters, vol. 36, pages 228–234, 2014. 25

[Berretti *et al.* 2010] Stefano Berretti, Alberto Del Bimbo, Pietro Pala, Boulbaba Ben Amor and Mohamed Daoudi. *A set of selected SIFT features for 3D facial expression recognition.* In Pattern Recognition (ICPR), 2010 20th International Conference on, pages 4125–4128. IEEE, 2010. 37, 90, 92, 107, 108

[Bettadapura 2012] Vinay Bettadapura. *Face expression recognition and analysis: the state of the art.* arXiv preprint arXiv:1203.6722, 2012. 29, 45, 89

[Blanz & Vetter 1999] Volker Blanz and Thomas Vetter. *A morphable model for the synthesis of 3D faces.* In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 32, 33

[Blanz & Vetter 2003] Volker Blanz and Thomas Vetter. *Face recognition based on fitting a 3D morphable model.* IEEE Transactions on pattern analysis and machine intelligence, vol. 25, no. 9, pages 1063–1074, 2003. 32

[Boutellaa *et al.* 2015] Elhocine Boutellaa, Abdenour Hadid, Messaoud Bengherabi and Samy Ait-Aoudia. *On the use of Kinect depth data for identity, gender and ethnicity classification from facial images.* Pattern Recognition Letters, vol. 68, pages 270–277, 2015. 26

[Chan *et al.* 2007] Chi-Ho Chan, Josef Kittler and Kieron Messer. *Multi-scale local binary pattern histograms for face recognition.* In International Conference on Biometrics, pages 809–818. Springer, 2007. 60

[Chang *et al.* 2005] Kyong I Chang, Kevin W Bowyer and Patrick J Flynn. *An evaluation of multimodal 2D+ 3D face biometrics.* IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 4, pages 619–624, 2005. 91

[Chao *et al.* 2015] Wei-Lun Chao, Jian-Jiun Ding and Jun-Zuo Liu. *Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection.* Signal Processing, vol. 117, pages 1–10, 2015. 37, 41

[Cheng *et al.* 2015] Shiyang Cheng, Ioannis Marras, Stefanos Zafeiriou and Maja Pantic. *Active nonrigid ICP algorithm.* In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, volume 1, pages 1–8. IEEE, 2015. viii, 33, 34

[Chu *et al.* 2013] Wen-Sheng Chu, Chun-Rong Huang and Chu-Song Chen. *Gender classification from unaligned facial images using support subspaces.* Information Sciences, vol. 221, pages 98–109, 2013. 14

[Chu *et al.* 2014] Baptiste Chu, Sami Romdhani and Liming Chen. *3D-aided face recognition robust to expression and pose variations.* In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1907–1914. IEEE, 2014. viii, 34

[Cootes *et al.* 1995] Timothy F Cootes, Christopher J Taylor, David H Cooper and Jim Graham. *Active shape models-their training and application.* Computer vision and image understanding, vol. 61, no. 1, pages 38–59, 1995. 29

[Dalal & Triggs 2005] Navneet Dalal and Bill Triggs. *Histograms of oriented gradients for human detection.* In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893. IEEE, 2005. 25

[Danelakis *et al.* 2015] Antonios Danelakis, Theoharis Theoharis and Ioannis Pratikakis. *A survey on facial expression recognition in 3D video sequences.*

Multimedia Tools and Applications, vol. 74, no. 15, pages 5577–5615, 2015. 29, 45

[Deb *et al.* 2002] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal and TAMT Meyarivan. *A fast and elitist multiobjective genetic algorithm: NSGA-II.* IEEE transactions on evolutionary computation, vol. 6, no. 2, pages 182–197, 2002. 41

[Dhall *et al.* 2011] Abhinav Dhall, Roland Goecke, Simon Lucey and Tom Gedeon. *Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark.* In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pages 2106–2112. IEEE, 2011. 42

[Dorai & Jain 1997] Chitra Dorai and Anil K. Jain. *COSMOS-A representation scheme for 3D free-form objects.* IEEE transactions on pattern analysis and machine intelligence, vol. 19, no. 10, pages 1115–1130, 1997. 23

[Edelman *et al.* 1997] Shimon Edelman, Nathan Intrator and Tomaso Poggio. *Complex cells and object recognition.* 1997. 51

[Edwards *et al.* 1998] Gareth J Edwards, Christopher J Taylor and Timothy F Cootes. *Interpreting face images using active appearance models.* In Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pages 300–305. IEEE, 1998. 30

[Ekman & Friesen 1978] P Ekman and W Friesen. *A technique for the measurement of facial movement.* Facial Action Coding System, 1978. 4

[eth 8 31] *2000 Census of Population - Public Law 94-171 Redistricting Data File: Race*, 2009-08-31. 4

[Fang *et al.* 2011] Tianhong Fang, Xi Zhao, Omar Ocegueda, Shishir K Shah and Ioannis A Kakadiaris. *3D facial expression recognition: A perspective on promises and challenges.* In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 603–610. IEEE, 2011. 29, 45, 91

[Farkas *et al.* 2005] Leslie G Farkas, Marko J Katic and Christopher R Forrest. *International anthropometric study of facial morphology in various ethnic groups/races.* Journal of Craniofacial Surgery, vol. 16, no. 4, pages 615–646, 2005. 6

[Fasel & Luettin 2003] Beat Fasel and Juergen Luettin. *Automatic facial expression analysis: a survey.* Pattern recognition, vol. 36, no. 1, pages 259–275, 2003. 29, 45

[Fletcher *et al.* 2004] P Thomas Fletcher, Conglin Lu, Stephen M Pizer and Sarang Joshi. *Principal geodesic analysis for the study of nonlinear statistics of shape.* IEEE transactions on medical imaging, vol. 23, no. 8, pages 995–1005, 2004. 23

[Freeman & Tenenbaum 1997] Bill Freeman and Josh Tenenbaum. *Separating Style and Content.* Advances in neural information processing systems, vol. 9, 1997. 31

[Freund & Schapire 1995] Yoav Freund and Robert E Schapire. *A desicion-theoretic generalization of on-line learning and an application to boosting.* In European conference on computational learning theory, pages 23–37. Springer, 1995. 26, 61

[Friedman *et al.* 2000] Jerome Friedman, Trevor Hastie, Robert Tibshirani*et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors).* The annals of statistics, vol. 28, no. 2, pages 337–407, 2000. 40, 110

[Goldfeather & Interrante 2004] Jack Goldfeather and Victoria Interrante. *A novel cubic-order algorithm for approximating principal direction vectors.* ACM Transactions on Graphics (TOG), vol. 23, no. 1, pages 45–63, 2004. 100

[Golomb *et al.* 1990] B. A. Golomb, D. T. Lawrence and T. J. Sejnowski. *SexNet: A Neural Network Identifies Sex from Human Faces.* In Proceedings of the

1990 Conference on Advances in Neural Information Processing Systems 3, NIPS-3, pages 572–577, 1990. vii, 12

[Gong *et al.* 2009] Boqing Gong, Yueming Wang, Jianzhuang Liu and Xiaoou Tang. *Automatic facial expression recognition on a single 3D face by exploring shape deformation.* In Proceedings of the 17th ACM international conference on Multimedia, pages 569–572. ACM, 2009. 39, 90, 103, 107, 108, 109

[Goodfellow *et al.* 2013] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee*et al. Challenges in representation learning: A report on three machine learning contests.* In International Conference on Neural Information Processing, pages 117–124. Springer, 2013. 42

[Graves *et al.* 2008] Alex Graves, Christoph Mayer, Matthias Wimmer, J Schmidhuber and Bernd Radig. *Facial expression recognition with recurrent neural networks.* In Proceedings of the International Workshop on Cognition for Technical Systems, Munich, germany, 2008. 43

[Gu & Yau 2008] Xianfeng David Gu and Shing-Tung Yau. Computational conformal geometry. International Press Somerville, Mass, USA, 2008. 35

[Gu *et al.* 2004] Xianfeng Gu, Yalin Wang, Tony F Chan, Paul M Thompson and Shing-Tung Yau. *Genus zero surface conformal mapping and its application to brain surface mapping.* IEEE Transactions on Medical Imaging, vol. 23, no. 8, pages 949–958, 2004. 35

[Guo & Mu 2010] Guodong Guo and Guowang Mu. *A study of large-scale ethnicity estimation with gender and age variations.* In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pages 79–86. IEEE, 2010. vii, 15, 16, 27

[Guo & Mu 2014] Guodong Guo and Guowang Mu. *A framework for joint estimation of age, gender and ethnicity on a large database.* Image and Vision Computing, vol. 32, no. 10, pages 761–770, 2014. 17, 27

[Guo *et al.* 2010a] Zhenhua Guo, Lei Zhang and David Zhang. *A completed modeling of local binary pattern operator for texture classification.* IEEE Transactions on Image Processing, vol. 19, no. 6, pages 1657–1663, 2010. 57, 78, 82

[Guo *et al.* 2010b] Zhenhua Guo, Lei Zhang and David Zhang. *Rotation invariant texture classification using LBP variance (LBPV) with global matching.* Pattern recognition, vol. 43, no. 3, pages 706–719, 2010. 57

[Guyon & Elisseeff 2003] Isabelle Guyon and André Elisseeff. *An introduction to variable and feature selection.* Journal of machine learning research, vol. 3, no. Mar, pages 1157–1182, 2003. 52

[Hadid & Pietikäinen 2013] Abdenour Hadid and Matti Pietikäinen. *Demographic classification from face videos using manifold learning.* Neurocomputing, vol. 100, pages 197–205, 2013. 25, 27

[Hadid *et al.* 2015] Abdenour Hadid, Juha Ylioinas, Messaoud Bengherabi, Mohammad Ghahramani and Abdelmalik Taleb-Ahmed. *Gender and texture classification: A comparative analysis using 13 variants of local binary patterns.* Pattern Recognition Letters, vol. 68, pages 231–238, 2015. 19

[Han *et al.* 2009] Xia Han, Hassan Ugail and Ian Palmer. *Gender classification based on 3D face geometry features using SVM.* In CyberWorlds, 2009. CW'09. International Conference on, pages 114–118. IEEE, 2009. 22, 27

[Happy & Routray 2015] SL Happy and Aurobinda Routray. *Automatic facial expression recognition using features of salient facial patches.* IEEE transactions on Affective Computing, vol. 6, no. 1, pages 1–12, 2015. 37

[Hinton & Salakhutdinov 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. *Reducing the dimensionality of data with neural networks.* Science, vol. 313, no. 5786, pages 504–507, 2006. 42

[Hosoi *et al.* 2004] Satoshi Hosoi, Erina Takikawa and Masato Kawade. *Ethnicity estimation with facial images.* In Automatic Face and Gesture Recognition,

2004. Proceedings. Sixth IEEE International Conference on, pages 195–200. IEEE, 2004. 17, 27

[Hu *et al.* 2010] Yuan Hu, Jingqi Yan and Pengfei Shi. *A fusion-based method for 3D facial gender classification.* In Computer and automation engineering (ICCAE), 2010 The 2nd International Conference on, volume 5, pages 369–372. IEEE, 2010. 23, 27

[Huang *et al.* 2006] Yonggang Huang, Yunhong Wang and Tieniu Tan. *Combining Statistics of Geometrical and Correlative Features for 3D Face Recognition.* In BMVC, pages 879–888. Citeseer, 2006. 57

[Huang *et al.* 2011a] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang and Liming Chen. *Local binary patterns and its application to facial image analysis: a survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 6, pages 765–781, 2011. 51

[Huang *et al.* 2011b] Di Huang, Wael Ben Soltana, Mohsen Ardabilian, Yunhong Wang and Liming Chen. *Textured 3d face recognition using biological vision-based facial representation and optimized weighted sum fusion.* In CVPR 2011 WORKSHOPS, pages 1–8. IEEE, 2011. 50

[Huang *et al.* 2012a] Di Huang, Mohsen Ardabilian, Yunhong Wang and Liming Chen. *3-d face recognition using elbp-based facial description and local feature hybrid matching.* IEEE Transactions on Information Forensics and Security, vol. 7, no. 5, pages 1551–1565, 2012. 59

[Huang *et al.* 2012b] Di Huang, Mohsen Ardabilian, Yunhong Wang and Liming Chen. *Oriented gradient maps based automatic asymmetric 3D-2D face recognition.* In 2012 5th IAPR International Conference on Biometrics (ICB), pages 125–131. IEEE, 2012. viii, 51, 54, 55, 56

[Huang *et al.* 2014] Di Huang, Chao Zhu, Yunhong Wang and Liming Chen. *HSOG: a novel local image descriptor based on histograms of the second-*

order gradients. IEEE Transactions on Image Processing, vol. 23, no. 11, pages 4680–4695, 2014. 92, 93, 97, 99

[Huang *et al.* 2015] Xiaohua Huang, Su-Jing Wang, Guoying Zhao and Matti Piteikainen. *Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection.* In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 1–9, 2015. 44

[Huang *et al.* 2016] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng and Matti Pietikäinen. *Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns.* Neurocomputing, vol. 175, pages 564–578, 2016. 44

[Huynh *et al.* 2012] Tri Huynh, Rui Min and Jean-Luc Dugelay. *An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data.* In Asian Conference on Computer Vision, pages 133–145. Springer, 2012. 80

[Ji *et al.* 2008] Zheng Ji, Wen-Yun Yang, Si Wu and Bao-Liang Lu. *Encoding human knowledge for visual pattern recognition.* Knowledge-Based & Intelligent Engineering Systems, submitted for publication, 2008. 19

[Jia & Cristianini 2015] Sen Jia and Nello Cristianini. *Learning to classify gender from four million images.* Pattern Recognition Letters, vol. 58, pages 35–41, 2015. 19, 27

[Joshi *et al.* 2007] SH Joshi, E Klassen, A Srivastava and IH Jermyn. *A novel representation for efficient computation of geodesics between n-dimensional curves.* In IEEE CVPR, pages 1–7, 2007. 38

[Jung *et al.* 2015] Heechul Jung, Sihaeng Lee, Sunjeong Park, Injae Lee, Chunghyun Ahn and Junmo Kim. *Deep temporal appearance-geometry network for facial expression recognition.* arXiv preprint arXiv:1503.01532, 2015. 42

[Kakadiaris *et al.* 2007] Ioannis A Kakadiaris, Georgios Passalis, George Toderici, Mohammed N Murtuza, Yunliang Lu, Nikos Karampatziakis and Theoharis Theoharis. *Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pages 640–649, 2007. 23

[Kashima *et al.* 2008] Hisashi Kashima, Jianying Hu, Bonnie Ray and Moninder Singh. *K-means clustering of proportional data using L1 distance*. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1–4. IEEE, 2008. 59

[Khorrami *et al.* 2016] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli and Thomas S Huang. *How Deep Neural Networks Can Improve Emotion Recognition on Video Data*. arXiv preprint arXiv:1602.07377, 2016. 43

[Kim *et al.* 2006] Hyun-Chul Kim, Daijin Kim, Zoubin Ghahramani and Sung Yang Bang. *Appearance-based gender classification with Gaussian processes*. Pattern Recognition Letters, vol. 27, no. 6, pages 618–626, 2006. 14

[Kim *et al.* 2011] Tae-Hyun Kim, Dong-Chul Park, Dong-Min Woo, Taikyeong Jeong and Soo-Young Min. *Multi-class classifier-based adaboost algorithm*. In International Conference on Intelligent Science and Intelligent Data Engineering, pages 122–127. Springer, 2011. 63

[Kittler *et al.* 1998] Josef Kittler, Mohamad Hatef, Robert P. W. Duin and Jiri Matas. *On Combining Classifiers*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pages 226–239, March 1998. 24

[Kumar 2009] B Vinay Kumar. *Face expression recognition and analysis: the state of the art*. Course Paper, Visual Interfaces to Computer, 2009. 89, 90

[Lanitis *et al.* 1997] Andreas Lanitis, Christopher J. Taylor and Timothy F. Cootes. *Automatic interpretation and coding of face images using flexible models*.

IEEE Transactions on Pattern Analysis and machine intelligence, vol. 19, no. 7, pages 743–756, 1997. 30, 39, 89

[Le *et al.* 2011] Vuong Le, Hao Tang and Thomas S Huang. *Expression recognition from 3D dynamic faces using robust spatio-temporal shape features.* In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 414–421. IEEE, 2011. 40, 41, 42

[Lei *et al.* 2014] Zhen Lei, Matti Pietikäinen and Stan Z Li. *Learning discriminant face descriptor.* IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 2, pages 289–302, 2014. 51

[Lemaire *et al.* 2011] Pierre Lemaire, Boulbaba Ben Amor, Mohsen Ardabilian, Liming Chen and Mohamed Daoudi. *Fully automatic 3D facial expression recognition using a region-based approach.* In Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, pages 53–58. ACM, 2011. 38, 107, 108

[Li *et al.* 2010] Xiaoli Li, Qiuqi Ruan and Yue Ming. *3D facial expression recognition based on basic geometric features.* In IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS, pages 1366–1369. IEEE, 2010. 37, 41

[Li *et al.* 2011a] Huibin Li, Di Huang, Pierre Lemaire, Jean-Marie Morvan and Liming Chen. *Expression robust 3D face recognition via mesh-based histograms of multiple order surface differential quantities.* In 2011 18th IEEE International Conference on Image Processing, pages 3053–3056. IEEE, 2011. 93, 99, 102

[Li *et al.* 2011b] Huibin Li, Jean-Marie Morvan and Liming Chen. *3d facial expression recognition based on histograms of surface differential quantities.* In International Conference on Advanced Concepts for Intelligent Vision Systems, pages 483–494. Springer, 2011. 90, 92

[Li *et al.* 2012a] Bing Li, Xiao-Chen Lian and Bao-Liang Lu. *Gender classification by combining clothing, hair and facial component classifiers.* Neurocomputing, vol. 76, no. 1, pages 18–27, 2012. 18

[Li *et al.* 2012b] Huibin Li, Liming Chen, Di Huang, Yunhong Wang and Jean-Marie Morvan. *3D facial expression recognition via multiple kernel learning of multi-scale local normal patterns.* In Pattern Recognition (ICPR), 2012 21st International Conference on, pages 2577–2580. IEEE, 2012. 38, 41, 90, 107, 108

[Li *et al.* 2013] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao and Matti Pietikäinen. *A spontaneous micro-expression database: Inducement, collection and baseline.* In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pages 1–6. IEEE, 2013. 43

[Li *et al.* 2015] Huibin Li, Di Huang, Jean-Marie Morvan, Yunhong Wang and Liming Chen. *Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors.* International Journal of Computer Vision, vol. 113, no. 2, pages 128–142, 2015. 93, 99

[Lian *et al.* 2005] Hui-Cheng Lian, Bao-Liang Lu, Erina Takikawa and Satoshi Hosoi. *Gender recognition using a min-max modular support vector machine.* In International Conference on Natural Computation, pages 438–441. Springer, 2005. 17, 27

[Liao *et al.* 2009] Shu Liao, Max WK Law and Albert CS Chung. *Dominant local binary patterns for texture classification.* IEEE transactions on image processing, vol. 18, no. 5, pages 1107–1118, 2009. 57

[Liu & Fieguth 2012] Li Liu and Paul Fieguth. *Texture classification from random features.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pages 574–586, 2012. 82, 85

[Liu *et al.* 2015a] Mengyi Liu, Shaoxin Li, Shiguang Shan and Xilin Chen. *Au-inspired deep networks for facial expression feature learning.* Neurocomputing, vol. 159, pages 126–136, 2015. 42

[Liu *et al.* 2015b] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao and Xiaolan Fu. *A main directional mean optical flow feature for spontaneous micro-expression recognition.* IEEE Transactions on Affective Computing, 2015. 44

[Lowe 2004] David G Lowe. *Distinctive image features from scale-invariant keypoints.* International journal of computer vision, vol. 60, no. 2, pages 91–110, 2004. 97

[Lu & Jain 2004] Xiaoguang Lu and Anil K Jain. *Ethnicity identification from face images.* In Defense and Security, pages 114–123. International Society for Optics and Photonics, 2004. vii, 13, 27

[Lu & Lin 2007] Huchuan Lu and Hui Lin. *Gender recognition using adaboosted feature.* In Third International Conference on Natural Computation (ICNC 2007), volume 2, pages 646–650. IEEE, 2007. 17, 27

[Lu *et al.* 2006] Xiaoguang Lu, Hong Chen and Anil K Jain. *Multimodal facial gender and ethnicity identification.* In International Conference on Biometrics, pages 554–561. Springer, 2006. vii, 24, 25, 27, 50, 51, 64, 66, 80, 82, 84

[Luo & Lu 2006] Jun Luo and Bao-Liang Lu. *Gender recognition using a min-max modular support vector machine with equal clustering.* In International Symposium on Neural Networks, pages 210–215. Springer, 2006. 17

[Lyle *et al.* 2010] Jamie R Lyle, Philip E Miller, Shrinivas J Pundlik and Damon L Woodard. *Soft biometric classification using periocular region features.* In Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, pages 1–7. IEEE, 2010. 18, 27, 66

[Lyons *et al.* 1998] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi and Jiro Gyoba. *Coding facial expressions with gabor wavelets.* In Automatic Face

and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pages 200–205. IEEE, 1998. 36, 90

[Maalej *et al.* 2010] Ahmed Maalej, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava and Stefano Berretti. *Local 3D shape analysis for facial expression recognition.* In Pattern Recognition (ICPR), 2010 20th International Conference on, pages 4129–4132. IEEE, 2010. 39, 40, 41, 90

[Mäkinen & Raisamo 2008] Erno Mäkinen and Roope Raisamo. *An experimental comparison of gender classification methods.* Pattern Recognition Letters, vol. 29, no. 10, pages 1544–1556, 2008. 26

[Mehrabian 1972] Albert Mehrabian. Nonverbal communication. Transaction Publishers, 1972. 4

[Meilă 2006] Marina Meilă. *The uniqueness of a good optimum for k-means.* In Proceedings of the 23rd international conference on Machine learning, pages 625–632. ACM, 2006. 58

[Mery & Bowyer 2014] Domingo Mery and Kevin Bowyer. *Recognition of facial attributes using adaptive sparse representations of random patches.* In European Conference on Computer Vision, pages 778–792. Springer, 2014. vii, 20, 22

[Mery & Bowyer 2015] Domingo Mery and Kevin Bowyer. *Automatic facial attribute analysis via adaptive sparse representation of random patches.* Pattern Recognition Letters, vol. 68, pages 260–269, 2015. 22, 27

[Mian *et al.* 2007] Ajmal Mian, Mohammed Bennamoun and Robyn Owens. *An efficient multimodal 2D-3D hybrid approach to automatic face recognition.* IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 11, pages 1927–1943, 2007. 91

[Michaeli 2013] Dov Michaeli. The doctor weighs in (on everything). CreateSpace Independent Publishing Platform, 2013. 2

[Moghaddam & Yang 2000] Baback Moghaddam and Ming-Hsuan Yang. *Gender classification with support vector machines.* In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pages 306–311. IEEE, 2000. 13, 27

[Mpiperis *et al.* 2008] Iordanis Mpiperis, Sotiris Malassiotis and Michael G Strintzis. *Bilinear models for 3-D face and facial expression recognition.* IEEE Transactions on Information Forensics and Security, vol. 3, no. 3, pages 498–511, 2008. vii, 30, 32, 39, 90, 107, 108

[Ocegueda *et al.* 2011a] Omar Ocegueda, Tianhong Fang, Shishir K Shah and Ioannis A Kakadiaris. *Expressive maps for 3D facial expression recognition.* In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pages 1270–1275. IEEE, 2011. 38, 41, 90

[Ocegueda *et al.* 2011b] Omar Ocegueda, Shishir K Shah and Ioannis A Kakadiaris. *Which parts of the face give out your identity?* In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 641–648. IEEE, 2011. 38

[Ojala *et al.* 1994] Timo Ojala, Matti Pietikainen and David Harwood. *Performance evaluation of texture measures with classification based on Kullback discrimination of distributions.* In Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on, volume 1, pages 582–585. IEEE, 1994. 17

[Ojala *et al.* 2002] Timo Ojala, Matti Pietikainen and Topi Maenpaa. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.* IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pages 971–987, 2002. 56, 57, 60

[O'toole *et al.* 1997] Alice J O'toole, Thomas Vetter, Nikolaus F Troje and Heinrich H Bülthoff. *Sex classification is better with three-dimensional head*

structure than with image intensity information. Perception, vol. 26, no. 1, pages 75–84, 1997. 22

[Pantic & Rothkrantz 2000] Maja Pantic and Leon J. M. Rothkrantz. *Automatic analysis of facial expressions: The state of the art.* IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 12, pages 1424–1445, 2000. 89, 90

[Passalis *et al.* 2005] G Passalis, IA Kakadiaris, T Theoharis, G Toderici and N Murtuza. *Evaluation of 3D face recognition in the presence of facial expressions: an annotated deformable model approach.* In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, pages 171–171. IEEE, 2005. 33

[Patel *et al.* 2015] Bhavik Patel, RP Maheshwari and R Balasubramanian. *Multi-quantized local binary patterns for facial gender classification.* Computers & Electrical Engineering, 2015. 19

[Pfister *et al.* 2011] Tomas Pfister, Xiaobai Li, Guoying Zhao and Matti Pietikäinen. *Recognising spontaneous facial micro-expressions.* In 2011 International Conference on Computer Vision, pages 1449–1456. IEEE, 2011. 44

[Phillips *et al.* 2005] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min and William Worek. *Overview of the face recognition grand challenge.* In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 947–954. IEEE, 2005. 64

[Platt *et al.* 1999] John Platt*et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Advances in large margin classifiers, vol. 10, no. 3, pages 61–74, 1999. 24

[Poggio & Girosi 1990] Tomaso Poggio and Federico Girosi. *Networks for approximation and learning.* Proceedings of the IEEE, vol. 78, no. 9, pages 1481–1497, 1990. 13

[Poggio *et al.*] Brunelli Poggio, R. Brunelli and T. Poggio. *HyberBF Networks for Gender Classification.* 13

[Prasad & Domke 2005] V Shiv Naga Prasad and Justin Domke. *Gabor filter visualization.* J. Atmos. Sci, vol. 13, 2005. 14

[Rai & Khanna 2014] Preeti Rai and Pritee Khanna. *A gender classification system robust to occlusion using Gabor features based (2D) 2 PCA.* Journal of Visual Communication and Image Representation, vol. 25, no. 5, pages 1118–1129, 2014. 17, 27

[Ramanathan *et al.* 2006] Subramanian Ramanathan, Ashraf Kassim, YV Venkatesh and Wu Sin Wah. *Human facial expression recognition using a 3D morphable model.* In 2006 International Conference on Image Processing, pages 661–664. IEEE, 2006. 32, 90

[Reale *et al.* 2013] Michael Reale, Xing Zhang and Lijun Yin. *Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis.* In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pages 1–8. IEEE, 2013. 112

[Reid & Nixon 2011] Daniel A Reid and Mark S Nixon. *Using comparative human descriptions for soft biometrics.* In Biometrics (IJCB), 2011 International Joint Conference on, pages 1–6. IEEE, 2011. 1

[Riesenhuber & Poggio 1999] Maximilian Riesenhuber and Tomaso Poggio. *Hierarchical models of object recognition in cortex.* Nature neuroscience, vol. 2, no. 11, pages 1019–1025, 1999. 15

[Rosato *et al.* 2008] Matthew Rosato, Xiaochen Chen and Lijun Yin. *Automatic registration of vertex correspondences for 3D facial expression analysis.* In Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on, pages 1–7. IEEE, 2008. 36, 38, 41

[Rumelhart *et al.* 1988] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. *Learning representations by back-propagating errors.* Cognitive modeling, vol. 5, no. 3, page 1, 1988. 42

[Sandbach *et al.* 2011] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic and Daniel Rueckert. *A dynamic approach to the recognition of 3d facial expressions and their temporal models.* In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 406–413. IEEE, 2011. 40

[Sandbach *et al.* 2012a] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic and Daniel Rueckert. *Recognition of 3D facial expression dynamics.* Image and Vision Computing, vol. 30, no. 10, pages 762–773, 2012. 40, 42

[Sandbach *et al.* 2012b] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic and Lijun Yin. *Static and dynamic 3D facial expression recognition: A comprehensive survey.* Image and Vision Computing, vol. 30, no. 10, pages 683–697, 2012. v, 29, 34, 39, 45, 91

[Sariyanidi *et al.* 2015] Evangelos Sariyanidi, Hatice Gunes and Andrea Cavallaro. *Automatic analysis of facial affect: A survey of registration, representation, and recognition.* IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 6, pages 1113–1133, 2015. 36

[Savran *et al.* 2010] Arman Savran, Bülent Sankur and M Taha Bilge. *Facial action unit detection: 3D versus 2D modality.* In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pages 71–78. IEEE, 2010. 91

[Scherer 2005] Klaus R Scherer. *What are emotions? And how can they be measured?* Social science information, vol. 44, no. 4, pages 695–729, 2005.

[Sha *et al.* 2011] Teng Sha, Mingli Song, Jiajun Bu, Chun Chen and Dacheng Tao. *Feature level analysis for 3D facial expression recognition.* Neurocomputing, vol. 74, no. 12, pages 2135–2141, 2011. 38, 41

[Shalev-Shwartz *et al.* 2011] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro and Andrew Cotter. *Pegasos: Primal estimated sub-gradient solver for svm.* Mathematical programming, vol. 127, no. 1, pages 3–30, 2011. 19

[Shan & Gritti 2008] Caifeng Shan and Tommaso Gritti. *Learning Discriminative LBP-Histogram Bins for Facial Expression Recognition.* In BMVC, pages 1–10, 2008. 60, 61

[Shan *et al.* 2006] Caifeng Shan, Shaogang Gong and Peter W McOwan. *A comprehensive empirical study on linear subspace methods for facial expression analysis.* In 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), pages 153–153. IEEE, 2006. 37

[Shan *et al.* 2009] Caifeng Shan, Shaogang Gong and Peter W McOwan. *Facial expression recognition based on local binary patterns: A comprehensive study.* Image and Vision Computing, vol. 27, no. 6, pages 803–816, 2009. 37, 41, 90

[Shan 2012] Caifeng Shan. *Learning local binary patterns for gender classification on real-world face images.* Pattern Recognition Letters, vol. 33, no. 4, pages 431–437, 2012. 18, 27

[Shbib & Zhou 2015] Reda Shbib and Shikun Zhou. *Facial expression analysis using active shape model.* Int. J. Signal Process. Image Process. Pattern Recognit, vol. 8, no. 1, pages 9–22, 2015. 30

[Shih 2013] Huang-Chia Shih. *Robust gender classification using a precise patch histogram.* Pattern Recognition, vol. 46, no. 2, pages 519–528, 2013. 18, 27

[Smith & Hancock 2008] William AP Smith and Edwin R Hancock. *Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics.* International Journal of Computer Vision, vol. 76, no. 1, pages 71–91, 2008. 23

[Song *et al.* 2014] Mingli Song, Dacheng Tao, Shengpeng Sun, Chun Chen and Stephen J Maybank. *Robust 3d face landmark localization based on local*

coordinate coding. IEEE Transactions on Image Processing, vol. 23, no. 12, pages 5108–5122, 2014. 91

[Soyel & Demirel 2007] Hamit Soyel and Hasan Demirel. *Facial expression recognition using 3D facial feature distances*. In International Conference Image Analysis and Recognition, pages 831–838. Springer, 2007. 37, 90, 107, 108

[Soyel & Demirel 2008] Hamit Soyel and Hasan Demirel. *3D facial expression recognition with geometrically localized facial features*. In Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on, pages 1–4. IEEE, 2008. 37, 41, 90, 107, 108

[Soyel & Demirel 2010] Hamit Soyel and Hasan Demirel. *Optimal feature selection for 3D facial expression recognition using coarse-to-fine classification*. Turkish Journal of Electrical Engineering & Computer Sciences, vol. 18, no. 6, pages 1031–1040, 2010. 41

[Srivastava & Roy 2009] Ruchir Srivastava and Sujoy Roy. *3D facial expression recognition using residues*. In TENCON 2009-2009 IEEE Region 10 Conference, pages 1–5. IEEE, 2009. 38

[Sun & Yin 2008] Yi Sun and Lijun Yin. *Facial expression recognition based on 3D dynamic range model sequences*. In European Conference on Computer Vision, pages 58–71. Springer, 2008. 42

[Sun *et al.* 2008] Yi Sun, Michael Reale and Lijun Yin. *Recognizing partial facial action units based on 3D dynamic range data for facial expression recognition*. In Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, pages 1–8. IEEE, 2008. 112

[Tan & Triggs 2007] Xiaoyang Tan and Bill Triggs. *Enhanced local texture feature sets for face recognition under difficult lighting conditions*. In International Workshop on Analysis and Modeling of Faces and Gestures, pages 168–182. Springer, 2007. 57

[Tan & Triggs 2010] Xiaoyang Tan and Bill Triggs. *Enhanced local texture feature sets for face recognition under difficult lighting conditions.* IEEE transactions on image processing, vol. 19, no. 6, pages 1635–1650, 2010. 82

[Tang & Huang 2008a] Hao Tang and Thomas S Huang. *3D facial expression recognition based on automatically selected features.* In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pages 1–8. IEEE, 2008. 37, 90, 107, 108

[Tang & Huang 2008b] Hao Tang and Thomas S Huang. *3D facial expression recognition based on properties of line segments connecting facial feature points.* In Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, pages 1–6. IEEE, 2008. 37, 41, 90, 107, 108

[Tekguc *et al.* 2009] Umut Tekguc, Hamit Soyel and Hasan Demirel. *Feature selection for person-independent 3D facial expression recognition using NSGA-II.* In Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on, pages 35–38. IEEE, 2009. 37, 41

[Tenenbaum & Freeman 2000] Joshua B Tenenbaum and William T Freeman. *Separating style and content with bilinear models.* Neural computation, vol. 12, no. 6, pages 1247–1283, 2000. 31

[Toderici *et al.* 2010] George Toderici, Sean M O'malley, George Passalis, Theoharis Theoharis and Ioannis A Kakadiaris. *Ethnicity-and gender-based subject retrieval using 3-D face-recognition techniques.* International Journal of Computer Vision, vol. 89, no. 2-3, pages 382–391, 2010. 23, 27, 64, 80, 81

[Trucco & Fisher 1995] Emanuele Trucco and Robert B Fisher. *Experiments in curvature-based segmentation of range data.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 2, pages 177–182, 1995. 38

[Tsalakanidou & Malassiotis 2009] Filareti Tsalakanidou and Sotiris Malassiotis. *Robust facial action recognition from real-time 3D streams.* In Computer Vision and Pattern Recognition Workshops, pages 4–11, 2009. 30

[Tsalakanidou & Malassiotis 2010] Filareti Tsalakanidou and Sotiris Malassiotis. *Real-time 2D+ 3D facial action and expression recognition.* Pattern Recognition, vol. 43, no. 5, pages 1763–1775, 2010. 37, 91

[Varma & Zisserman 2009] Manik Varma and Andrew Zisserman. *A statistical approach to material classification using image patch exemplars.* IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 11, pages 2032–2047, 2009. 82, 85

[Viola & Jones 2001] Paul Viola and Michael Jones. *Rapid object detection using a boosted cascade of simple features.* In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, pages I–511. IEEE, 2001. 61

[Viola & Jones 2004] Paul Viola and Michael J Jones. *Robust real-time face detection.* International journal of computer vision, vol. 57, no. 2, pages 137–154, 2004. 19

[Vu & Caplier 2012] Ngoc-Son Vu and Alice Caplier. *Enhanced patterns of oriented edge magnitudes for face recognition and image matching.* IEEE Transactions on Image Processing, vol. 21, no. 3, pages 1352–1365, 2012. 51

[Wan & Aggarwal 2014] Shaohua Wan and JK Aggarwal. *Spontaneous facial expression recognition: A robust metric learning approach.* Pattern Recognition, vol. 47, no. 5, pages 1859–1868, 2014. 37, 41

[Wang & Yin 2007] Jun Wang and Lijun Yin. *Static topographic modeling for facial expression recognition and analysis.* Computer Vision and Image Understanding, vol. 108, no. 1, pages 19–34, 2007. 30, 39, 90

[Wang *et al.* 2006] Jun Wang, Lijun Yin, Xiaozhou Wei and Yi Sun. *3D facial expression recognition based on primitive surface feature distribution.* In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1399–1406. IEEE, 2006. 38, 41, 64, 90, 91, 92, 107, 108

[Wang *et al.* 2007] Peng Wang, Christian Kohler, Fred Barrett, Raquel Gur, Ruben Gur and Ragini Verma. *Quantifying facial expression abnormality in schizophrenia by combining 2D and 3D features.* In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007. 37, 38, 92

[Wang *et al.* 2011] Su-Jing Wang, Chun-Guang Zhou, Na Zhang, Xu-Jun Peng, Yu-Hsin Chen and Xiaohua Liu. *Face recognition using second-order discriminant tensor subspace analysis.* Neurocomputing, vol. 74, no. 12, pages 2142–2156, 2011. 44

[Wang *et al.* 2014a] Su-Jing Wang, Hui-Ling Chen, Wen-Jing Yan, Yu-Hsin Chen and Xiaolan Fu. *Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine.* Neural processing letters, vol. 39, no. 1, pages 25–43, 2014. 43

[Wang *et al.* 2014b] Yandan Wang, John See, Raphael C-W Phan and Yee-Hui Oh. *Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition.* In Asian Conference on Computer Vision, pages 525–537. Springer, 2014. 44

[Wang *et al.* 2015] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang and Jianhua Tao. *Micro-expression recognition using color spaces.* IEEE Transactions on Image Processing, vol. 24, no. 12, pages 6034–6047, 2015. 44

[Wolberg 1998] George Wolberg. *Image morphing: a survey.* The visual computer, vol. 14, no. 8, pages 360–372, 1998. 35

[Wolf & Shashua 2003] Lior Wolf and Amnon Shashua. *Learning over sets using kernel principal angles.* Journal of Machine Learning Research, vol. 4, no. Oct, pages 913–931, 2003. 14

[Wright *et al.* 2009] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry and Yi Ma. *Robust face recognition via sparse representation.* IEEE trans-

actions on pattern analysis and machine intelligence, vol. 31, no. 2, pages 210–227, 2009. 20

[Wu *et al.* 2003] Bo Wu, Haizhou Ai and Chang Huang. *LUT-based Adaboost for gender classification.* In International Conference on Audio-and Video-Based Biometric Person Authentication, pages 104–110. Springer, 2003. 26

[Wu *et al.* 2009] Jing Wu, William AP Smith, Edwin R Hancock and Michal Kawulok. *Extracting gender discriminating features from facial needle-maps.* In 2009 16th IEEE International Conference on Image Processing (ICIP), pages 2449–2452. IEEE, 2009. 23

[Wu *et al.* 2010] Jing Wu, William AP Smith and Edwin R Hancock. *Facial gender classification using shape-from-shading.* Image and Vision Computing, vol. 28, no. 6, pages 1039–1048, 2010. 23, 80

[Wu *et al.* 2011] Jing Wu, William AP Smith and Edwin R Hancock. *Gender discriminating models from facial surface normals.* Pattern Recognition, vol. 44, no. 12, pages 2871–2886, 2011. 23, 27

[Xiong & De la Torre 2013] Xuehan Xiong and Fernando De la Torre. *Supervised descent method and its applications to face alignment.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 532–539, 2013. 95

[Yan *et al.* 2007] Shuicheng Yan, Huan Wang, Xiaoou Tang and Thomas Huang. *Exploring feature descritors for face recognition.* In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, volume 1, pages I–629. IEEE, 2007. 60

[Yan *et al.* 2013] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang and Xiaolan Fu. *CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces.* In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pages 1–7. IEEE, 2013. 43

[Yan *et al.* 2014a] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen and Xiaolan Fu. *CASME II: An improved spontaneous micro-expression database and the baseline evaluation.* PloS one, vol. 9, no. 1, page e86041, 2014. 43

[Yan *et al.* 2014b] Wen-Jing Yan, Su-Jing Wang, Yong-Jin Liu, Qi Wu and Xiaolan Fu. *For micro-expression recognition: Database and suggestions.* Neurocomputing, vol. 136, pages 82–87, 2014. 44

[Yang & Ai 2007] Zhiguang Yang and Haizhou Ai. *Demographic classification with local binary patterns.* In International Conference on Biometrics, pages 464–473. Springer, 2007. 18, 27, 66

[Yang & Wang 2007] Hong Yang and Yiding Wang. *A LBP-based face recognition method with Hamming distance constraint.* In Image and Graphics, 2007. ICIG 2007. Fourth International Conference on, pages 645–649. IEEE, 2007. 57

[Yang *et al.* 2015a] Xudong Yang, Di Huang, Yunhong Wang and Liming Chen. *Automatic 3D facial expression recognition using geometric scattering representation.* In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. IEEE, 2015. 38

[Yang *et al.* 2015b] Xudong Yang, Di Huang, Yunhong Wang and Liming Chen. *Automatic 3d facial expression recognition using geometric scattering representation.* In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, volume 1, pages 1–6. IEEE, 2015. 107, 109

[Yin *et al.* 2006a] Lijun Yin, Xiaozhou Wei, Peter Longo and Abhinesh Bhuvanesh. *Analyzing facial expressions using intensity-variant 3D data for human computer interaction.* In 18th International Conference on Pattern Recognition (ICPR'06), volume 1, pages 1248–1251. IEEE, 2006. 39, 41

[Yin *et al.* 2006b] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang and Matthew J Rosato. *A 3D facial expression database for facial behavior research.* In 7th international conference on automatic face and gesture recognition (FGR06), pages 211–216. IEEE, 2006. 31, 102

[Yu & Zhang 2015] Zhiding Yu and Cha Zhang. *Image based static facial expression recognition with multiple deep network learning.* In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 435–442. ACM, 2015. 43

[Yueh *et al.* 2015] Mei-Heng Yueh, Xianfeng David Gu, Wen-Wei Lin, Chin-Tien Wu and Shing-Tung Yau. *Conformal Surface Morphing with Applications on Facial Expressions.* arXiv preprint arXiv:1504.00097, 2015. viii, 35, 39, 47

[Zalewski & Gong 2004] Lukasz Zalewski and Shaogang Gong. *Synthesis and recognition of facial expressions in virtual 3d views.* In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pages 493–498. IEEE, 2004. 30, 89

[Zeng *et al.* 2009] Zhihong Zeng, Maja Pantic, Glenn I Roisman and Thomas S Huang. *A survey of affect recognition methods: Audio, visual, and spontaneous expressions.* IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pages 39–58, 2009. 29, 45

[Zeng *et al.* 2013] Wei Zeng, Huibin Li, Liming Chen, Jean-Marie Morvan and Xianfeng David Gu. *An automatic 3D expression recognition framework based on sparse representation of conformal images.* In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pages 1–8. IEEE, 2013. 36, 38, 90, 107

[Zhang & Wang 2009] Guangpeng Zhang and Yunhong Wang. *Multimodal 2D and 3D facial ethnicity classification.* In Image and Graphics, 2009. ICIG'09. Fifth International Conference on, pages 928–932. IEEE, 2009. 25, 27, 66

[Zhang 1994] Zhengyou Zhang. *Iterative point matching for registration of free-form curves and surfaces.* International journal of computer vision, vol. 13, no. 2, pages 119–152, 1994. 71

[Zhao & Pietikainen 2007] Guoying Zhao and Matti Pietikainen. *Dynamic texture recognition using local binary patterns with an application to facial expressions.* IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pages 915–928, 2007. 25

[Zhao & Pietikainen 2008] Guoying Zhao and Matti Pietikainen. *Principal appearance and motion from boosted spatiotemporal descriptors.* In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pages 1–8. IEEE, 2008. 61

[Zhao *et al.* 2010] Xi Zhao, Di Huang, Emmanuel Dellandrea and Liming Chen. *Automatic 3D facial expression recognition based on a Bayesian Belief Net and a Statistical Facial Feature Model.* In Pattern Recognition (ICPR), 2010 20th International Conference on, pages 3724–3727. IEEE, 2010. 37, 41, 90, 91, 107, 108

[Zhen *et al.* 2015] Qingkai Zhen, Di Huang, Yunhong Wang and Liming Chen. *Muscular movement model based automatic 3d facial expression recognition.* In International Conference on Multimedia Modeling, pages 522–533. Springer, 2015. 38, 90, 107, 108, 109

[Zheng & Lu 2011] Ji Zheng and Bao-Liang Lu. *A support vector machine classifier with automatic confidence and its application to gender classification.* Neurocomputing, vol. 74, no. 11, pages 1926–1935, 2011. 19

[Zhu *et al.* 2009] Ji Zhu, Hui Zou, Saharon Rosset and Trevor Hastie. *Multi-class adaboost.* Statistics and its Interface, vol. 2, no. 3, pages 349–360, 2009. 63

[Zulqarnain Gilani *et al.* 2015] Syed Zulqarnain Gilani, Faisal Shafait and Ajmal Mian. *Shape-based automatic detection of a large number of 3d facial land-*

marks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4639–4648, 2015. 112