3D face analysis : landmarking, expression recognition and beyond

Xi Zhao

THESE

pour obtenir le grade de
DOCTEUR DE L'ECOLE CENTRALE DE LYON
Spécialité : Informatique

présentée et soutenue publiquement par

XI ZHAO
le 13 septembre 2010

3D Face Analysis:
Landmarking, Expression Recognition and beyond

Ecole Doctorale InfoMaths

Directeur de thèse : Liming CHEN
Co-directeur de thèse : Emmanuel DELLANDRÉA

JURY

Prof. Bulent Sankur	Université Bogazici	Rapporteur
Prof. Maurice Milgram	Université UMPC	Rapporteur
Prof. Alice Caplier	Université INP	Examinateur
Prof. Dimitris Samaras	Université Stony Brook	Examinateur
Prof. Mohamed Daoudi	Université Telecom Lille	Examinateur
Prof. Liming Chen	Ecole Centrale de Lyon	Directeur de thèse
Dr. Emmanuel Dellandréa	Ecole Centrale de Lyon	Co-directeur de thèse

Numéro d'ordre : 2010-21

Acknowledgment

I wish to express my deep and sincere gratitude to my supervisor, Prof. Liming Chen. His wide knowledge and his serious attitude towards research have been of great value both for my Ph.D study and for my future academic career. At the same time, his understanding, encouragement and care also give me emotional support throughout my three-year Ph.D life.

I am deeply grateful to my supervisor, Dr. Emmanuel Dellandréa, for his constructive and detailed supervision during my PhD study, and for his important help throughout this thesis. His logical way of thinking and carefulness on the research have affected me to a large extent.

I wish to express my warm and sincere appreciation to Prof. Bulent Sankur, University of Bogzici, and Prof. Maurice Milgram, University of UMPC, for their detailed, valuable and constructive comments, which help to improve the quality of this work greatly.

I warmly thank Mohsen Ardabilian, Christian Vial, Colette Vial, and Isabelle Dominique for their support in all aspects of my lab life.

I owe my gratitude to Kun Peng, Xiao Zhongzhe, Aliaksandr Paradzinets, Alain Pujol, Yan Liu, Huanzhang Fu, Chu Duc Nguyen, Karima Ouji, Przemyslaw Szeptycki, Kiryl Bletsko, Gang Niu, Xiaopin Zhong, Jing Zhang, Ying Hu, Di Huang, Chao Zhu, Huibin Li, Yu Zhang, Boyang Gao, Ningning Liu and Tao Xu. The valuable discussions and communications with them not only help me to solve difficulties both in academic and personal aspects, but also make my life so pleasant and happy in these three years.

I owe my loving thankfulness to my parents Jinsheng Zhao and Yaxian Dang, and my wife Zhenmei Zhu. Without their encouragement and understanding it would have been impossible for me to finish my PhD study.

I give my sincere appreciation to the China Scholarship Council for the financial support.

Ecully, France, Sep. 2010
Xi ZHAO

Contents

List of Tables

List of Figures

Resumé

Cette thèse de doctorat est dédiée à l'analyse automatique de visages 3D, incluant la détection de points d'intérêt et la reconnaissance de l'expression faciale. En effet, l'expression faciale joue un rôle important dans la communication verbale et non verbale, ainsi que pour exprimer des émotions. Ainsi, la reconnaissance automatique de l'expression faciale offre de nombreuses opportunités et applications, et est en particulier au coeur d'interfaces homme-machine "intelligentes" centrées sur l'être humain. Par ailleurs, la détection automatique de points d'intérêt du visages (coins de la bouche et des yeux, ...) permet la localisation d'éléments du visage qui est essentielle pour de nombreuses méthodes d'analyse faciale telle que la segmentation du visage et l'extraction de descripteurs utilisée par exemple pour la reconnaissance de l'expression. L'objectif de cette thèse est donc d'élaborer des approches de détection de points d'intérêt sur les visages 3D et de reconnaissance de l'expression faciale pour finalement proposer une solution entièrement automatique de reconnaissance de l'activité faciale incluant l'expression et les unités d'action (ou *Action Units*).

Dans ce travail, nous avons proposé un réseau de croyance bayésien (Bayesian Belief Network ou BBN) pour la reconnaissance d'expressions faciales ainsi que d'unités d'action. Un modèle statistique de caractéristiques faciales (Statistical Facial feAture Model ou SFAM) a également été élaboré pour permettre la localisation des points d'intérêt sur laquelle s'appuie notre BBN afin de permettre la mise en place d'un système entièrement automatique de reconnaissance de l'expression faciale. Nos principales contributions sont les suivantes. Tout d'abord, nous avons proposé un modèle de visage partiel déformable, nommé SFAM, basé sur le principe de l'analyse en composantes principales. Ce modèle permet d'apprendre à la fois les variations globales de la position relative des points d'intérêt du visage (configuration du visage) et les variations locales en terme de texture et de forme autour de

chaque point d'intérêt. Différentes instances de visages partiels peuvent ainsi être produites en faisant varier les valeurs des paramètres du modèle. Deuxièment, nous avons dévoloppé un algorithme de localisation des points d'intérêt du visage basé sur la minimisation d'une fontion objectif décrivant la corrélation entre les instances du modèle SFAM et les visages requête. Troisièmement, nous avons élaboré un réseau de croyance bayésien (BBN) dont la structure décrit les relations de dépendance entre les sujets, les expressions et les descripteurs faciaux. Les expressions faciales et les unités d'action sont alors modélisées comme les états du noeud correspondant à la variable expression et sont reconnues en identifiant le maximum de croyance pour tous les états. Nous avons également proposé une nouvelle approche pour l'inférence des paramètres du BBN utilisant un modèle de caractéristiques faciales pouvant être considéré comme une extension de SFAM. Finalement, afin d'enrichir l'information utilisée pour l'analyse de visages 3D, et particulièrement pour la reconnaissance de l'expression faciale, nous avons également élaboré un descripteur de visages 3D, nommé SGAND, pour caractériser les propriétés géométriques d'un point par rapport à son voisinage dans le nuage de points représentant un visage 3D.

L'efficacité de ces méthodes a été évaluée sur les bases FRGC, BU3DFE et Bosphorus pour la localisation des points d'intérêt ainsi que sur les bases BU3DFE et Bosphorus pour la reconnaissance des expressions faciales et des unités d'action.

Mots-clés : Visage 3D, reconnaissance de l'expression faciale, reconnaissance des unités d'action, localisation de points d'intérêt, modèle statistique de caractéristiques faciales, réseau de croyance bayésien.

Abstract

This Ph.D thesis work is dedicated to automatic facial analysis in 3D, including facial landmarking and facial expression recognition. Indeed, facial expression plays an important role both in verbal and non verbal communication, and in expressing emotions. Thus, automatic facial expression recognition has various purposes and applications and particularly is at the heart of "intelligent" human-centered human/computer(robot) interfaces. Meanwhile, automatic landmarking provides a prior knowledge on location of face landmarks, which is required by many face analysis methods such as face segmentation and feature extraction used for instance for expression recognition. The purpose of this thesis is thus to elaborate 3D landmarking and facial expression recognition approaches for finally proposing an automatic facial activity (facial expression and action unit) recognition solution.

In this work, we have proposed a Bayesian Belief Network (BBN) for recognizing facial activities, such as facial expressions and facial action units. A Statistical Facial feAture Model (SFAM) has also been designed to first automatically locate face landmarks so that a fully automatic facial expression recognition system can be formed by combining the SFAM and the BBN. The key contributions are the followings. First, we have proposed to build a morphable partial face model, named SFAM, based on Principle Component Analysis. This model allows to learn both the global variations in face landmark configuration and the local ones in terms of texture and local geometry around each landmark. Various partial face instances can be generated from SFAM by varying model parameters. Secondly, we have developed a landmarking algorithm based on the minimization an objective function describing the correlation between model instances and query faces. Thirdly, we have designed a Bayesian Belief Network with a structure describing the casual relationships among subjects, expressions and facial features. Facial expression or action units are modelled as the states of the expression node and are recognized

by identifying the maximum of beliefs of all states. We have also proposed a novel method for BBN parameter inference using a statistical feature model that can be considered as an extension of SFAM. Finally, in order to enrich information used for 3D face analysis, and particularly 3D facial expression recognition, we have also elaborated a 3D face feature, named SGAND, to characterize the geometry property of a point on 3D face mesh using its surrounding points.

The effectiveness of all these methods has been evaluated on FRGC, BU3DFE and Bosphorus datasets for facial landmarking as well as BU3DFE and Bosphorus datasets for facial activity (expression and action unit) recognition.

Keywords: 3D face, facial expression recognition, action unit recognition, face landmarking, statistical facial feature model, Bayesian belief network.

Chapter 1

Introduction

1.1 Research topic

Human face contains important and rich visual information for identification and communication, particularly for expressing emotion. Thus, analysing human face benefits a wide variety of applications from public security to personal emotion understanding, and from human computer interface (HCI) to robotics.

A problem of interest dealing with face analysis is facial expression recognition. Indeed, the traditional HCI that neglects facial expression excludes important information which can stimulate computer/robot to initialize proactive and socially appropriate behaviour during the communication process. This interaction between human and computers/robots is computer-based. It emphasizes the transmission of explicit information from texts, voices and gestures but ignores implicit information about the user. However, studies on human interaction paradigm suggest that facial expression contributes more than 50 percent to the effect of the spoken message as a whole while verbal part of a message contributes less than 10 percent to the effect of the message. Moreover, facial expression is one of the bases for understanding human emotional state, which is expressed through the contraction of face muscles resulting in facial appearance and geometry changes. Therefore, facial expression is an important cue for understanding emotions and its automatic recognition is a fundamental step for elaborating "intelligent" human/computer interactions.

Among various aspects of face analysis, we mainly focus on face landmarking and facial expression recognition problems in 3D. Indeed, landmarking, which consists in automatically detecting points of interest on the face, is a fundamental step for further processing and is an important part in an automatic facial analysis system,

particularly for facial expression recognition which is the second important contribution of our work. Moreover, we have proposed contributions related to new 3D face features, and to face tracking in 2D videos for people counting.

1.2 Problems and objective

Since images and videos can be easily accessed, 2D face analysis has been at the heart of many research works for several years, such as detection, tracking, recognition and expression recognition. However, head pose and illumination variations impose strong hurdles on these problems. Recently, 3D face has emerged as a promising solution in face processing and analysis. There are several reasons that the 3D face has gained many interests. Firstly, 3D facial geometry is invariant to pose and illumination conditions so that exploitation of 3D geometry can tackle various problems encountered by 2D face analysis. Secondly, 3D face carries ample information on both geometry and texture. This helps to improve the recognition accuracy and the analysis of subtle facial motions. Finally, exploring the relationship between texture and geometry of 3D face provides auxiliary means for 2D face analysis, making it possible to reconstruct 3D shape from 2D face.

While 3D faces are theoretically reputed to be insensitive to lighting condition changes, they still require to be pose normalized and correctly registered for further face analysis. As most of the existing registration techniques assume that some 3D face landmarks are available, a reliable localization of them is capital. There exist many landmarking methods in 2D, such as PDM, ASM, AAM and CLM, which can be applied to the texture maps of 3D faces. Landmarking on 3D faces is then realized by mapping those 2D points to 3D face meshes. However, landmarking directly on texture of 3D faces still encounters the pose and lighting problems. Moreover, 3D face scans synthesized from stereovision systems generally do not include this kind of mapping. Thus, most of existing 3D landmarking methods are based on 3D geometry. They obtain a high accuracy when locating shape-salient landmarks like nose tip. Unfortunately, the accuracy dramatically decreases when they attempt to locate other landmarks such as mouth corners or eyebrow corners,

which either are distributed on non-rigid regions of face or do not have salient local shape. Consequently, the number of landmarks provided is limited and the locating accuracy is not robust to conditions that cause changes on face geometry, such as expression and occlusion. Therefore, our objective in this work is to develop an approach robust to expression and occlusion making use of both shape and texture properties so that a larger number of shape-salient and non-shape-salient landmarks can be located.

Compared to 2D methods, facial expression analysis in 3D is more promising because 3D faces carry rich information and they are insensitive to lighting and head pose conditions. This explains the more recent significant increase in the number of studies dealing with 3D expression recognition. These studies are either based on facial feature or based on deformable face models, both of which have their own drawbacks. On the one hand, feature based approaches generally rely on a large number of landmarks which are used either for feature extraction or for face segmentation. Most of their performance highly depends on the landmark precision but automatically located landmarks do not have a sufficiently high precision. Therefore, these approaches require manually located landmarks and are not suitable for automatic systems. On the other hand, model based approaches deform 3D deformable models to fit testing faces by minimizing energy functions on face texture and/or geometry and recognize expression using fitted model parameters. This limits the usage of those interesting features which have better discrimination power on expression to enhance the recognition results. Therefore, our goal is to propose an efficient approach that is applicable to real world use cases. To do so, three requirements should be fulfilled: being robust to the imprecision of automatically located landmarks, being easily extended to recognize more expressions other than the universal expressions, and being flexible to integrate new features easily to enrich the face knowledge for better recognition when more expressions other than the six universal expressions or facial Action Units are considered.

1.3 Our approach

Human face properties are carried by rigid parts corresponding to the skeleton and by non-rigid parts corresponding to face muscles whose configuration mainly depends on emotion. Since the variations of these properties influence both face shape and texture, the problem of interest turns to be how we can learn these variations and apply them to the face analysis. Statistical models have been widely used for this purpose, which learn the major variations of face by Principle Component Analysis (PCA) and synthesize new face instances by a combination of learnt variation modes. This process can be understood as the construction of a face space, whose bases are eigenvectors of variations from PCA and new face instances can be associated with a point in this space. Previous statistical models of 3D faces are generally built on the whole face, such as the 3D Morphable Model. The drawback is that they cannot be properly fitted into new face scans when faces are partially occluded and thus have local shape deformations. In order to solve this problem, we propose in this thesis to build statistical face models from local regions configured by a global morphology, and apply them to face landmarking and expression recognition.

Specifically, for face landmarking, to overcome the accuracy decrease when locating non shape-salient landmarks, we consider both geometry and texture information so that all landmarks can be featured prominently. The statistical models are learnt from training faces with all kinds of the universal expressions in order to include the expression variations. By doing so, face instances with expression can be generated for landmarking on those faces with expression. Moreover, our statistical models are built from local regions so that it is still applicable for partially occluded faces by fitting itself based on unoccluded face regions.

In order to develop a facial expression recognition method being robust to automatic landmark errors, features are extracted from local shape and local texture around landmark locations. A graphical model Bayesian Belief Network (BBN) has been designed to estimate the emotion states or facial action units. The structure of this BBN allows integrating different features and expressions in a flexible manner. The BBN parameters are computed thanks to our proposed statistical feature mod-

els. By combining this graphical model with our automatic landmarking approach, we are able to implement a fully automatic and efficient system for facial expression recognition.

1.4 Our contributions

These two main contributions of this PhD thesis involve 3D face landmarking and 3D facial expression recognition. Moreover, we also contribute to new 3D face features, and to face tracking in 2D videos containing drastic changes on face scale for people counting.

3D face landmarking: Landmarking is essential for most of the face processing and analysis methods. We target on locating a large number of feature points on 3D faces under challenging conditions, i.e., expression and occlusion, so that automatically detected landmarks can be used for facial expression analysis. We have intentionally chosen those landmarks that can be used for face registration and facial expression recognition. To do so, we have proposed a 3D statistical facial feature model (SFAM), which learns both global variations in 3D face morphology and local ones around each landmark in terms of local texture and geometry respectively by PCA. By varying control parameters of their corresponding sub-models in the SFAM, we can thus generate different 3D partial face instances. The fitting of SFAM into an input face is based on an optimization of an objective function describing the correlation between model instances and faces. It also contains a set of parameters modeling local occlusion for which we proposed an automatic detection.

Facial expression recognition: In order to flexibly add expression/action unit classes and combine features from different representations (morphology, texture and shape) to improve the recognition rate, we have designed a Bayesian Belief Network (BBN) for 3D facial expression recognition. The structure of the BBN describes the casual relationships among subject, expression and facial features. Facial expression or action unit are modeled as the states of the expression node and are recognized by identifying the maximum of beliefs of all states that are inferred from the feature evidence on the target face. Different from the other graphic models for expression

analysis, we have proposed a novel method to compute BBN parameters based on a statistical feature model, which can be considered as an extension of SFAM. Systematical studies have been conducted to evaluate the effectiveness of the BBN under different configurations as well as comparisons with other classifiers like SVM and SRC. The combination of BBN used for expression recognition and SFAM used for landmarking provides a fully automatic facial expression system that is applicable to real world use cases.

Other contributions dealing with face analysis:

New Feature for 3D facial expression: Pose-invariant features for 3D faces can be a shortcut for face analysis because it avoids the procedure of face alignment. However, previous proposed curvature based features, such as shape index and HK curvature, are sensitive to facial surface noise. We have proposed a novel feature which derives only from point clouds of 3D faces to describe the local shape properties. It is easy to be implemented and can be used to enrich information used for face analysis.

Face tracking and people counting: A problem of interest regarding face analysis in image sequences is face tracking, which has numerous interesting applications such as people counting, the problem we have addressed in this work. In order to increase the face tracking accuracy under the scenarios where face scale varies dramatically, such as faces moving towards the camera, we have improved Kalman filter by integrating 3D information to better predict the face position. The improved Kalman filter is further combined with a kernel based object tracking algorithm so that the combined tracker is more robust to head pose variation and illumination. This tracker is initialized by an Adaboost face detector and outputs tracked face trajectories. People counting is then performed thanks to a trajectory classification algorithm we proposed that is based on the histogram of trajectory angles.

1.5 Organization of the thesis

This thesis is organized as follows:

Chapter 2 focuses on facial landmarking. We first introduce the background

about facial landmarking. Both facial landmarking methods in 2D and 3D are presented in the related works section. Then, we propose our 2.5D facial landmarking approach in the following section. Section 2.4 presents the SFAM with a fitting algorithm for 3D face landmarking. An occlusion detection and classification method is also proposed here so that this approach is applicable to landmarking on partially occluded faces. We draw our conclusion on facial landmarking at the end of this chapter.

Chapter 3 covers 3D facial expression recognition. We start the chapter by first introducing the development of facial expression analysis and motivations for facial expression recognition. Section 3.2 presents the emotion theories, facial expression properties and facial expression interpretations. We make a review of the state-of-the-art dealing with the classification of facial expressions in both 2D and 3D. In section 3.4, we present a 3D facial expression recognition approach based on a local geometry-based feature. This feature, named SGAND, is proposed in conjunction with a pose estimation algorithm for 3D faces since the face direction is required for the feature extraction. Section 3.5 presents our graphical model, BBN, for recognizing facial expressions and AUs with an uniform structure. A fully automatic facial expression recognition system is also presented in this section. Conclusions are drawn in section 3.6.

Chapter 4 presents a minor contribution: a people counting system which is based on face detection and tracking. The related works and the system framework are introduced in sections 4.1 and 4.2. Section 4.3 presents our face tracker whereas section 4.4 describes the people counter. Experimental results are given in section 4.5 and a conclusion is drawn in section 4.6.

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

Chapter 2

3D Face Landmarking

2.1 Introduction

A problem of interest concerning face analysis is landmarking which consists in locating facial landmarks. Facial landmarks are points of correspondence on faces that matches between and within populations [Mardia & Dryden 1998]. They have consistent reproducibility even in adverse conditions such as facial expression or occlusion [Farkas 1994]. These facial landmarks generally include nose tip, inner eye corners, outer eye corners, mouth corners, etc. They are not only characterized by their own properties in terms of local texture and local shape but also by their structural relationships which result from the global face morphology.

Uses of landmarks for face analysis are numerous and include important applications such as face alignment, registration, reconstruction, recognition and expression recognition. For example, irises are often located in 2D face images for normalizing the face scale and in-plane rotation; the triangle constructed by the nose tip and the inner corners of eyes can be used for aligning 3D faces; a set of landmarks are generally required from 2D face images and their corresponding peers on 3D face models for 3D face reconstruction. Furthermore, various features are extracted such as Gabor response and Local Binary Pattern for face recognition and facial expression recognition, either around facial landmarks or on face regions segmented by landmarks. An example of face alignment using landmarks on 2D face images is illustrated in Fig.2.1 .

2.5D/3D faces have recently emerged as a major solution in face processing and analysis to deal with pose and lighting variations [Bowyer *et al.* 2006]. A 2.5D face is a simplified three-dimensional (x, y, z) face representation that contains at most one

Figure 2.1: Face Alignment using landmarks [Huang *et al.* 2007]

depth (z) value for every point in the (x, y) plane. While 2.5D/3D face models are theoretically reputed to be insensitive to lighting condition changes, they still require to be pose normalized and correctly registered for further face analysis, such as 3D face matching [Lu *et al.* 2006, Zeng *et al.* 2008], tracking [Sun & Yin 2008], recognition [Gokberk *et al.* 2008] [Kakadiaris *et al.* 2007], and facial expression analysis [Niese *et al.* 2008]. As most of the existing registration techniques assume that some 2.5D/3D face landmarks are available, a reliable localization of these facial landmarks is essential.

When automatically locating landmarks on faces, approaches generally face the challenges of head pose, illumination, facial expression and occlusion. Head pose variations not only influence the facial appearance in images or video sequences but also cause self-occlusion where some landmarks are hidden. Illumination changes, including variations of lighting intensity and lighting source position, affect either pixel values over the whole face or those on face parts. Facial expressions due to facial muscle contractions cause non rigid deformations on face texture and shape, especially in the mimic parts of faces such as mouth regions. Occlusion is usually caused by face clusters, hand gesture or accessories such as glasses and masks. Some landmarks may become hidden and thus changes may appear concerning the local texture and shape around other landmarks. Although some approaches handle the variations of head pose and illumination in their learning stages for a better precision and robustness, these two problems are still challenging in 2D environment. Thus, 3D faces have gained interest since their processing and analysis may have the ability to deal with pose and lighting variations. Indeed, landmarks on 3D faces can be located through the face mesh analysing using curvature information or other geometry-based features. However, expression and occlusion remain open problems

for landmarking on 3D faces.

In this chapter, we focus on locating facial landmarks on 3D face scans, including those affected by the presence of facial expression and/or occlusion. We are convinced that local feature information and the structural relationships are jointly important for reliable face landmarking. Thus, contrary to the existing methods which rely on geometric information for 3D landmarking, we propose to solve the problem with statistical approaches using both local shape and texture of 3D faces as well as their global structure.

The remainder of this chapter is organized as follows. A review of landmarking algorithms in 2D and 3D is presented in section 2.2. Then, we present our statistical approach for landmarking on 2.5 faces in section 2.3. In section 2.4, we describe our statistical facial feature model and its application for locating landmarks on 3D faces with expression and occlusion. Finally we draw a conclusion in section 2.5.

2.2 Related works

Even if our work is focused on 3D face landmarking, a review of approaches developed in 2D is essential to understand their foundations and challenges that extension studies in 3D can aim at. Thus, current 2D face landmarking approaches are presented in subsection 2.2.1 followed by subsection 2.2.2 where state-of-the-art approaches for 3D face landmarking are introduced.

2.2.1 Face landmarking in 2D

Face landmarking has been extensively studied on 2D face images. These approaches can be divided into feature-based and structure-based categories.

2.2.1.1 2D face databases

2D landmarking approaches are tested on a variety of datasets due to the different usage of landmarks and different pose, illumination, scale conditions. Thus, it is hard to directly compare the efficiency of landmarking methods.

Databases for 2D face analysis are dedicated to different research fields, mainly including face detection, face tracking, face recognition, and facial expression recognition. It is hard to include all datasets here so that we only present some representative works. MIT Face Database [Database a], FERET Database [Phillips *et al.* 1998], Yale Database [Database b] are among those mainly for face detection; DXM2VTS [Teferi & Bigun 2007], A Video Database of Moving Faces and People [O'Toole *et al.* 2005] for face tracking; FERET [Phillips *et al.* 1998], Yale Database [Database b] for face recognition; Cohn-Kanade AU-Coded Facial Expression Database [Kanade *et al.* 2000], PIE database [Sim *et al.* 2003] and JAFFE database [Lyons *et al.* 1998] for facial expression analysis. A detailed description on these datasets and more datasets on 2D face analysis can be found following the link: http://www.face-rec.org/databases/.

2.2.1.2 Feature-based approaches

Feature-based approaches associate features extracted around landmarks in order to relocate them in new images. Features used for representing point properties include color [Zhao *et al.* 2008] [Talafova & Rozinaj 2007] , local intensity value [Beumer *et al.* 2006], Gabor wavelets [Feris *et al.* 2002] [Celiktutan *et al.* 2008] [Shih & Chuang 2004], gradient orientations [Yun & Guan 2009], discrete cosine transform (DCT) [Akakin *et al.* 2007], Scale-Invariant Feature Transform (SIFT) [Asbach *et al.* 2008], Speed Up Robust Features (SURF) [Kim & Dahyot 2008], etc. Landmarks are searched either based on some priori knowledge on face or based on two-class classifiers and similarity scores. The locating results may be further constrained or refined by geometrical information, like line properties between landmarks (distance or angles) or a global geometrical model.

These feature-based approaches can be classified into three categories:

- Prior knowledge based strategies: [Zhao *et al.* 2008] uses the prior geometry distances between eye regions as well as between the month region and eye regions. In this approach, eye centers are located by calculating the sum of RGB components difference and the mouth is supposed to be in a region lo-

cated by a certain distance below the eye centers. The mouth center is then located from color distribution. An correction rate of 95.52% for locating eye centers and mouth center is reached. When the Euclidean distance between an automatic eye center and its corresponding manual eye center is less than 0.25 of the Euclidean distance (pixel) between two manual eye centers (d_{rl}), automatic eye centers are considered as correctly detected. When the Euclidean distance between an automatic mouth center and the manual mouth center is less than 0.12 of d_{rl}, automatic mouth center is considered as correctly detected. [Talafova & Rozinaj 2007] search landmarks at the assumed subregions in a face image using human skin chromaticity and face morphological characteristics. Author gives the average locating errors of 5 pixels without indicating the resolution of testing face images. [Wang *et al.* 2009] uses the horizontal and vertical projection curves of gray values for locating eye, nose and mouth. The approach has been tested on JAFFE dataset without providing quantitative analysis on landmarking result.

- Similarity score based strategies: [Feris *et al.* 2002] performs a brute-force search within a limited window for a position that minimizes the score computed in a wavelet subspace. An detection rate over 94% for eight landmarks (four eye corners, two nostrils and two mouth corners) has been reached with an precision of 3 pixels on selected images (resolution of 640*480) from the Yale and FERET dataset. In [Beumer *et al.* 2006], the likelihood that a scanned pixel is the position of the landmark is evaluated by Approximate Maximum Discrimination Analysis (AMDA) using the texture values in a region surrounding landmarks. They name this method 'the MLLL-BILBO combination'. Tested on the texture images from part of FRGC dataset, 'the MLLL-BILBO combination' achieves an average errors of 0.042 for right eye, 0.046 for left eye, 0.058 for nose and 0.037 for mouth. This error is computed as the Euclidean distance (pixel) between automatic landmark and its manual equivalent divided by the inter-ocular distance. In [Akakin *et al.* 2007], each facial landmark is selected from the peaks in the corresponding matching score

map, which is computed by correlating DCT template with the DCT vector of the test block. Tested on texture images from FRGC dataset, at least 83% of outer eye corners, 90% of nose tip and 70% of mouth corners are correctly detected. The correct detection is evaluated as if the Euclidean distance in terms of pixels of an automatic landmark from the true position is less than 0.1 of the inter-ocular distance. [Asbach *et al.* 2008] compares potential landmarks with vertices on a normalized face mesh using SIFT and SURF description. They conclude that scale invariant Harris interest points with SURF descriptions are the most promising combination for locating landmarks. [Celiktutan *et al.* 2008] models facial landmarks redundantly by four different feature, i.e., DCT, Non-negative Matrix Factorization, Independent Component Analysis and Gabor Wavelets. Matching scores are later fused for selecting candidate points. over 94.8% of inner eye corners, over 93.8% and 89.7% of outer eye corners and inner eyebrow corners are correctly detected using a criterion of 0.1 of inter-ocular distance for correct detection.

- Classifier based strategies: another method presented in [Beumer *et al.* 2006] uses the Viola-Jones detector to classify a combination of Haar-like features on local texture around landmarks. Tested on the texture images from part of FRGC dataset, the approach achieves an average errors of 0.032 for right eye, 0.033 for left eye, 0.063 for nose and 0.041 for mouth. This error is computed as the Euclidean distance (pixel) between automatic landmark and its manual equivalent divided by the inter-ocular distance. Compared with 'the MLLL-BILBO combination' presented in the same paper, Viola-Jones detectors perform better on upper part of face but worse on lower part of face. In [Hanif *et al.* 2008], neural networks are trained to locate specific landmarks such as eyes and mouth corners in orientation-free face images. They report a mean location error of 0.12 over four landmarks. The normalized localization error is defined as the mean Euclidean distance between the detected landmark and the equivalent normalized with respect to the inter-ocular distance. In [Yun & Guan 2009], Adaboost classifiers are trained to classify gra-

dient orientation histograms calculated from the direction of interest points in their neighborhood. Tested on 480 image sequences of 120 subjects from Cohn-Kanade dataset and another database, their approach achieves a good performance of 90.69% average recognition rate for landmarking 26 landmarks with a criterion of 5 pixels between automatic landmarks and ground truth with a image resolution of 640*480. [Kim & Dahyot 2008] uses Support Vector Machines (SVM) to classify SURF local descriptors of landmarks. Tested on their own collected dataset, they achieve correct detection rates over 90% for detecting eye and mouth and 72% for detecting nose on images (resolution of 130*140) without providing a clear definition on correct detection.

The located landmarks or candidates may be further filtered or constrained by geometry information. In [Beumer *et al.* 2006], a shape correction algorithm is performed after AMDA to relocate those landmarks which do not comply with the constrains in a pre-trained shape model. In [Akakin *et al.* 2007], a probabilistic graph model chooses the best combination of located landmarks. In [Celiktutan *et al.* 2008], a structural completion method uses a graph composed of 12 landmark points to recover the missing landmarks. [Kim & Dahyot 2008] eliminates the wrongly classified descriptors based on geometrical constraints on relationship of facial component positions.

2.2.1.3 Structure-based approaches

Structure-based approaches are usually implemented via fitting a face model composed of both face shape and texture feature. Instead of extracting features from texture or shape separately, these methods use texture and shape knowledge simultaneously in the locating process. There exist many popular face models or approaches such as Active Shape Model and Active Appearance Model. Here we review some representative studies.

- Active Shape Model (ASM): [Cootes *et al.* 1995] have first proposed to represent the shape of an object using landmark points and to learn shape variations using Principle Component Analysis (PCA). Each landmark is associated with

the corresponding local texture. ASM is fitted by minimizing an objective function based on the sum of local texture similarity. [Park & Shin 2008] uses ASM to detect four landmarks for facial recognition. To apply ASM which is learnt from objects with great shape variations, [Xu & Ma 2008] introduce a Procrustes analysis technique to match feature landmark points in training and strengthening the edge in searching face profile. Tested on 240 images with a resolution of 640*480, over 95% for all 58 landmarks are correctly detected without detailing the correction criterion. For more accurate extraction of feature, [Sun & Xie 2008] combines both the local texture constraint and the global texture constraint in building and fitting ASM. Tested on 240 images (resolution of 640*480) from IMM database, the average error (measured as the Euclidean distance between an automatic landmark and its manual equivalent) over all 240 images for each of 58 landmarks is between two to four pixels in this method. [Huang *et al.* 2007] build separate Gaussian models for shape components instead of using statistical analysis on the global shape so that more detailed local shape deformations can be preserved. The nonlinear inter-relationships among the shape components are described by the Gaussian Process Latent Variable Model. Tested on 377 images (640*480) of group 1 from extended Yale Database, over 89.7% of all landmarks are localized with a precision of 12 pixels. [Faling *et al.* 2009] project a 3D shape model into images during the searching process with a 3D transform method to handle the great variance of head pose. Tested on their own dataset, the average errors (measured as the Euclidean distance between an automatic landmark and its manual equivalent) over all 200 images for ten landmarks is between three to five pixels for nearly frontal face views.

- Active Appearance model (AAM): AAM, originally proposed in [Cootes *et al.* 1998], learns a statistical shape model similar to ASM as well as a model from shape-normalized texture. The two models are further combined to form an appearance model by removing their correlations. The searching algorithm uses the residual between the current estimation of

appearance and the input image to drive an optimization process. Based on AAM, [Hou *et al.* 2001] proposes a direct appearance model which uses texture information directly in the prediction of the shape and in the estimation of position and appearance for faster convergence and higher accuracy. To locate landmarks robustly against the expression changes and scale variations, [Zhu & Zhao 2009] fits AAM with Lucas-Kanade algorithm, which minimizes the square sum of the difference between a template and an input image with respect to the warp parameters. They claim that about 90% of the test images give good results for locating 60 landmarks without a definition of the good criterion. [Yu & Yan 2009] applies AAM in automatically locating 17 landmarks for face recognition.

- Other Approaches combining local and global information: [Tu & Lien 2009] use Singular Value Decomposition (SVD) to combine two related classes, shape and texture, in a single eigenspace, named Direct Combined Model (DCM) algorithm. It estimates the facial shape directly by applying the significant texture-to-shape correlations. Tested on 450 images with resolution of 640*480 from their own database, they achieve an average error of 2 pixels for all 84 landmarks over images with the frontal pose. This average error increases to 8 pixels when head pose varies to 35 degrees. [Cristinacce & Cootes 2008] build a Constrained Local Model by learning global shape variations, local texture variations around landmarks, and their correlation. Correlation Meshes describing the similarity between local instances and local regions around landmarks are computed and used for optimizing a fitting function driven by shape parameters. Results have proved the efficiency of CLM compared to AAM. The fitting of this model is improved by using an optimization in the form of subspace constrained mean-shifts [Saragih *et al.* 2009]. [Kozakaya *et al.* 2008] proposes a weighted vector concentration approach, which integrates the global shape vector and locally normalized Histogram of Oriented Gradients (HOG) descriptor. Both the global and local information are combined in landmarking by solving a single weighted objective function. Tested on 1918 images

from FERET dataset, they achieve landmarking error between 0.03 and 0.07 for 14 landmarks. This error is measured as the average of Euclidean distance (normalized by inter ocular distance) between an automatic landmark and its manual equivalent over all tested images. The mean error over all 14 landmarks is 0.05.

In order to overcome the lighting problem, the above studies perform either an intensity normalization process [Cootes *et al.* 1995, Cootes *et al.* 1998], extract illumination-insensitive features, such as facial component contour and corner [Wang *et al.* 2009], or include illumination variations in their learnt facial models [Cristinacce & Cootes 2008]. Meanwhile, in order to overcome the head pose problem caused by the in-plane rotation and face scale, [Celiktutan *et al.* 2008] and [Xu & Ma 2008] detect landmark candidates under multi-directions and multi-scales. However, locating landmarks remains too challenging 2D approaches when dealing with faces with out-plane rotation and partially illuminated. Theoretically, 2D landmarking methods can be applied to 3D faces by locating landmarks on 2D texture maps and then using correspondence from the scanner systems to map those points onto 3D face meshes. However, due to the aforementioned limitations, dedicated landmarking approaches are necessary for locating landmarks on 3D faces.

2.2.2 Face landmarking in 3D

3D face landmarking approaches can be classified into three categories: 3D face mesh-based, 2.5D range map-based and geometry and texture combination-based approaches.

2.2.2.1 3D face databases

Most of 3D face analysis focused on 3D face recognition. Moreover, most of face recognition studies on 3D faces are conducted on the FRGC dataset [Phillips *et al.* 2005]. Thus, as a preprocessing step for face recognition, landmarking methods are often tested on this dataset. Other 3D datasets such as BU3DFE

[Yin *et al.* 2006] and Bosphorus database [Savran *et al.* 2008] are also used for testing 3D landmarking methods.

2.2.2.2 3D face mesh-based approaches

In this category, studies rely on face point cloud and triangulation. [Conde & Serrano 2005] calculate the spin images [Johnson 1997] based on 3D face meshes and uses Support Vector Machine (SVM) as classifiers to locate the nose tip and inner eye points. [Xu *et al.* 2006] define the 'effective energy' describing local surface feature, and designs a hierarchical filtering scheme to filter both this feature and a local shape statistical feature. SVM is then used to locate nose tip. Tested on 280 scans from 3DPEF dataset, a correction rate up to 99.3% has been achieved with a precision of 20mm on the unnormalized face scans. Nose tip is also detected in [Bevilacqua *et al.* 2008] by applying a generalized Hough Transform. In [Chang *et al.* 2006], nose tip and nose saddles are detected thanks to a calculation of the local curvature surface. Descriptors of local shape are calculated in [Huertas & Pears 2008] from inner product of points and their local plane normal. The search for nose tip and inner corner of eyes are based on matching these descriptors with ones from training, constrained by a graph model. Tested on 1507 non-normalized face scans from FRGCv2 dataset, over 90% of eye-corners and the nose-tip are located with a precision of 12mm and 15mm. Other studies fit a normalized face mesh with prior landmark knowledge to a new face for locating its landmarks. [Kakadiaris *et al.* 2007] build an annotated deformable face model on face mesh. By fitting the model to new faces, landmarks are naturally implied to the correspondence of the model annotation. [Irfanoglu *et al.* 2004] find a dense correspondence between an annotated base mesh to new faces. A similar surface model is built by [Hutton *et al.* 2003], however, their surface model can be considered as a 3D point distribution model (PDM). A recent work done by Nair [Nair & Cavallaro 2009] builds another 3D PDM model for face detection, landmarking and registration. In order to be insensitive to expressions, its points are located only in nasal and ocular regions. Tested on non-normalized face scans from BU3DFE dataset, their methods locates inner corners of eyes (left and right),

outer corners of eyes (left and right) and nose tip with mean errors (absolute distances between automatic landmarks and manual equivalent) of 11.89mm, 12.11mm, 19.38mm, 20.46mm and 8.83mm. Because features related to 3D face meshes are only based on the geometry information, it is hard to distinguish geometry-non-salient points e.g., eyebrow corners, and therefore the points that can be located are limited. Moreover, shape variations like expression, occlusion and self-occlusion can easily handicap this branch of landmarking approaches.

2.2.2.3 Range map-based approaches

Range maps from 3D faces can be considered as 2D images with pixel values corresponding to the linear transformation of Z coordinates in 3D. Thus, besides calculating curvature and shape index, popular features in 2D, like Gabor wavelet, can also be extracted from this representation. [Lu *et al.* 2004] and [Colbry *et al.* 2005] first find nose tip by closest Z value to camera and then find other landmarks using shape index within eye, mouth and chin regions. Tested on non-normalized 113 scans from their own database, the mean error for locating the five landmarks is 10mm. [Colbry & Stockman 2007] propose a canonical face depth map on range image and locates nose tip and inner eye corners based on this representation. [Colombo *et al.* 2006] and [Szeptycki *et al.* 2009] compute Gaussian (K) and Mean (H) curvature for each point in range image and set threshold on curvature to isolate candidate regions for nose tip and eye inner corners. The candidate landmarks are further filtered according to the shape of the triangles they compose in [Szeptycki *et al.* 2009]. Tested on 1600 face scans from FRGC dataset, over 99% of the three landmarks are localized with a precision of 10mm in [Szeptycki *et al.* 2009]. HK curvature can also be used on range images of full 3D head scans [Li *et al.* 2002], which locate six landmarks when point curvature properties fulfil a set of empirical conditions. In [Segundo *et al.* 2007], nose tip, nose corners and eye corners are initially located using HK curvature and their positions are then corrected by finding salient points on projection curves of range images. Over 99% of all these landmarks are correctly detected. However, no specific criterion on good detection has been provided. [D'House *et al.* 2007] use Gabor wavelets on range map for a coarse

detection of landmarks and then apply Iterative Closest Points (ICP) algorithm to enhance the location precision. They achieved 99.37% of correct nose tip location with a precision of 10mm on FRGC v1.0 database, but their accuracy on the outer corners of the eyes is relatively lower. [Dibeklioglu *et al.* 2008] propose a Gaussian Mixture Models liked statistical model (MoFA), describing local gradient feature distribution around each landmark. This model produces a likelihood map for each landmark on new faces and the highest value in this map is located as landmark. Tested on FRGC dataset, their approach achieves over 90%, 99%, 99% and 87% of detection rates for outer eye corners, inner eye corners, nose tip and mouth corners respectively with a precision of three pixels on texture maps (resolution of 480*640). [Koudelka *et al.* 2005] develop an accurate approach by computing radial symmetry maps, gradient and zero-crossing maps from range maps. Landmarks are chosen by using a series of heuristic constraints. Over 97% of all five landmarks (nose tip, eye inner corners, mouse center and sellion) are localized with a precision of 10mm on FRGCv1 dataset. Compared with the first category, this branch of approaches can localize more points with higher accuracy, because faces in range maps are normally in frontal pose and facial landmarks can be represented with 2D features. Nevertheless, the drawbacks of these methods is their sensitivity to face scale and head pose variations. Moreover, they have difficulty to locate non-salient points in geometry and points in non-rigid face regions with the presence of expressions.

2.2.2.4 Approaches based on a combination of facial geometry and texture

Due to the above reasons, a single face representation may not provide enough information for localizing some landmarks consistently. However, the perfect matching of range map and texture map from scanner systems ensures the combination correctness of multi-representation. Accumulating evidence derived from different face representations has the potential to make the feature extraction richer and more robust. [Boehnen & Russ 2004] compute the eye and mouth maps based on both color and range information and selects potential feature candidates of inner corner of eyes, nose tip and sub-nasal. A 3D geometric-based confidence of candidates

is computed to aid in the selection of best location. Tested on FRGC dataset, their approach achieve 99.6% of correction location on FRGCv2 dataset without providing a criterion for correct detection. [Wang *et al.* 2002] use "point signature" representation to code face mesh as well as Gabor jets of landmarks from 2D texture images. In [Jahanbin *et al.* 2008a] and [Jahanbin *et al.* 2008b], an extended elastic bunch graph is proposed to locate landmarks, in which Gabor Wavelet coefficients are used to model local appearance in texture map and local shape in range map. Tested on ADIR dataset, their approach achieve a correct detection rate of 98% for 11 landmarks with a precision of $0.06m_e$, where m_e is the inter-ocular distance in terms of mm. In [Salah & Akarun 2006] and [Salah *et al.* 2007], the distribution of Gabor response of texture images around landmarks are featured and modeled by MoFA. 3D surface normals are used to remove illumination effects from texture images. They locate landmarks on downsampled images for a coarse locating. Tested on FRGCv1 dataset, they achieve correct detection rate of 83.4%, 97.2%, 98% and 37.8% for outer eye corners, inner eye corners, nose and mouth corners with a precision of 3 pixels on downsampled images with a resolution of 60*80 pixels. A fine detection can be found in [Akakin *et al.* 2006], where larger search regions on the original texture images and range images are cropped to produce higher dimensional Gabor-jets. The results prove obvious increase in locating accuracy when two stage detection is performed. [Lu & Jain 2005] fit a statistical model constructed as the average 3D position of landmarks based on pre-detected nose tip, and then computes and fuses shape index response (range) and cornerness response (texture) in local regions to locate seven points. They extend this method to be insensitive to head pose variation in [Lu & Jain 2006]. Tested on non-normalized face scans from FRGC dataset, seven landmarks are located with mean errors between 3.6mm to 7.9mm. The error is measured as the Euclidean distance (in 3D) between the automatically extracted landmarks and the manually labelled ones. These landmarks are further tuned by Iterative Cloud Point algorithm in [Lu *et al.* 2006]. These approaches demonstrate the robustness in landmarking both geometry salient and appearance salient landmarks. However, since they extract features directly from 2D texture and range maps, pose and scale remains strong difficulties for methods

in this category.

2.2.3 Discussion

Despite the increasing amount of related literature, face landmarking is still an open problem. Indeed, current face landmarking techniques need to increase both accuracy and robustness, especially in the presence of lighting variations, head pose, scale changes, facial expressions, self-occlusion and occlusion by accessories such as hair, moustache and eyeglasses [Salah *et al.* 2007]. This chapter aims at proposing a general framework for precise 3D face landmarking robust to facial expression and partial occlusion.

Face landmarking has been extensively studied on 2D facial texture images as discussed in previous section. An interesting approach is 2D statistical models such as the popular Active Appearance Model [Cootes *et al.* 2001] or more recently the Constrained Local Model (CLM) [Cristinacce & Cootes 2008] which carry out statistical analysis both on facial appearance and 2D shape. However, these approaches, while working on 2D facial texture images, inherit the sensibility to lightening and pose changes.

Works on 3D face landmarking are rather recent. Most of them try to best embed a priori knowledge on landmarks on 3D face, computing response of local 3D shape-related features, such as spin image [Kakadiaris *et al.* 2007], effective energy [Xu *et al.* 2006], Gabor filtering [D'House *et al.* 2007] [Colbry *et al.* 2005], generalized Hough Transform [Bevilacqua *et al.* 2008], local gradients [Dibeklioglu *et al.* 2008], HK curvature [Colombo *et al.* 2006], shape index [Lu *et al.* 2006] and curvedness index [Nair & Cavallaro 2009], radial symmetry [Koudelka *et al.* 2005], etc. While these approaches enable rather accurate shape prominent landmark detection such as the nose tip or the inner corners of eyes, their localization precision drastically decreases for other less prominent landmarks.

As current 3D imaging systems can deliver registered range and texture images, a straightforward way for more discriminative local features is to accumulate evidence from both the two face representations, i.e. face shape and texture. For

instance, [Boehnen & Russ 2004] compute the eye and mouth maps based on both color and range information. [Wang *et al.* 2002] use "point signature" representation coding 3D face mesh as well as Gabor jets of landmarks from 2D texture image. In [Salah *et al.* 2007] [Jahanbin *et al.* 2008b], Gabor wavelet coefficients are used to model local appearance in texture map and local shape in range map around each landmark while [Lu & Jain 2006] propose to compute and fuse shape index response (range) and cornerness response (texture) in local regions around seven landmarks.

As the combinations of candidate landmarks resulting from shape and/or texture related descriptors are generally important, some authors also propose to make use of structural relationships between landmarks, for instance through heuristics [Nair & Cavallaro 2009], a 3D geometric-based confidence [Boehnen & Russ 2004], an extended elastic bunch graph [Jahanbin *et al.* 2008b], or a simple mean model constructed as the average 3D position of landmarks from a learning dataset [Lu & Jain 2005]. However, there is no approved technique which best takes into account both configuration relationships between landmarks and the local properties in terms of geometric shape and/or texture around each landmark.

Few of the aforementioned studies address the issue of face landmarking in the presence of facial expression or occlusion. [Nair & Cavallaro 2009] experiment their 3D Point Distribution Model to locate five landmarks (the two outer eye points, the two inner eye points and the nose tip) under facial expressions with a locating accuracy ranging from 8.83 mm for nose tip to 20.46 mm for the right outer eye point. However, these five landmarks are all located on face regions stable to facial expressions. [Dibeklioglu *et al.* 2008] study 3D facial landmarking under expression, pose and occlusion variations. However, only one landmark, the nose tip, was considered in their work which is not sufficient for further accurate face analysis.

In this chapter, we address the facial landmarking problem in 3D with presence of expression and occlusion, aiming at locating a sufficient number of landmarks with good accuracy for other face analysis, especially for facial expression recognition. In order to do so, we propose a general learning-based framework for 3D face landmarking which combines configuration relationships among the landmarks and

their local properties of texture and geometry. Based on this principle, we propose two approaches in section 2.3 and section 2.4 for landmarking 2.5D faces and 3D faces respectively. Statistical face models are trained by applying Principle Component Analysis (PCA) to face landmark configurations, local texture and local shape around each landmark from training faces. Two different fitting algorithms are proposed to fit the face models to new faces so that landmark locations can be found by searching the closest points to known landmarks on fitted models. Provided with 3D training faces with expressions, our models are able to learn the expression variations and generate instances with these variations so that the accuracy in fitting faces with expression can be increased. Moreover, in order to overcome the occlusion problem, a classification system allowing to detect occluded faces and the type of occlusion has been proposed, so that occlusion information can taken into account during the fitting process of our second approach presented in section 2.4.

2.3 A 2.5D face landmarking method

In this section, we propose a statistical learning-based approach for 2.5D face landmarking. Taking benefit from the rich information contained in 3D face data, our model is built not only based on facial texture but also based on the geometry variations from a training face set. Specifically, a variety of face shape on texture maps are analyzed and learnt as the global configuration of landmarks. Meanwhile, variations on local texture and range are also learnt from the scale-free patches around each landmark. Thus, the statistical model is made up of a global face shape model, a texture model and a range model. New patch instances can then be synthesized by varying the model parameters. When fitted for a best match to a new 3D face, this statistical model delivers the location of the landmarks on the texture map of the input 3D face. The fitting process is the optimization of the global shape in order to reach the highest correlation in both texture and range between local patches from the input face and instances synthesized from our texture and range models.

2.3.1 Methodology

2.3.1.1 Preprocessing face scans

3D face scans delivered by the current 3D imaging systems are usually noisy and may contain holes and spikes, as shown in Fig. 2.2. In order to remove these noises, we perform the following operations to enhance the quality of 3D face scans:

1. Median Cut: spikes are detected by checking the discontinuity of points and are removed by the application of a median filter.

2. Hole Filling: holes are located by a morphological reconstruction and filled by cubic interpolation.

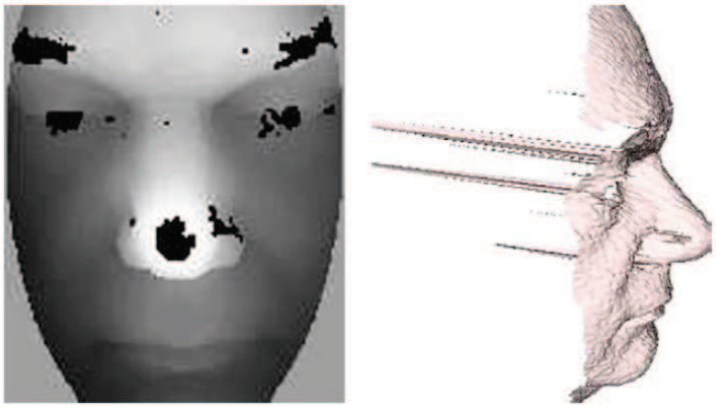


Figure 2.2: Holes and spikes on a 3D face scan

Although faces are scanned from the frontal view, there still remain variations in head pose which disturb the learning of global shape variations and consequently also may perturb the learning of local shape and texture variations. To compensate head pose, faces are first translated near to the origin of the coordinate system by subtracting the gravity center of the point cloud. Then, Iterative Closest Point (ICP) algorithm [Zhang 1994] is used to minimize the difference between two point clouds of the new face and an arbitrarily selected face which holds a frontal and straight pose, as illustrated in Fig. 2.3.

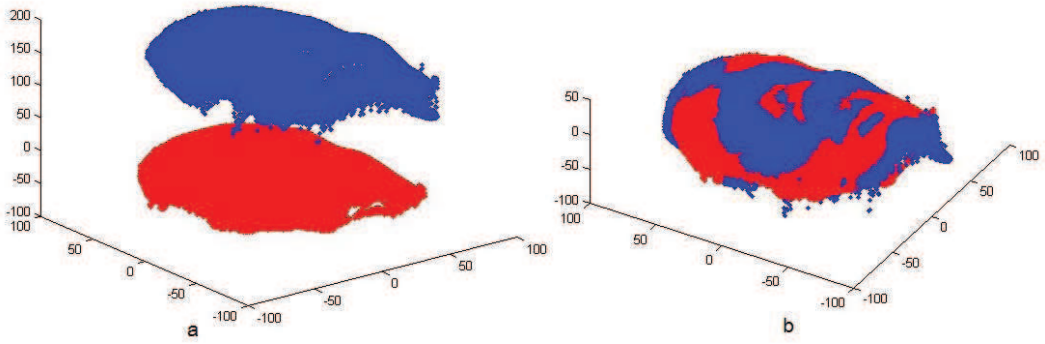


Figure 2.3: Point clouds of the preset face (red) and new face (blue) before ICP alignment (a) and after ICP alignment (b).

2.3.1.2 Creating scale-free patches in range and texture

Contrary to other methods which directly use texture and range maps for extracting local information, we process a local remesh on 3D point clouds. This is because the distance between subjects and the scanner affects the face scale of 3D face scans, which influences the 3D point cloud density and resolution of face in texture and range maps. Directly sampling from the texture map is sensitive to the scale variation and thus creates local patches covering different areas of face parts around landmarks. Therefore, we create the scale-free local patches with uniform scale among all faces to normalize face scale in local regions.

We consider 15 landmarks on each 3D face model in a learning dataset which need to be manually labelled as illustrated in Fig. 2.4. They are automatically associated with the corresponding 2D landmarks in the texture maps. 3D coordinates of points in the neighborhood of each landmark and their associated intensities are sampled. The neighborhood is centered at the corresponding landmark with a fixed length and width on the XY plane. The number of sampled points varies with the face scale.

Uniform grids for texture and range respectively are created around each landmark with a fixed size (15*15 in this work, a compromise between accuracy and efficiency) as shown in Fig. 2.5. We can benefit two factors from the uniform grids. Firstly, because the distances between subjects and the 3D scanner are different so

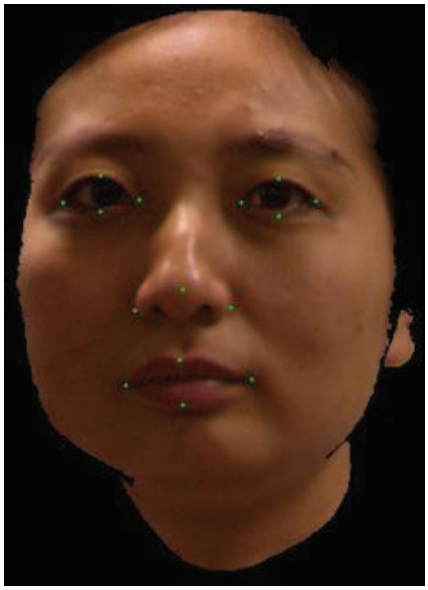


Figure 2.4: Manually labelled landmarks on a frontal 2.5D face.

that the number of points varies. This leads to variation on the density of point clouds of 3D faces. By resampling local points, this variation can be normalized. Secondly, there exists a nature correspondence of resampled points on grids centered at a specific landmark of different faces. This find the point-to-point correspondence among faces easily and efficiently.

Specifically, the centers of grids have the (x, y) values of their corresponding landmark. The intervals of grids on X,Y dimensions are fixed to 1mm. The range values (resp. intensity values) on the grids are interpolated from range values of sampled points in the local regions (resp. the intensity values). The interpolation methods used for range values and intensity values are different. Triangle-based linear interpolation is used for the intensity values, which computes the current intensity value based on the weighted distances from the point to three vertices of the triangle covering the point. The Biharmonic Spline Interpolation [Sandwell 1987] is used for the range values. The grids are then projected into 2D along Z direction and then range and texture patches around the landmark can be obtained. This process is repeated for all landmarks on a face.

Intensities and range values are then concatenated on all patches into two vectors G and Z as in eq.2.1 and eq.2.2 where m is the total number of points on all grids,

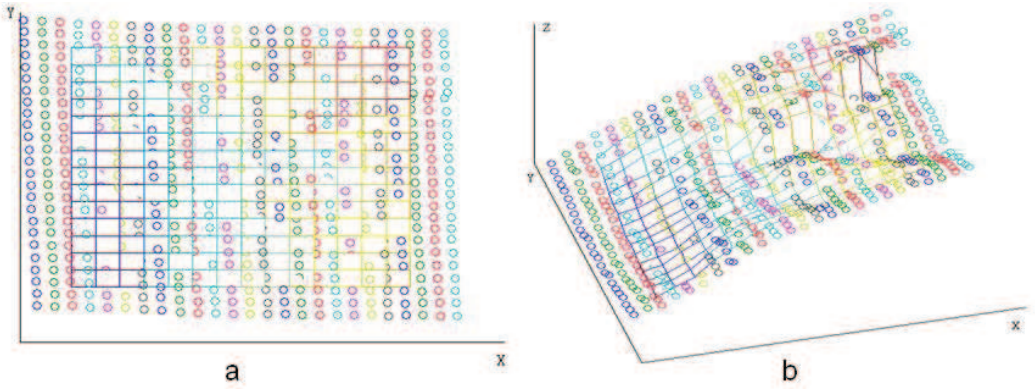


Figure 2.5: Creation of uniform grid in a local region associated with the left corner of the left eye from two viewpoints (a) and (b). Circles are the sampled points from the 3D face model and the grid composed of the interpolated points. The interpolation is also performed for intensity values.

(3375 here).

$$G = (g_1, g_2, ..., g_m)^T \tag{2.1}$$

$$Z = (z_1, z_2, ..., z_m)^T \tag{2.2}$$

2.3.1.3 Building a statistical landmark configuration model

Since 3D scanner systems create a point-to-point matching between pixels in 2D texture and vertex in 3D point cloud, the landmarks in 3D space can be mapped into the 2D texture map. Therefore, their positions can be obtained in the texture map and they are also concatenated into a vector X, as in eq. 2.3 where N is the number of landmarks.

$$X = (x_1, y_1, x_2, y_2..., x_N, y_N)^T \tag{2.3}$$

Then, all X vectors are normalized from training faces using a procrustes analysis in order to remove 2D global variations [Cootes *et al.* 1995]:

1. Select one shape to be the approximated mean shape.

2. Align the shapes to the approximated mean shape. First, calculate the centroid of each shape. Then, align all shapes centroid to the origin and normalize each shape scale. Finally, rotate each shape to align with the newest approximate mean.

3. Calculate the new approximate mean from the aligned shapes.

4. Repeat step 2 and 3 until the approximated mean converges.

Principal Component Analysis (PCA) is then applied where 95% major components have been preserved. Taken PCA on the set $\{X_i\}$ for example, the analysis is as follows:

1. Compute the mean of the data,

$$\bar{X} = \frac{1}{N_x} \sum_{i=1}^{N_x} X_i \tag{2.4}$$

2. Compute the covariance of the data,

$$\Sigma = \frac{1}{N_x - 1} \sum_{i=1}^{N_x} (X_i - \bar{X})(X_i - \bar{X})^T \tag{2.5}$$

3. Compute the eigenvectors, ϕ_i and corresponding eigenvalues λ_i of Σ (sorted in descending order),

4. Retain a number of eigenvectors ϕ_i to compose P_x so that the model represents some proportion (eg the sum of first eigenvalues reaches 95% of the sum of all eigenvalues) of the total variance of the data.

The same process is applied for the training sets of $\{X_i\}$, $\{G_i\}$, $\{Z_i\}$ to build the following three linear models (eq.2.6-2.8).

$$X = \bar{X} + P_x b_x \tag{2.6}$$

$$G = \bar{G} + P_g b_g \tag{2.7}$$

$$Z = \bar{Z} + P_z b_z \tag{2.8}$$

where $\bar{X}$, $\bar{G}$, $\bar{Z}$ are the mean 2D shape, mean normalized intensity and mean range value respectively; P_x, P_g, P_z are sets of modes of shape, intensity and range value variation respectively; b_x, b_g, b_z are sets of parameters of 2D shape, intensity and range values respectively. The dimensions of P_x, P_g, P_z are respectively $(2 * N, n_x)$, (m, n_g), (m, n_z), where n_x, n_g and n_z are the number of eigenvectors preserved, N is the number of landmarks and m is the total number of points from all grids around the landmarks.

In 2D statistical models such as ASM and AAM [Dryden & Mardia 1998, Cootes & C.J.Taylor 2004], the assumption that the control parameters in the model follow the Gaussian distribution has been proved to be efficient in many cases. Thus, following these studies, we assume that b_i from PCA where $b_i \in \{b_x, b_z, b_g\}$ are independent and Gaussian distributed with a zero mean and a standard derivation σ_i^j, where j refers to each parameter of b_i. Figures 2.6, 2.7 and 2.8, illustrate the first two modes ($j \in \{1, 2\}$) at their left and right ending variation (-$3\sigma_i^j$,+$3\sigma_i^j$), namely -3std and +3std, respectively for 2D shape variation, texture variation and range variation.

Thus, the statistical model built here includes three sub-models: shape, texture and range models. Similar to other statistical models, our model can be trained with different training sets and thus can learn different variation modes. The more diverse training faces are provided, the more comprehensive variations the model includes. For example, if we provide a training set including faces with different expressions and illumination conditions, the model is learnt with expression and illumination variations. However, if the model is trained with neutral faces under a single illumination condition, it only contain variations due to subject physiognomy.

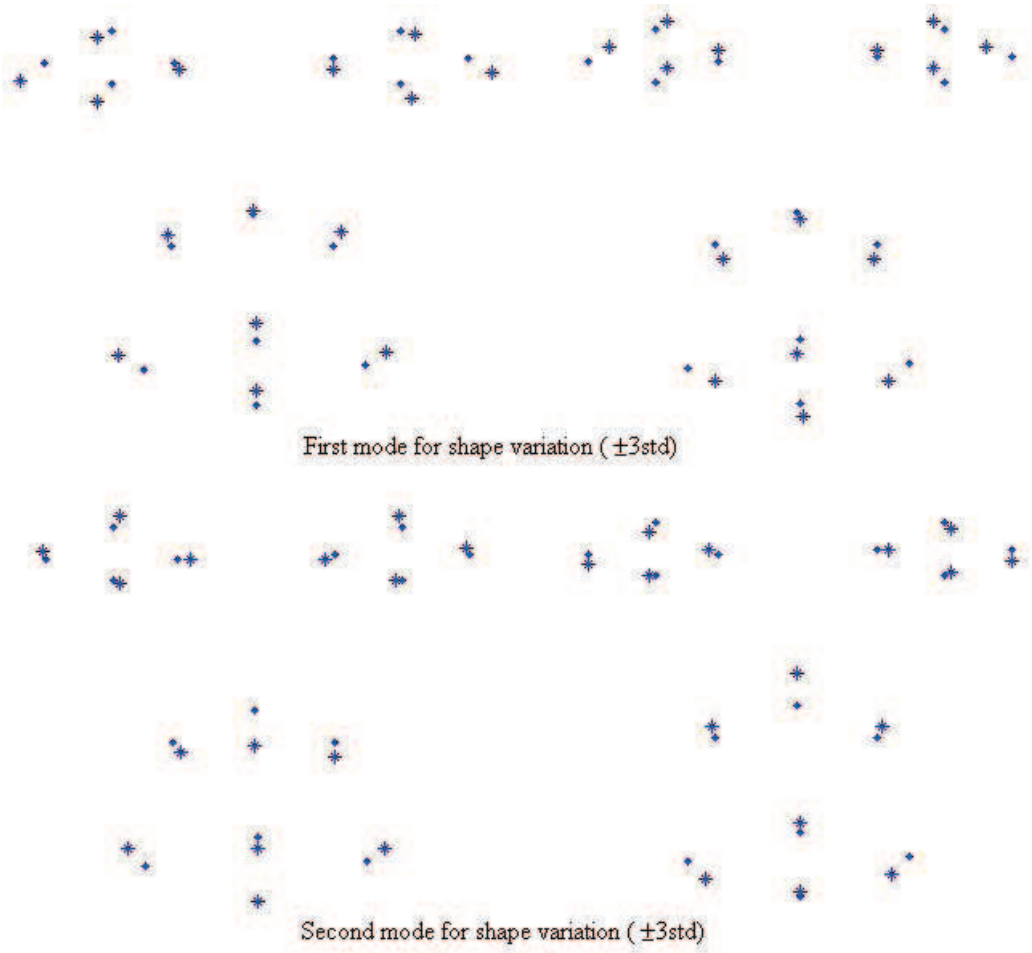
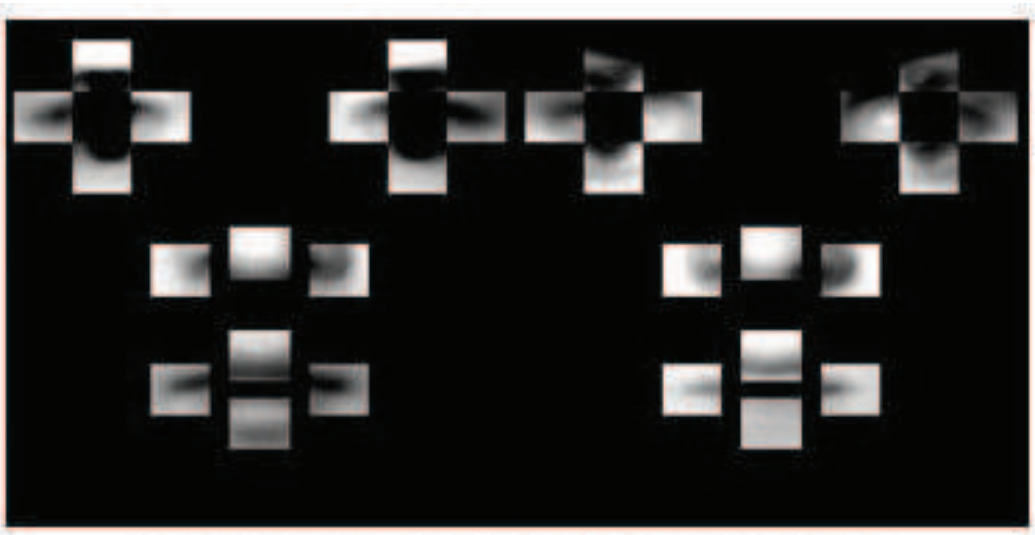
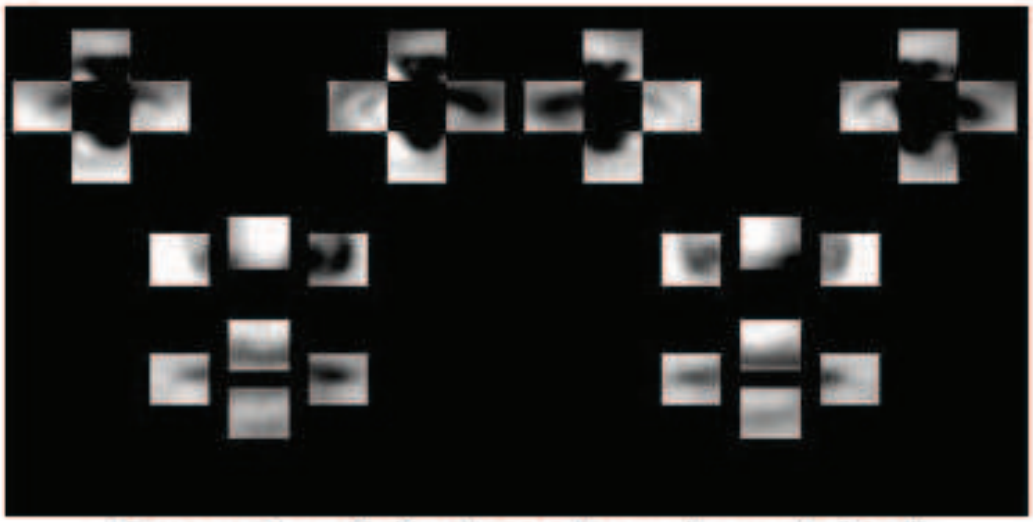


Figure 2.6: First two modes of shape variation in 2D. Points represented by '*' are current shapes while points represented by '.' are mean shape. The first variation mode mostly explains the shape changes along the horizontal direction while the second variation mode mostly explains the shape changes along the vertical direction.

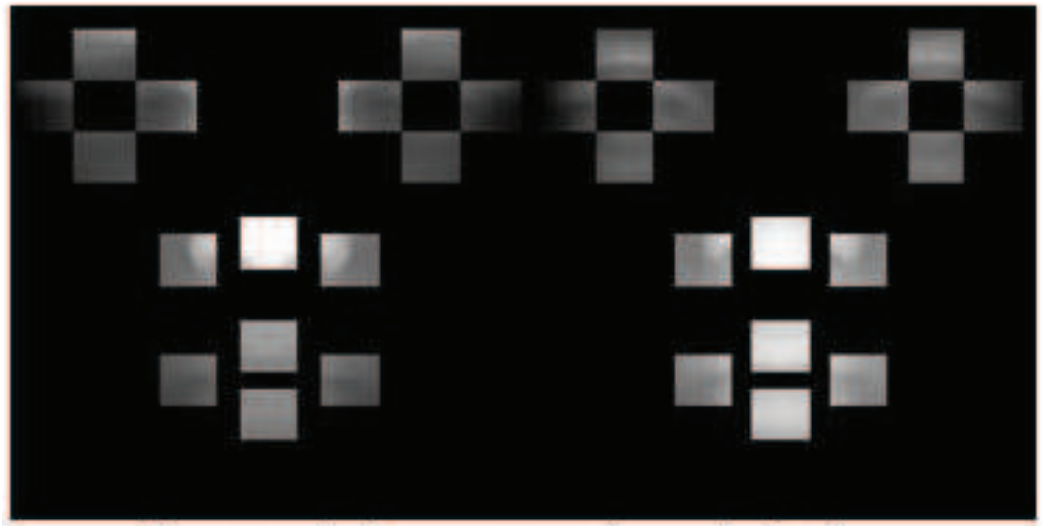


First mode for intensity variance(±3std)

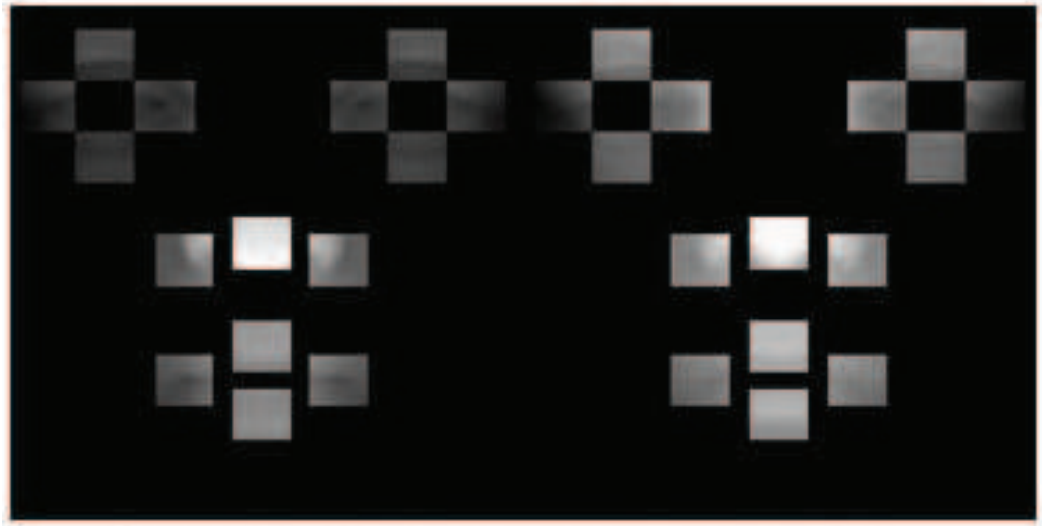


Second mode for intensity variance(±3std)

Figure 2.7: First two modes of texture variation. The first variation mode mostly explains the intensity changes in the eyebrow region and the mouth region, while the second variation mode explains the intensity changes in the nose region.



First mode for range variance(±3std)



Second mode for range variance(±3std)

Figure 2.8: First two modes of range variation. The first variation mode mostly explains the range value changes in the lower part of face, while the second variation mode explains the range value changes in the upper part of face.

2.3.1.4 Estimating instances from a face

P_x, P_g, P_z in eq. 2.6-2.8 contain the variation modes of shape, texture and range. Thus, given the parameters b_x, the 2D shape can be generated by using eq. 2.6. In order to transform it into a 2D image coordinate system, 3 more parameters are required, namely a translation parameter (C_x ,C_y), a scale parameter α and an in-plane rotation parameter ρ as described in eq.2.9.

$$X = \alpha \cdot (R(\rho) \cdot (\bar{X} + P_x b_x) + C) \tag{2.9}$$

where X is the created shape instance and $R(\rho)$ is the rotation matrix. The shape transformation parameters and shape parameters (b_x) are concatenated into a vector $\Theta = (b_x^T|C^T|\alpha|\rho)^T$.

Given a shape instance X and a preprocessed 3D face scan, the 2D points in X can be mapped back into 3D space based on the correspondence from the scanner system and vectors G and Z (eq. 2.1-2.2) can be obtained through the same process as the one described in section 2.3.1.2. They are further used to estimate b_g and b_z, as follows:

$$b_g = P_g^T(G - \bar{g}) \tag{2.10}$$

$$b_z = P_z^T(Z - \bar{z}) \tag{2.11}$$

However, in order to constrain the possible deformations, a boundary ($\pm 3\sigma_i^j$) is set for the corresponding parameter in b_i ($b_i \in \{b_z, b_g\}$). Thus, any b_i^j, which exceeds its boundary is replaced by its closest boundary. Then, texture and range instances $\hat{G}$ and $\hat{Z}$ can be generated according to eq. 2.7-2.8 using these constrained b_g and b_z.

2.3.1.5 Model fitting

Given a new 2.5D face, the landmarking problem is how to best fit our learnt statistical model on this face. This can be considered as an optimization problem

with an objective function depending on variable Θ.

Thus, by initializing Θ, a starting 2D shape instance $\hat{X}$ can be generated, and texture and range instances $\hat{G}$ and $\hat{Z}$ can be created as described in section 2.3.1.4. A normalized correlation is further computed for texture F_G (eq. 2.12) and range F_Z (eq. 2.13) on all local regions respectively, allowing to obtain the objective function in eq. 2.14.

$$F_G = \sum_{i=1}^{N} \left\langle \frac{G_i}{\|G_i\|}, \frac{\hat{G}_i}{\left\|\hat{G}_i\right\|} \right\rangle \tag{2.12}$$

$$F_Z = \sum_{i=1}^{N} \left\langle \frac{Z_i}{\|Z_i\|}, \frac{\hat{Z}_i}{\left\|\hat{Z}_i\right\|} \right\rangle \tag{2.13}$$

where $\langle \cdot, \cdot \rangle$ is the inner product and $\|\cdot\|$ is the L^2 norm. N is the number of landmarks.

$$f(\Theta) = -(F_G + F_Z) \tag{2.14}$$

The optimization of $f(\Theta)$ is performed based on the Nelder-Mead simplex algorithm [Nelder & Mead 1965].

The fitting procedure is as follows:

1. The initial shape instance $\hat{X}^0$ is created from the vector Θ^0 where b_x are zeros and C^T, S, ρ are preset.

2. Scale-free patches on range Z^k and texture G^k are interpolated as described in section 2.3.1.2.

3. Texture and range instances $\hat{G}^k$, $\hat{Z}^k$ are estimated following eq. 2.11 in section 2.3.1.4.

4. The function f^k is computed following the eq. 2.12, eq. 2.13, eq. 2.14.

5. Taking the Θ^k as variables and f^k as the objective function value, the optimization algorithm predicts Θ^{k+1} which leads to a lower value of the objective

function.

6. X^{k+1} is computed following the eq. 2.9 and compared with X^k to check the convergence. If convergence is not reached, go to 2.

In order to initialize C^T and S, an Adaboost face detector [Viola & Jones 2002] can be applied on the 2D texture maps and outputs a box containing faces. Thus, these two parameters can be estimated by the center and length of the box respectively. However, in our implementation, this initialization is performed thanks to face masks obtained from the scanner system which has the advantage to be accurate and much simpler. ρ are preset to zero. In order to constrain the deformations and to ensure that the shape instance is plausible, b_x^j parameters are also limited within the boundary $\pm 3\sigma_x^j$. All trespassing b_x^j values are replaced by their closest boundary.

Note that it is not necessary to perform a size normalization before and during the fitting process since three parameters which project shape instances into the image coordinate are optimized during this fitting process. Moreover, there is no photometric normalization done before the model fitting since the objective function computes the correlation between scale free patches and their estimated instances. This process has more tolerance to illumination conditions compared to those directly extracting features on the images for landmarking.

2.3.2 Experimental results

2.3.2.1 Database

The datasets we have used are FRGC v1.0 and v2.0 [Phillips *et al.* 2005]. The first version of the FRGC dataset contains 953 face scans from 275 people, captured under controlled illumination conditions and generally neutral expressions. However, these 953 face scans have slight head pose variation and scale variation. The second version of the FRGC database contains 4,007 face scans from 466 persons. These 3D face scans were captured under different illumination conditions and contain various facial expressions, including happiness or surprise, etc..

All faces have been manually labelled by our research group with 15 landmarks as illustrated in Fig. 2.4. These manually labelled landmarks can be used as ground truth for learning or quality assessment of automatic landmark location. In our experiments, the whole FRGC v1.0 dataset is first cleaned by filtering out several badly captured face models. It is further divided into two parts, the first half part (452 faces) is used for training, and the second one (462 faces) for testing our algorithm. Subjects in the training set are different from those in the testing set. For comparison purpose, we also applied to this testing set the curvature analysis based method developed within our team [Szeptycki *et al.* 2009]. However, only 9 landmarks can be used for comparison between these two techniques as the curvature analysis-based method can not locate the other 6 landmarks which do not have prominent curvature properties. In order to assess the generality of our statistical model which is learnt from 3D face models from FRGV v1.0, 1400 faces are randomly selected from FRGC v2.0 dataset as an extended testing set. The precision in all tests is measured as the mean locating error ($P_r = \sum_{i=1}^{N} d_i$) where d_i is the 3D Euclidean distance between a landmarks automatically located and its corresponding manually labelled landmark.

2.3.2.2 Results

Fig. 2.9 displays the accumulative precision of all landmarks located by our model on the testing set from FRGC v1.0 dataset. As we can see, our model can locate 97% cases in 10mm precision and 100% in 20mm precision for all landmarks. Our method achieves its best location result for landmark 13 (see legend in Fig. 2.9) with a 100% accuracy in the precision of 9mm, and the worst one for landmark 7 which displays 100% accuracy only in the precision of 19mm. Fig. 2.10 shows the precision curves displayed by the curvature analysis based method [Szeptycki *et al.* 2009] on the same testing set. As we can see from the figure, while the nose tip and inner corner of eyes, having each prominent geometric feature, are better located by the curvature analysis-based method, our statistical model displays better precision on all the other landmarks.

The first two rows in Table 2.1 shows the mean and std of locating error for each

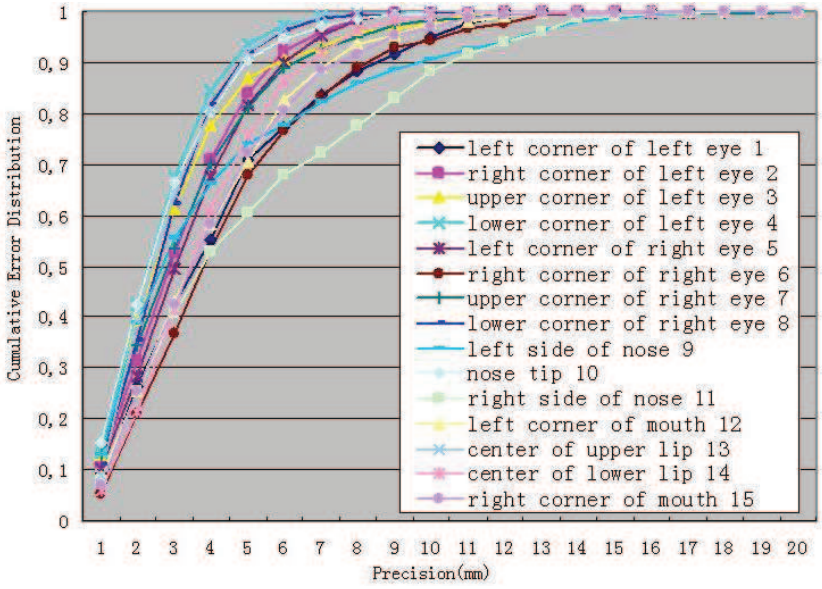


Figure 2.9: Precision curves for all landmarks located by our method

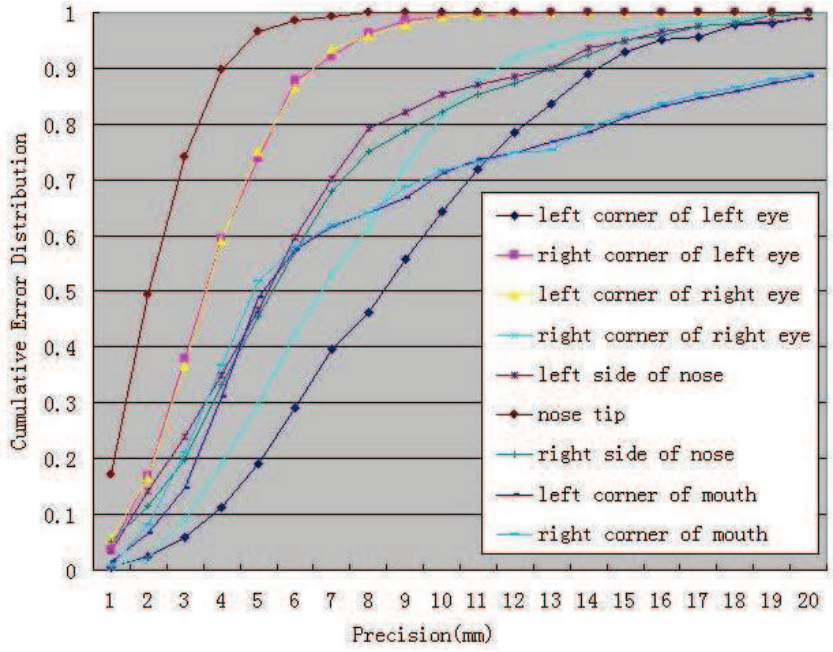


Figure 2.10: Precision curves for all landmarks located by the method in [Szeptycki *et al.* 2009]

Table 2.1: Mean and deviation of locating errors for all landmarks using FRGC v1.0 (mm)

	1	2	3	4	5	6	7
Mean	4,15	3,11	2,98	2,50	3,30	4,38	3,28
Std	2,82	1,90	2,23	1,51	2,04	2,81	2,43
Mean	8.76	3.85	-	-	3.84	7.16	-
Std	4.24	2.02	-	-	2.03	3.46	-
8	9	10	11	12	13	14	15
2,72	4,00	2,68	4,93	3,91	2,72	3,76	3,95
1,57	3,61	1,85	3,76	2,50	1,51	2,07	2,56
-	6.07	2.27	6.29	8.68	-	-	8.44
-	4.18	1.35	4.27	7.47	-	-	7.47

The first group of mean and std of locating errors are from this approach, and the second group are from the method in [Szeptycki *et al.* 2009]. Both tests are done on the same testing data set. When landmarking results are not available for the point, the symbol "-" is used.

landmark (d_i) from our method while the following rows are the results achieved by the curvature analysis-based method. The database that has been used is FRGC v1.0. The table is indexed by the landmark number referring to the legend in Fig. 2.9. As we can see, mean locating errors of all landmarks are less than 5mm. Notice that, as mentioned previously, except the nose tip, the mean and std of locating errors from our method are smaller than the ones from the curvature analysis-based method.

Table 2.2 shows the experimental results on 1400 face models randomly selected from FRGC v2.0 dataset. Recall that our statistical model was trained on selected face models in FRGC v1.0 only having controlled illumination and neutral expression while the 1400 face models randomly selected from FRGC v2.0 dataset display facial expressions and drastic illumination changes. As we can see from the table, the mean error in locating all landmarks only increases by 1mm compared with the experimental results on face models from FRGC v1.0 dataset.

The time for localizing landmarks on a face used by our algorithm (coded in Matlab) varies from 18min to 25min on a desktop PC with Intel Pentium4 1.8GHz and 1 Go RAM. Two steps are time consuming: firstly, it takes over 1500 iterations for the simplex algorithm to reach the convergence, which is more robust to local

Table 2.2: Mean and deviation of locating errors for all landmarks using FRGC v2.0 (mm)

	1	2	3	4	5	6	7
Mean	5.22	4.36	4.07	3.24	3.78	4.97	4.21
Std	3.14	2.21	2.32	1.67	1.91	2.83	2.55
8	9	10	11	12	13	14	15
3.10	6.65	4.88	6.95	5.38	3.53	6.48	4.67
1.64	4.50	2.52	4.24	3.14	1.86	3.16	2.99

minimum but is slower compared to other optimization algorithm. Secondly, the interpolation on local regions includes point sampling and interpolating. The more density the point clouds of a face are, the more time the process takes. Although it takes less than 500ms in each iteration, the overall time which it accumulates is quite noticeable due to the large number of iteration.

2.3.2.3 Discussion

Compared to others 3D landmarking algorithms mainly based on a prior knowledge of facial geometry information [D'House *et al.* 2007] [Lu & Jain 2006] [Faltemier *et al.* 2008] [Xua *et al.* 2006] [Colbry *et al.* 2005] [Szeptycki *et al.* 2009], our learning based method enables locating a higher number of landmarks (15 points to 3-9 points) while keeping a better location precision for all landmarks.

To further evaluate the location precision of our learning-based method, we have also driven experiments to analyse the location precision on manually labelled landmarks. Thus, we have asked 11 people to label the 15 landmarks on 10 faces, and have computed the mean and std of locating errors. The result is given in Table 2.3. As we can see, our statistical model applied to face models from FRGC v1.0 dataset (Table 2.3.2.2) have close results as compared with the ones achieved by human operators both in mean value and std. Moreover, these two results are correlated with each other, as points 4 and 10 in both tables hold higher accuracy while point 9 and 1 in both tables hold lower accuracy. This can be explained by the fact that our approach makes use of manual landmarks during the training process, which thus may be affected by errors in manual landmarking.

Table 2.3: Mean and deviation of locating errors for individual manually labeled landmarks(mm)

	1	2	3	4	5	6	7
Mean	2,95	2,42	2,03	1,94	2,04	2,76	2,11
Std	1,48	1,05	1,38	0,85	1,07	1,58	1,64
8	9	10	11	12	13	14	15
1,84	3,80	1,90	4,50	1,98	1,99	3,04	2,06
0,81	1,98	1,04	2,15	1,10	1,19	1,53	1,31

There exist three major sources of errors in our experiments. Firstly, our method requires an exact match between texture images and range ones. Although several badly mismatched face scans have already been filtered out in FRGC v1.0, there are still many face scans containing mismatches to a certain extent, especially in FRGC v2.0. Secondly, the training set should contain the major variations of faces, so that our learnt statistical model can synthesize instances as close as possible to the testing faces, further leading to a better locating accuracy. In our last test, variation of illumination and expression are not learnt in training. Last, as shown in Table 2.3, manual labelling also leads to locating errors of landmarks which implies a divergence for the global minimum of the objective function during the fitting process. Thus, our approach could be improved with a better learning of the model by using a training set containing more face variations, especially in expression and lighting condition, and with higher manual landmarking accuracy.

2.3.3 Conclusion

We have presented in this section a learning-based statistical method for automatic 2.5D face landmarking. The proposed statistical model learns from a training set both variations of global face shape as well as the local ones in terms of scale-free texture and range patches around each landmark. The fitting of the model to a new face is considered as an optimization problem according to shape parameters, with an optimization function describing the similarity between the input face and synthesized instances. Experiments have shown that our method has the ability to locate a high number of landmarks with a high accuracy. Using the model learnt

from half of the faces available from FRGC v1.0 dataset and experimented on 1400 faces randomly selected from FRGC v2.0 with uncontrolled illumination and facial expressions, our method has reached an average of locating errors less than 7mm for all 15 landmarks.

This approach is dedicated to 2.5D face landmarking. However, when the full 3D face information is available, this method can be improved by considering the 3D morphology as the global landmark configuration instead of the 2D shape on texture maps of 3D faces. This is the purpose of the method we propose in next section.

2.4 A 3D face landmarking method

In this section, we propose a general learning-based framework for 3D face landmarking. Our approach relies on a statistical model, called 3D Statistical Facial feAture Model (SFAM), which learns both global variations in 3D face morphology and local ones around each 3D face landmark in local texture and geometry. Different from the approach presented in the previous section, this method is a full 3D method which uses 3D morphology as the global landmark configuration instead of 2D shape on texture maps of 3D faces. Moreover, the fitting algorithm is based on the computation of correlation meshes between local features and their instances, which is more efficient. In this approach, in order to deal with the facial expression problem, the model is learnt from 3D faces with expressions. Moreover, in order to handle the occlusion problem, the detection of occluded faces and the identification of the occlusion type are performed. This classification provides the occlusion state for every local region around a landmark which allows the fitting algorithm to be applicable even on partially occluded faces.

2.4.1 Statistical facial feature model

2.4.1.1 Preprocessing the training faces

In order to train a SFAM, the targeted anthropometric landmarks need to be manually labelled for each aligned frontal 3D face. Contrary to most of the existing

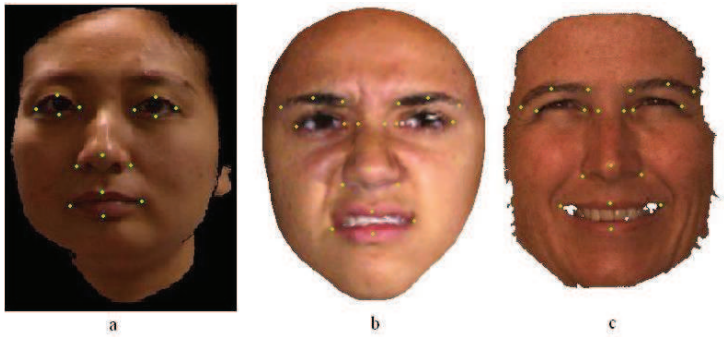


Figure 2.11: Two sets of landmarks are manually labelled on FRGC (a), BU-3DFE (b) and Bosphorus (c) datasets. Landmark set in (a) contains 15 landmarks, including nose tip and corners, inner and outer eye corners, mouth corners; landmark sets in (b) and (c) contain 19 landmarks, including corners and middles of eyebrows, inner and outer eye corners, nose saddles, nose tip and corners, left and right mouth corners and middles of upper and lower lips.

3D face landmarking algorithms, the set of our targeted landmarks can be easily changed provided a learning dataset. Through a statistical learning process, the local properties around landmarks along with their morphological relationships in training faces can be encoded independently of their locations and their number. To prove this, we have manually labelled two sets of landmarks on three different datasets, namely FRGC, BU-3DFE and Bosphorus datasets, as illustrated in Fig. 2.11. The landmark set for FRGC dataset is the same as the one described in the previous section.

The local regions around labelled landmarks are remeshed according to a principle similar to the one used for the creation of scale free local patches presented in subsection 2.3.1.2. Thus, points in local regions are first sampled and then interpolated on the uniform grids with the resolution of 1mm.

2.4.1.2 Modeling the configuration relationships of the landmarks as well as their local geometry and texture properties

Once a training 3D face has been preprocessed, 3D coordinates of all the landmarks, called 3D morphology, are concatenated into a vector S, describing the configuration relationships among local regions.

$$S = (x_1, y_1, z_1, x_2, y_2, z_2, ..., x_N, y_N, z_N)^T \tag{2.15}$$

where N is the number of landmarks, e.g. 15 or 19 in our work.

Two vectors G and Z are further generated, interpolated from local meshes as in eq. (2.16) and (2.17) similarly to G, Z in the method presented in the previous section. All Z vectors thus contain variations of local geometric shapes around landmarks while G vectors describe local texture properties. Alternatively, other local feature descriptors may also be computed from interpolated local 3D meshes and used, such as HK curvature, shape index, etc. for local shape, and Gabor jets, cornerness response, etc.

$$G = (g_1, g_2, ..., g_m)^T \tag{2.16}$$

$$Z = (z_1, z_2, ..., z_m)^T \tag{2.17}$$

Principal Component Analysis (PCA) is then applied to the three vector sets $\{S_i\}$, $\{G_i\}$, $\{Z_i\}$ and 95% of the variations in landmark configurations (morphology) as well as local texture and shape around each landmark are retained.

$$S = \bar{S} + P_s b_s \tag{2.18}$$

$$G = \bar{G} + P_g b_g \tag{2.19}$$

$$Z = \bar{Z} + P_z b_z \tag{2.20}$$

where $\bar{S}$, $\bar{G}$ and $\bar{Z}$ are respectively the mean landmark configuration, mean intensity and mean range value while P_s, P_g, P_z are respectively the three sets of corresponding variation modes. b_s, b_g, b_z are the sets of controlling parameters. All the individual parameters respectively in b_s, b_z and b_g are independent and follow Gaussian distributions with a zero mean and a standard deviation σ_i. The dimensions of P_s, P_g, P_z are respectively $(3 * N, n_s$, (m, n_g), $(m, n_z$, where n_s, n_g, n_z are the number of eigenvectors, N is the number of landmarks and m is the total number of points from all grids around the landmarks.

2.4.2 Locating landmarks

The SFAM based landmarking is performed through the optimization of an objective function which elegantly combines landmark configuration relationships with their local texture and shape features. The objective function is presented in section 2.4.2.1 whereas the fitting algorithm is given in section 2.4.2.3.

2.4.2.1 Objective function

Our objective function $f(b_s)$ is derived from a Bayesian approach by defining $f(b_s) = p(S|T, R, \psi)$, the probability to find the local texture and shape features at landmark configuration S given the 3D face with texture map T and range map R, as well as the learnt statistical model SFAM ψ given by eq.2.18, 2.19, and 2.20.

By following Bayes rule, we obtain:

$$\begin{aligned} p(S|T, R, \psi) &= p(T, R, S, \psi)/p(T, R, \psi) \\ &\propto p(T, R|S, \psi)p(S|\psi) \\ &\propto p(T|S, \psi)p(R|S, \psi)p(S|\psi) \end{aligned} \tag{2.21}$$

where $p(T|S, \psi)$, $p(R|S, \psi)$ are the probabilities of the face texture and range given a landmark configuration S and SFAM ψ. We assume the variable R and T from different face representations are independent within a local face region. $p(S|\psi)$ is the probability of a given landmark configuration estimated by SFAM.

Probabilities $p(T|S, \psi)$ and $p(R|S, \psi)$ can be estimated using a Gibbs-Boltzmann distribution as in eq. 2.22. This distribution has been widely used by PCA based statistical models in 2D, such as Constrained Local Model [Cristinacce & Cootes 2008], and proved to be efficient. This assumption is quite reasonable and results from the fact that the problem of 3D face landmarking is actually a Markov Random Field (MRF) which consists in assigning to each vertex of a 3D facial scan a label from a set of labels L. The set L encompasses all targeted landmarks (e.g., nose tip, eye corners) and a null value labeling any vertex which is not the location of any targeted landmark. Then, the theorem of the equivalence between MRFs and Gibbs distributions by Hammersley and Clifford [Li 2009] implies that the problem

of 3D face landmarking as described by the probabilities $p(T|S,\psi)$ and $p(R|S,\psi)$ are Gibbs-Boltzmann distributions [Duda *et al.* 2000].

$$\begin{aligned} p(S|T,R,\psi) &\propto \prod_{i=1}^{N} e^{-\alpha\eta_i} \prod_{i=1}^{N} e^{-\beta\gamma_i} \prod_{j=1}^{k} e^{-b_j^2/\lambda_j} \\ \log p(S|T,R,\psi) &\propto \sum_{i=1}^{N} -\alpha\eta_i + \sum_{i=1}^{N} -\beta\gamma_i - \sum_{j=1}^{k} \frac{b_j^2}{\lambda_j} \end{aligned} \tag{2.22}$$

where N is the number of local regions, η_i and γ_i are the similarities between instances and local regions, and α and β are weight factors. $p(S|\psi)$ can be considered as a penalty factor referred to [Cootes *et al.* 1995], where k is the number of landmark configuration or morphology modes, b_j is similar to b_s in eq. 2.18 and λ_j denotes the corresponding eigenvalues of the landmark configuration model.

We have extended the objective function to deal with face occlusion. Indeed, in the presence of occlusion, each local region around a landmark i will be associated with a probability of being uncovered m_i. The objective function is therefore rewritten as follows:

$$f(b_s) = m_i\alpha \sum_{i=1}^{N} F_{G^i}(s_i) + m_i\beta \sum_{i=1}^{N} F_{Z^i}(s_i) - \sum_{j=1}^{k} \frac{b_j^2}{\lambda_j} \tag{2.23}$$

where F_{G^i} and F_{Z^i} refer to eq. 2.26. m_i is the probability whether the region around the ith landmark is uncovered, thus being 0 if the local region is fully occluded and 1 if the local region is totally uncovered. s_i is landmark location from the morphology model.

The value of α and β can be determined by computing the ratio of $\sum_{i=1}^{N} F_{Gi}$ and $\sum_{j=1}^{k} \frac{b_j^2}{\lambda_j}$, $\sum_{i=1}^{N} F_{Zi}$ and $\sum_{j=1}^{k} \frac{b_j^2}{\lambda_j}$ separately when applied to verification faces with manually labelled landmarks.

In this work, we have made use of a simple occlusion classification algorithm which delivers a binary value for m_i which is 0 is if the local region is occluded and 1 if not.

2.4.2.2 Computation of the correlation meshes

In order to compute the F_{G^i} and F_{Z^i} factors in eq. 2.23, the correlation meshes are calculated in order to describe the similarity between instances and local face regions in both texture and shape modalities. It makes the optimization faster in the fitting process since those two factors can be directly obtained from the meshes instead of computing the objective function at each iteration.

The correlation meshes are illustrated in Fig. 2.12 and their computation is described as follows:

1. Given a new face scan, the closest point set S' to the landmark configuration S are computed on the face;

2. Texture and shape instances $\hat{G}$, $\hat{Z}$ are synthesized based on S':

   Based on points in S', we can obtain vectors G' and Z' through the local remeshing process as in the training phase. They are used to estimate b_g and b_z, as follows:

$$b_g = P_g^T(G' - \bar{g}) \tag{2.24}$$

$$b_z = P_z^T(Z' - \bar{z}) \tag{2.25}$$

   by limiting $b_i \in \{b_z, b_g\}$ to the range $\pm 3\sigma_i$ in order to constrain the possible deformations, any bi exceeding this boundary is replaced by its closest boundary. Then, texture and range instances $\hat{G}$ and $\hat{Z}$ can be generated according to eq. 2.19 and 2.20 using these constrained b_g and b_z.

3. Local regions around the points in S' are remeshed for both texture and range maps by using grids with a size of 51*51, as in the section 2.3.1.2;

4. For each local region i, a sliding window method is performed with the same size as the local grid size in SFAM (15*15). At each step j, a local range map Z and texture map G are extracted to compute the normalized correlation between them and $\hat{z}$, $\hat{g}$ respectively, which are the corresponding local parts

in $\hat{G}$ and $\hat{Z}$ (eq. 2.26). Then, the normalized correlations are set as the values of the window center on the corresponding meshes.

$$F_{G_j^i} = \left\langle \frac{g_j^i}{\left\| g_j^i \right\|}, \frac{\hat{g}_j}{\|\hat{g}_j\|} \right\rangle, F_{Z_j^i} = \left\langle \frac{z_j^i}{\left\| z_j^i \right\|}, \frac{\hat{z}_j}{\|\hat{z}_j\|} \right\rangle \tag{2.26}$$

$\langle \cdot, \cdot \rangle$ is the inner product and $\|\cdot\|$ is the L^2 norm.

2.4.2.3 Fitting algorithm

Before landmarking a 3D face through the fitting algorithm presented here, the occlusion algorithm described in section 2.4.3 is first applied to identify the occluded local regions and thus to set the corresponding m_i coefficient to zero. Therefore, only the unoccluded local regions will take part in the following fitting process. The algorithm works as follows:

1. Given a 3D face, its head pose is first compensated using ICP algorithm.

2. The morphology parameters b_s are optimized to minimize the distance between corresponding morphology instances and their closest points on the input face. The objective function is $f = \sum_{i=1}^{N} d_i$) where d_i is the 3D Euclidean distance between a point in a morphology instance and its closest points on the input face. i refers to the landmarks located on the rigid facial parts, such as those in the eyebrow, eye and nose regions.

3. The correlation meshes are computed as detailed in section 2.4.2.2, which is initialized by the optimized morphology from the step 2.

4. Morphology parameters are optimized to reach the maximum of the sum of values on two correlation meshes with the penalty factor. The objective function is $f(b_s)$ in eq.2.23 and its variables are b_s. Specifically, at each iteration, each point in morphology instances S^k generated from eq. 2.18 find its closest points on its associated local part in both correlation meshes of texture and shape. Correlation values from all local parts are summarized respectively and weighted as the sum of values on two correlation meshes (the first two factors

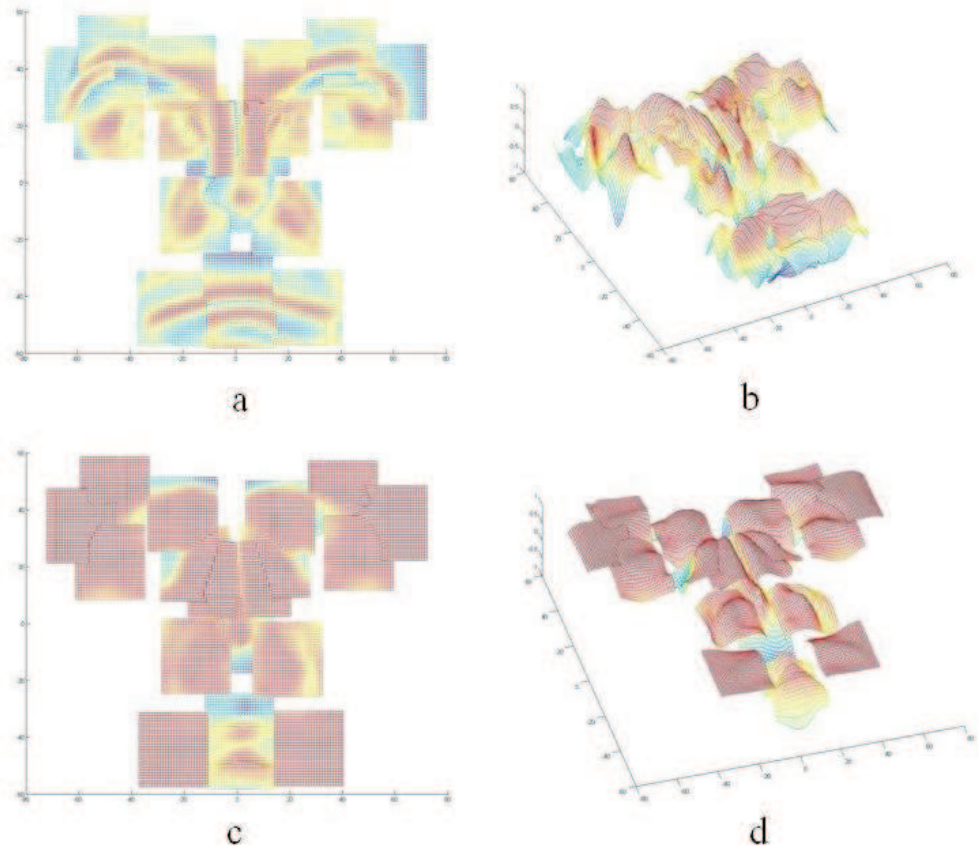


Figure 2.12: Correlation meshes from two viewpoints. Actually these meshes are in four dimension space, where the first three dimensions are x, y, z and the last one is correlation values. In these figures, we display the correlation values instead of z. (a) and (b) are the same correlation mesh from two point of views, describing the similarity of texture (intensity) instances from SFAM and texture (intensity) on the given face. (c) and (d) are the correlation mesh describing the similarity of shape (range) instances from SFAM and face shape (range). Red color corresponds to the high correlation and blue color corresponds to the low correlation.

in eq. 2.23). The penalty factor (the third factor in eq. 2.23) is computed accordingly.

The optimization in the step 2 and 4 is processed by Nelder-Meade simplex algorithm [Nelder & Mead 1965] because it is suitable for optimizing variables with dimension less than 30 and it is robustness to local minima. In each iteration in the simplex algorithm, the morphology parameters from the previous step are used for computing the value of objective function. Then, this value is compared with the function value in the previous step and simplex algorithm can thus predict a set of updated morphology parameters for the next step. The algorithm is converged when the morphology variation computed from morphology parameters between two consecutive steps is less than a threshold (5mm) or maximum iteration number is reached.

For partially occluded faces, occluded landmarks are excluded as well as their corresponding local meshes in the computation in the steps 2 and 4. Indeed, in wrong cases of occlusion classification, local non-face meshes lead the optimization to converge at a unpredictable point far from the desired minimum.

2.4.3 Occlusion detection and classification

Face analysis in the presence of partial occlusions, due to diverse factors such as hair, glasses, mustaches, scarf, etc. is a difficult problem. As far as 3D face landmarking is concerned, we are only interested in occlusions which may occur in local regions around landmarks. Thus, we have proposed a simple approach to classify occlusion type and give a set of binary values to local regions, corresponding to 'occluded' or 'unoccluded' states. Alternatively, we may have computed a probability associated with a local region being occluded or a measure indicating roughly how much a local region is occluded.

In order to perform occlusion detection, features from the range map are extracted since the presence of occlusion definitively changes the face shape in relevant local regions. Therefore, given an input face scan, its closest points s' to the mean landmark configuration (eq. 2.18) are computed. Then, 51*51 grids are used to

remesh local regions around these points only for range values, as in section 2.3.1.1.

For each local region i, a sliding window method is performed with the same size as the one of the local regions considered in SFAM. At each step j, a local depth map Z_α is computed and its local shape instance Z_β is calculated to further obtain a similarity map LS as follows:

$$b_{\alpha j} = P_{zi}^T (Z_{\alpha j} - \bar{z}_i) \tag{2.27}$$

$$Z_{\beta j} = \bar{z}_i + P_{zi} b_{\beta j} \tag{2.28}$$

$$LS_j^i = \left\langle \frac{Z_{\alpha j}}{\|Z_{\alpha j}\|}, \frac{Z_{\beta j}}{\|Z_{\beta j}\|} \right\rangle \tag{2.29}$$

P_{zi} is the submatrix composed of the rows in P_z associated with local region i. $\bar{z}_i$ is the subvector composed of rows in $\bar{z}$ also associated with local region i. $\langle \cdot, \cdot \rangle$ is the inner product and $\|\cdot\|$ is the L^2 norm. b_β is obtained by limiting b_α within a predefined boundary used to limit the possible deformations.

In case of occlusion, the local deformations are too large to be handled by the model. Thus, the instances Z_β generated from this model are quite different from the occluded local shape Z_α, which leads to a low similarity value in eq. 2.29. The LS_j^i describes the possibility to synthesize the local regions by learnt geometry variations. Therefore, this possibility decreases when a part of a face is occluded and thus contains non facial shape. This information is used for occlusion detection and classification.

Once LS has been computed for all points in a local region, a histogram from this similarity map is built. Then, histograms from all the local regions are further concatenated into a single feature, labelled with the occlusion type, such as occluded in the ocular region, occluded in the mouth region, occluded by glasses, or unoccluded. The distances between histograms are valued by the Euclidean distance, and the classification is performed by a simple K-NN classifier.

Since the available faces with occlusion in the dataset have certain patterns, as

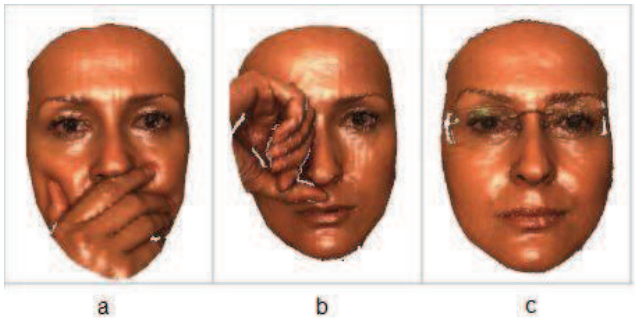


Figure 2.13: Different types of occlusion: a) occlusion in the mouth region, b) occlusion in the ocular region, c) occlusion caused by glasses.

shown in Fig. 2.13, we have preset a set of binary values indicating the occluded state in local regions for each type of occlusion. For example, for occlusion in the mouth region, the set of binary values {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0 } is used to initialize m_i, where the first two 0 correspond to the nose corners and last four 0 correspond to the mouth corners and lip middles. The classification leads to the list of local regions being occluded (m_i in eq. 2.23).

2.4.4 Experimentations

The SFAM model based framework for 3D face landmarking described so far has been experimented on three datasets, namely FRGC, BU-3DFE and Bosphorus datasets which are described in subsection 2.4.4.1 as well as the experimental setup. Then, the results are given in the following subsections.

2.4.4.1 Datasets and experimental setup

In order to test the soundness of our SFAM model-based 3D face landmarking framework, besides the Face Recognition Grand Challenge (FRGC) database used in section 2.3, we have made use of two other datasets, namely BU-3DFE database [Yin *et al.* 2006] and Bosphorus database [Savran *et al.* 2008].

The BU-3DFE database contains 100 subjects (56% female, 44% male). Each subject performs seven expressions in front of a 3D face scanner. Each of the six universal expressions (happiness, disgust, fear, angry, surprise and sadness) includes four levels of intensity. In our experiments, we have used the neutral faces and faces

with the 2 highest-level expressions from all subjects, thus 1300 face scans in total.

The Bosphorus dataset contains 4666 face scans from 105 subjects. This dataset contains not only many samples of the six universal facial expressions and many AUs, but also 3D face scans under realistic occlusions like glasses, hands around mouth and eye rubbing. Moreover, many male subjects have moustache and beard.

As illustrated in Fig. 2.11, 15 facial landmarks have been manually landmarked for the FRGC dataset and 19 for the BU-3DFE and Bosphorus datasets. They have been used as ground truth for learning the SFAM model and testing our landmark fitting algorithm. These three landmark sets contain some common landmarks, such as eye corners, mouth corners and contain landmarks from both face rigid and non-rigid regions.

2.4.4.2 Occlusion classification results

We have made use of the Bosphorus dataset including four kinds of occlusion, caused respectively by hair, glasses, hand near the mouth region and hand near the ocular region. An illustration of these types of occlusion can be found in Fig. 2.22. Occlusion caused by glasses occurs in front of two eyes, with changes mainly on local geometry. Occlusion caused by hand near the ocular region occurs generally in front of the right eye, with changes on both local geometry and local texture. As occlusions caused by hair generally do not occur on the landmark regions, this type of occlusion is excluded from our study. We consider for the experiments the other three types of occlusions and an unoccluded neutral face from each subject. We experimentally set K to five in the K-NN classifier and carried out a two-fold cross-validation. 347 face scans of 105 subjects have been used based on the data availability, where each subject contains at lease two scans out of four aforementioned types and at most four scans from each of the types. In each round, about faces scans from half of the subjects are used for training and the rest for testing. The subjects used in the training are different from those in the testing. After two round, scans from all subjects are used once for training and once for testing. The confusion matrix is given in table 2.4. As we can see, an average classification accuracy up to 93.8% can be achieved, which has been proved to be sufficient for further

Table 2.4: Confusion Matrix of occlusion classification

	Eye	Mouth	Glasses	Unoccluded
Eye	0.932	0.02	0.02	0.02
Mouth	0.01	0.97	0.02	0
Glasses	0.07	0.03	0.84	0.05
Unoccluded	0	0	0	1

'Eye' represents the occlusion caused by hand near ocular regions; 'Mouth' represents the occlusion caused by hand near the mouth regions; 'Glasses' represents the occlusion caused by glasses; 'Unoccluded' represents neutral faces without occlusion.

landmarking.

2.4.4.3 SFAM learning

We have made use of 452 face scans from FRGCv1 dataset to build our first SFAM model which learns local properties of 15 regions and their configuration relationships. The training face scans have limited illumination variations and do not contain facial expressions. Fig. 2.14 illustrates the first mode of configuration, local texture and local shape in the SFAM at their left and right boundaries (-3σ,3σ), namely -3std and +3std.

Moreover, we have used the face scans from 11 subjects in BU-3DFE dataset and the first 32 subjects in Bosphorus dataset to build our second and third SFAM respectively which capture global relationships and local properties from 19 landmarks. Every subject used for training has respectively 13 face scans in the case of the BU-3DFE dataset (a neutral face scan and 2 face scans for each of the six universal expressions in the intensity level 3 and 4), and 7 face scans including 6 basic expressions and the neutral one in the case of the Bosphorus dataset. Fig. 2.15, 2.16 and 2.17 illustrate the third SFAM learnt from Bosphorus dataset containing the first and second modes of configuration, local texture and local shape at their left and right boundaries (-3σ,3σ), namely -3std and +3std.

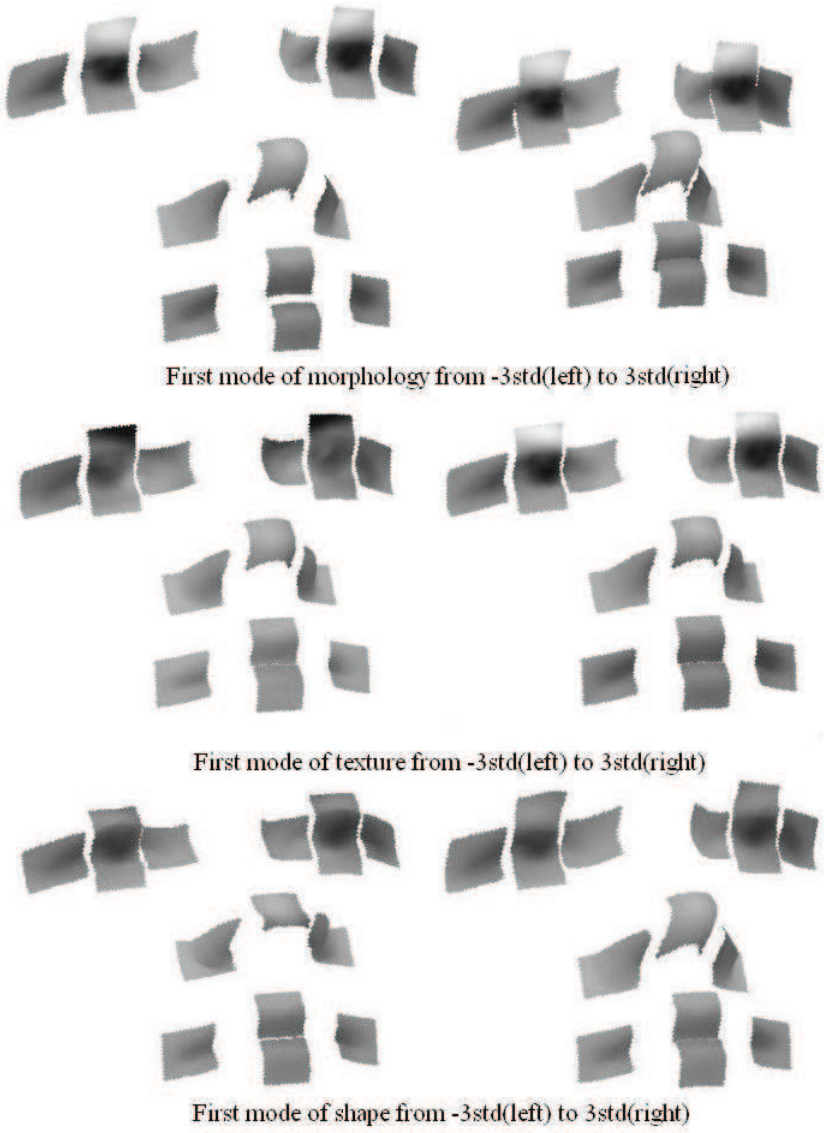


Figure 2.14: SFAM learnt from FRGCv1 dataset: first variation modes on the landmark configuration, local texture and local shape.First mode of morphology explains the landmark configuration variations in terms of face size; first mode of texture explains the intensity variation, especially in the eye region; first mode of shape explains the geometry variation in the upper part of face.

Figure 2.15: SFAM learnt from Bosphorus dataset: variations of the two first morphology modes. The first variation mode mostly explains the face morphology changes along the vertical direction, while the second variation mode explains the face morphology changes along the horizontal direction.

Figure 2.16: SFAM learnt from Bosphorus dataset: variations of the two first local texture modes. The first variation mode mostly explains the facial texture changes due to different skin color, while the second variation mode explains the facial texture changes in the eye and mouth regions.

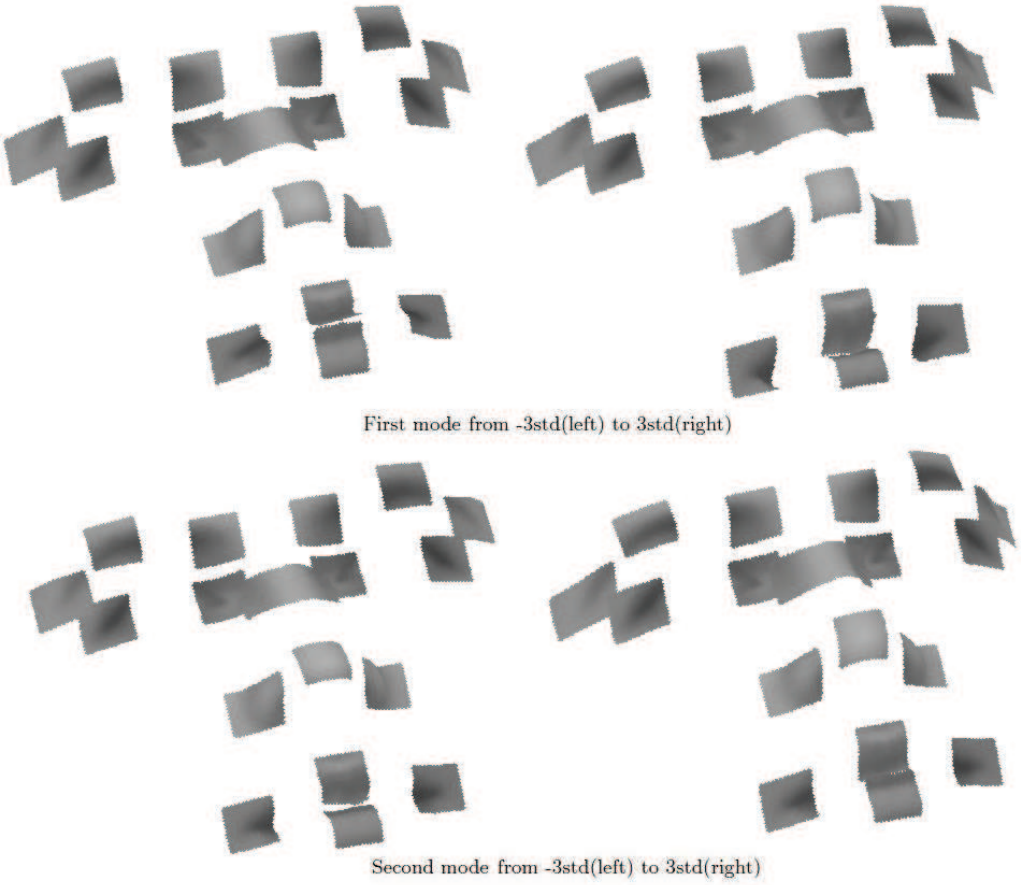


Figure 2.17: SFAM learnt from Bosphorus dataset: variations of the two first local geometry modes. The first variation mode mostly explains the face geometry changes in the lower part of face, while the second variation mode explains face geometry changes in the upper part of face.

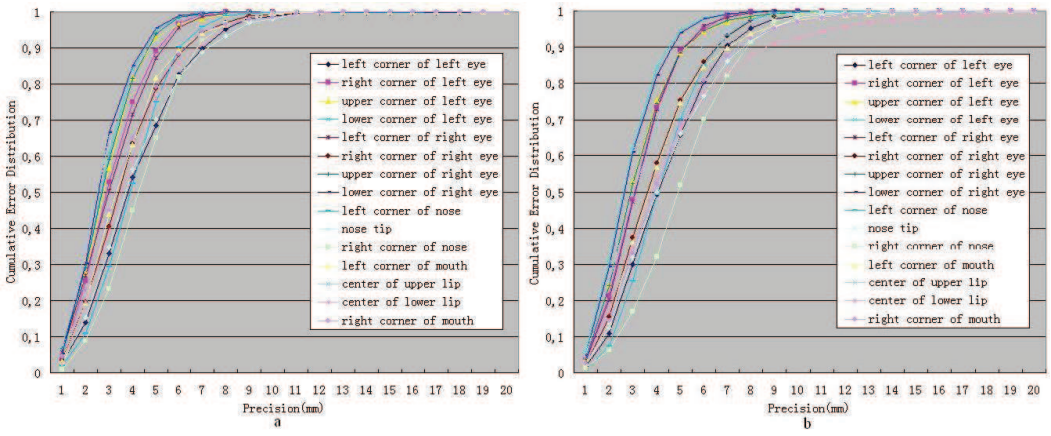


Figure 2.18: Cumulative error distribution of the precision for the 15 landmarks using FRGCv1 (a) and FRGCv2 (b).

Table 2.5: Mean error and standard deviation (mm) associated with each of the 15 landmarks on the FRGC dataset

	lcle	rcle	ucle	lwcle	lcre	rcre	ucre
Mean	4.17/4.31	3.07/3.21	2.92/3.17	2.76/2.75	3.15/3.24	3.67/3.89	2.84/3.18
Std	2.13/2.05	1.42/1.44	1.39/1.66	1.21/1.31	1.56/1.43	1.90/2.04	1.45/1.63
lwcre	lsn	nt	rsn	lcm	cul	cll	rcm
2.68/2.83	3.96/4.21	4.11/4.43	4.39/5.07	3.61/4.09	2.74/3.37	3.81/4.65	3.58/4.34
1.21/1.38	1.65/1.71	2.20/2.56	1.85/2.36	1.92/2.32	1.42/1.89	1.97/3.41	1.99/2.50

The index of the landmarks is the abbreviation of the legend in Fig. 2.18. The left number in each cell gives the result on FRGCv1 data while the right one the result on FRGCv2 data.

2.4.4.4 Results on landmarking

Using the learnt statistical models, the fitting algorithm for 3D face landmarking has been evaluated on 3 different experimental setups. In all these experiments, errors are calculated as the Euclidean distance between automatically located landmarks and the corresponding manual ones (ground truth). We do not set a general criterion or maximum allowed error to separate outliers in the following statistical results, which means almost all landmarking results are taken into consideration. A small number of landmarking results (around 20 face scans) which has mean errors over 20mm are excluded. The reason for this unreasonable error may be mostly due to the failure of ICP alignment.

Using the first SFAM, the fitting algorithm has first been experimented on the

Figure 2.19: Landmark locating examples from the FRGC dataset.

remaining roughly half FRGCv1 dataset not used for training, i.e. 462 face scans from subjects different from those in training. We have then tested SFAM on 1500 face scans randomly selected from the FRGCv2 dataset which contains illumination variations and facial expressions. Fig. 2.18 shows the cumulative distribution of the fitting accuracy for all 15 landmarks while Table 2.4.4.4 displays the mean and std of locating errors associated with each landmark. As we can see from the figure, most landmarks are automatically located within 9mm precision in both tests. Mean error and the corresponding standard deviation indicate that landmarks in the upper face region are located with better precision. A slight increase on mean error and the standard deviation in the second test is caused by uncontrolled illumination and facial expressions on tested face scans. Fig. 2.19 illustrates some landmark locating examples from these two experiments.

The third experiment has been carried out on the BU-3DFE dataset. Recall that 143 face scans from the first five male subjects and six female subjects have been used for training the second SFAM. 1157 face scans in total from the remaining 89 subjects are used for testing. Each testing subject has a neutral expression and six basic facial expressions with intensity level three and four. Fig 2.20 illustrates several locating examples with facial expression. Fig. 2.21 shows effect of expressions on landmarking accuracy. As we can see from this figure, landmarks with less deformation in expressions are better located, like eye corner, nose tip, nose corner. Mouth corners and the middle of lower lip are located with the worst precision

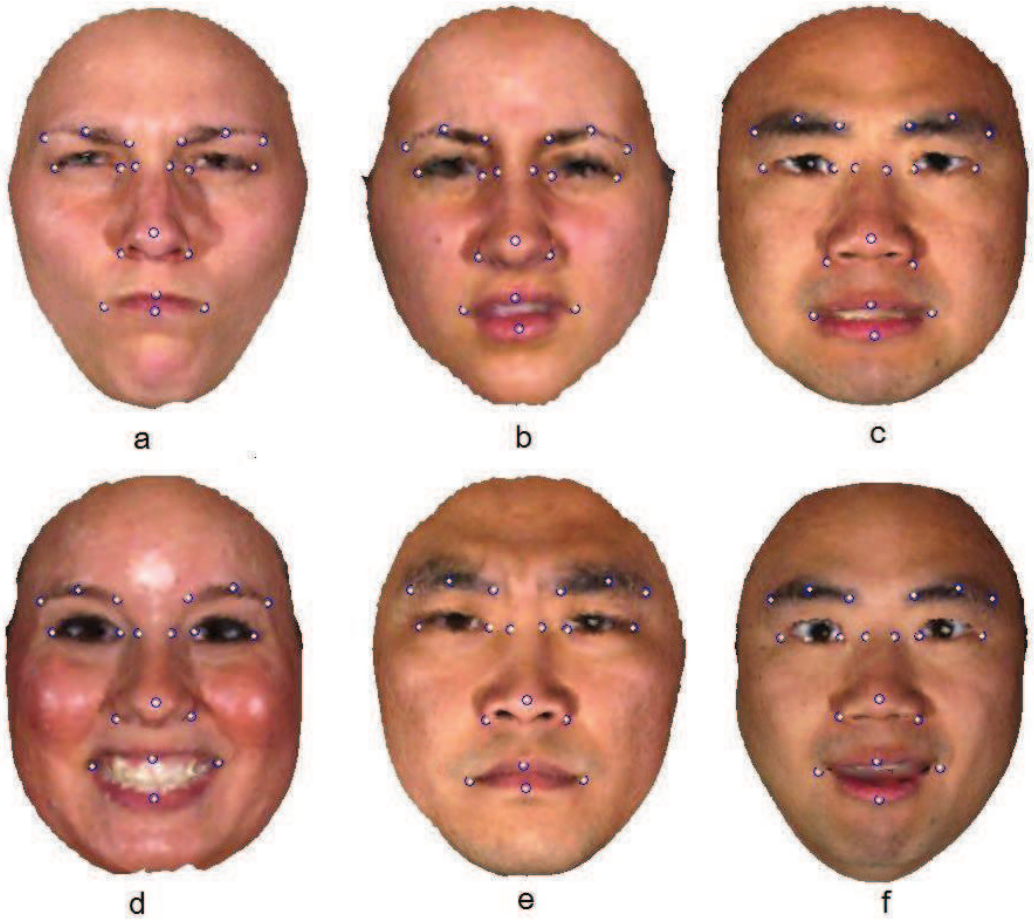


Figure 2.20: Landmarking examples from the BU-3DFE dataset with expressions of anger (a), disgust (b), fear (c), joy (d), sadness (e) and surprise (f).

Table 2.6: Mean error and the corresponding standard deviation (mm) of the 19 automatically located landmarks on the face scans, all expressions included, from the BU-3DFE dataset

	1	2	3	4	5	6	7	8	9
Mean	6.26	4.58	4.87	4.88	4.51	6.07	4.11	2.93	2.90
Std	3.72	2.82	2.99	2.97	2.77	3.35	1.89	1.40	1.36
10	11	12	13	14	15	16	17	18	19
4.07	3.30	3.27	3.32	4.04	3.62	7.15	4.19	7.52	8.82
2.00	1.70	1.56	1.94	1.99	1.91	4.64	2.34	4.75	7.12

The index of landmarks is as in Fig. 2.21.

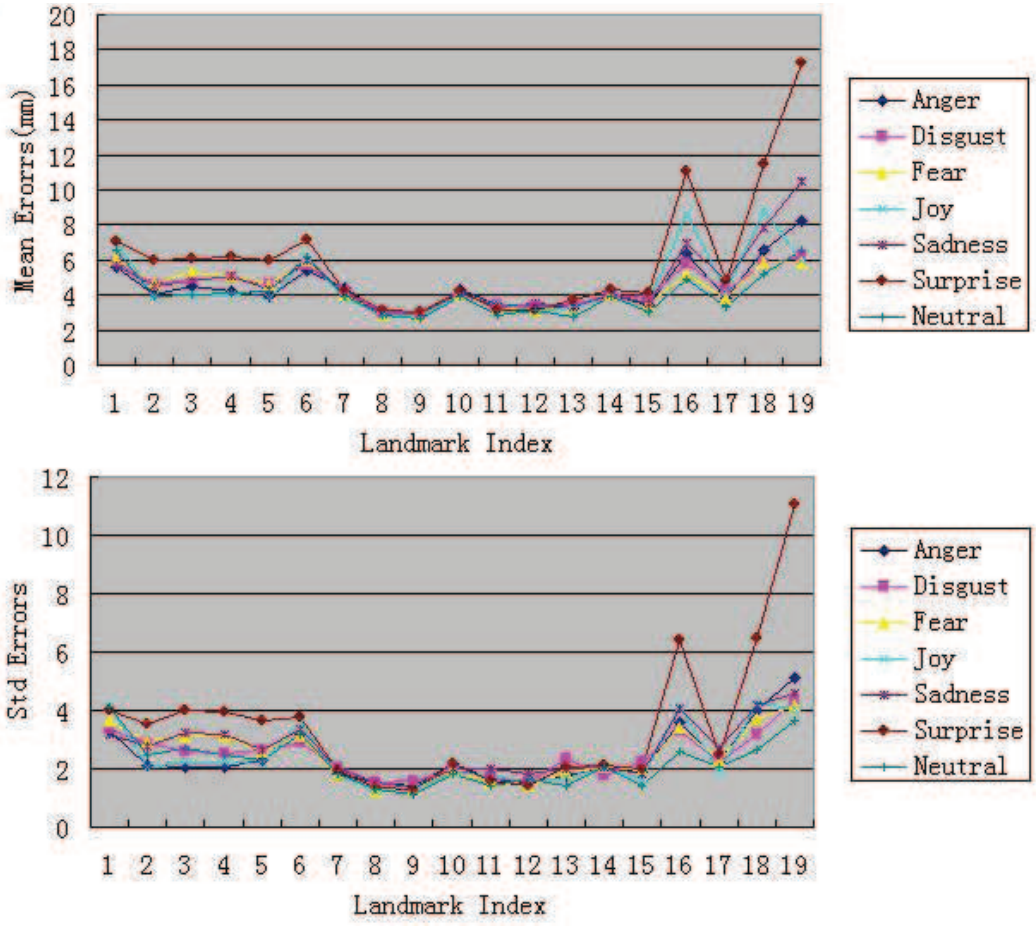


Figure 2.21: Landmarking accuracy on different expressions with the BU-3DFE dataset. (1: left corner of left eyebrow, 2: middle of left eyebrow, 3: right corner of left eyebrow, 4: left corner of right eyebrow, 5: middle of left eyebrow, 6: right corner of right eyebrow, 7: left corner of left eye, 8: right corner of left eye, 9: left corner of right eye, 10: right corner of right eye, 11: left nose saddle, 12: right nose saddle, 13: left corner of nose, 14: nose tip, 15: right corner of nose, 16: left corner of mouth, 17: middle of upper lip, 18: right corner of mouth, 19: middle of lower lip).

and the greatest standard deviation in face scans expressing surprise because of the significant deformation in this region induced by this emotional state. Table 2.6 summarizes the mean error along with the std of the landmarking algorithm with all expressions. The mean errors for all 19 landmarks stay within 10mm while most of standard deviations are lower than 5mm. The locating accuracy of landmarks in rigid face region is comparable to those of the corresponding landmarks located in FRGCv1.

The last experiment has tested the fitting algorithm using the third SFAM to locate 19 landmarks on 3D face scans under occlusion from the Bosphorus dataset. Fig. 2.22 illustrates several locating examples under occlusion. This experiment is carried out on 292 face scans from all the subjects different from the ones used for training in the Bosphorus dataset. In order to evaluate the efficiency of our proposed occlusion classifier for landmarking, the fitting algorithm is compared between the test with occlusion knowledge directly provided by the dataset and the test using occlusion knowledge from our proposed occlusion detection and classification algorithm (Table 2.7). The mean errors generally range from 6 to 11 mm, and more than 97% landmarks are located in 20mm precision in both configurations. Noting that this precision is considered as a criterion by some other works. Meanwhile, there exists an increase on mean error and std in average for the latter test, which is due to occlusion classification errors. However, these results remain acceptable, all the more since this automatic approach offers the ability to be generalized to datasets without occlusion information.

The time for localizing landmarks on a face used by this algorithm (coded in Matlab) varies from 10min to 16min on a desktop PC with Intel Pentium4 1.8GHz and 1 Go RAM. Similar to the previous algorithm, the simplex algorithm is used which is quite time consuming. The time consumed by the optimization in the step 2 is around 2 to 3 minutes. The computation of the correlation meshes saves time great, because it finishes the computation of the local interpolation once in the step 3 instead of computing them in each iteration as in the previous algorithm.

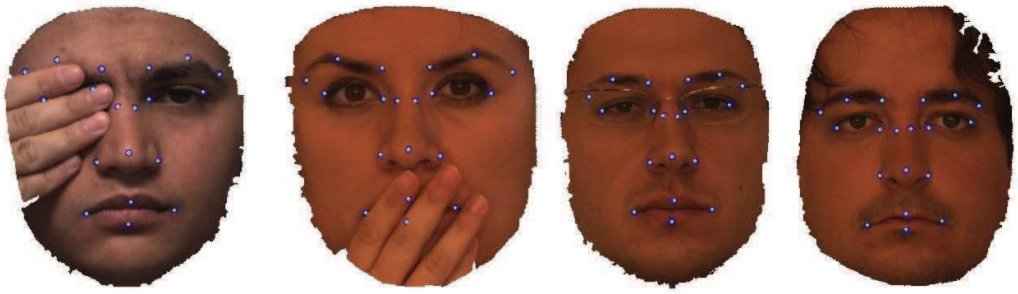


Figure 2.22: Landmarking examples from the Bosphorus dataset with occlusion. From left to right, faces are occluded in eye region, mouth region, by glasses and by hair.

Table 2.7: Mean error and the corresponding standard deviation (mm) associated with the each of the 19 automatically located landmarks on the face scans from the Bosphorus dataset under occlusion

	1	2	3	4	5	6
Mean	9.66/11.95	8.29/8.47	7.33/7.15	7.02/6.77	8.21/8.20	9.74/10.05
Std	6.08/8.85	3.92/4.39	3.41/3.36	3.23/3.38	4.27/4.45	5.23/6.08
	7	8	9	10	11	12
Mean	7.01/8.83	6.25/6.87	6.44/6.51	7.46/7.86	7.5/7.56	7.58/6.92
Std	3.77/6.37	3.42/4.21	3.08/3.58	3.56/4.73	3.60/3.88	3.63/4.02
13	14	15	16	17	18	19
6.35/7.19	8.46/8.39	8.03/7.79	7.96/9.75	8.67/9.01	8.21/9.65	10.41/10.61
3.11/2.99	3.64/3.64	3.31/3.36	4.18/6.28	4.84/4.93	4.25/4.97	5.37/5.61

The landmark indexes are as in Fig. 2.21. The left number in each cell represents the testing result using occlusion information provided by the dataset while the right one displays locating result using occlusion information provided by our occlusion detection and classification algorithm. In the latter tests, the knowledge of occlusion by hair (not considered by our occlusion detection and classification algorithm) was provided by the dataset

2.4.4.5 Failure cases and analysis

Fig. 2.23 illustrates several failure cases of landmarking under different conditions. The cases *a* and *b* are mainly due to the great deformation on the mouth region when face are displaying expressions. The morphology model in SFAM can not contain a specific mode for the deformation of a expression, however it generally learns variation modes from a mixture of expression and identity. Thus, when fitting SFAM on a face with great morphology deformation, like happiness and surprise, the fitting algorithm sometimes can not generate morphology instances which can approximate this extreme deformation. The cases *c* and *d* are mainly due to the information reduction in the fitting process when occlusion occurs. The occluded local parts are not considered in the fitting algorithm so that less part of correlation meshes are used in the objective function. Thus, the prediction of morphology parameters uses less information and is not as accurate and robust to local minimum as the prediction when no occlusion happens. Moreover, the missing values on occluded local correlation meshes introduce errors in the objective function as the weights α and β are determined when all local regions are considered.

2.4.4.6 Discussion

Compared to other 3D face landmarking algorithms in the literature, such as the ones in [D'House *et al.* 2007] [Lu & Jain 2006] [Faltemier *et al.* 2008] [Xua *et al.* 2006] [Colbry *et al.* 2005] [Jahanbin *et al.* 2008a] [Dibeklioglu *et al.* 2008], our SFAM-based approach is a general 3D landmarking framework which encodes the configuration relationships of the landmarks and their local properties in terms of texture and shape by a statistical learning instead of heuristic knowledge directly embedded within the algorithm. Our algorithm is thus more flexible and enables locating landmarks which are not necessary shape prominent or texture salient.

Most existing works on 3D face landmarking in the literature are only experimented on the FRGCv1 dataset. We can thus compare these results with the ones achieved in our first experiment described in the previous subsection.

Using the same dataset and a heuristic guided statistical method, Dibeklioglu et

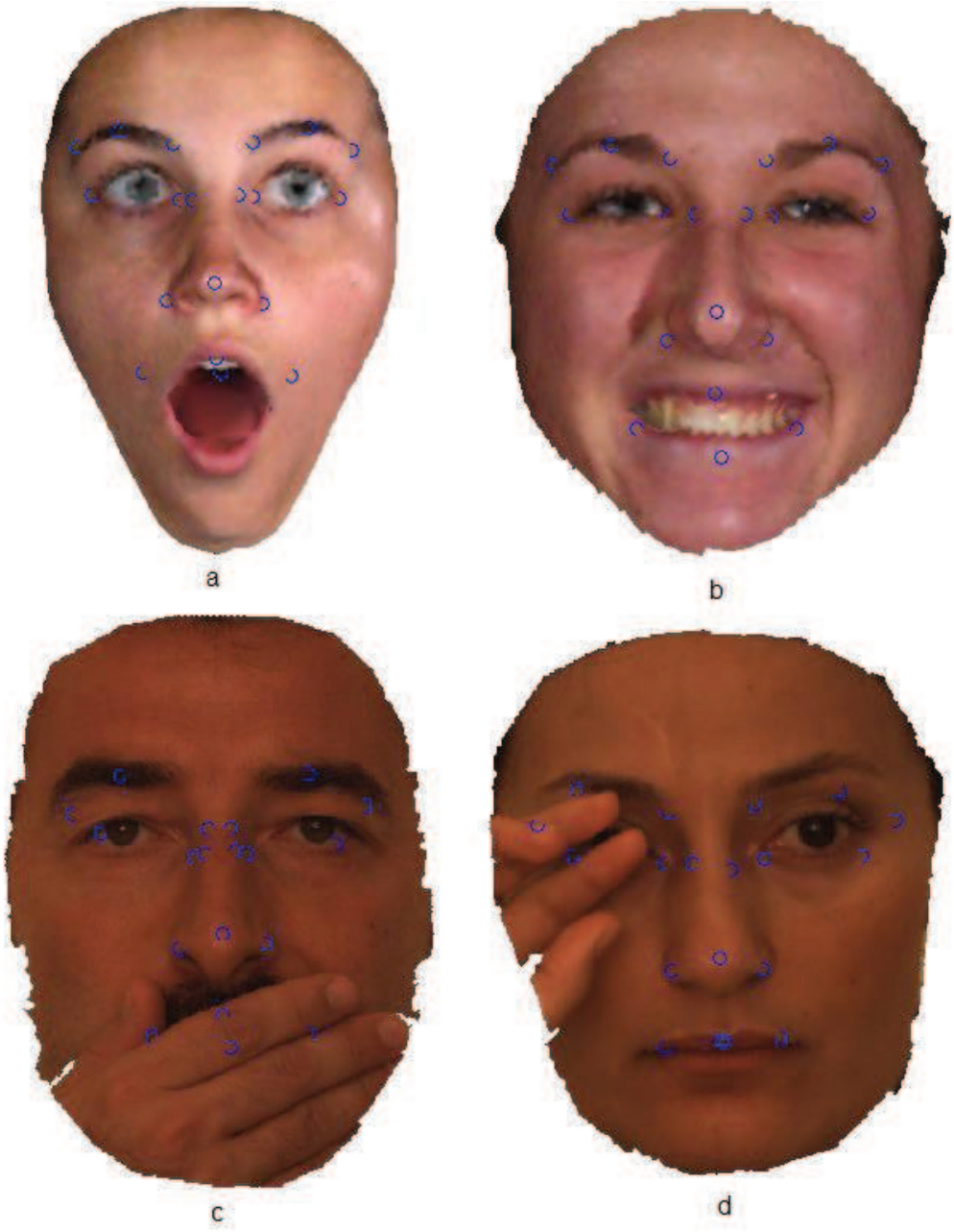


Figure 2.23: Some failure cases. a: failure case on face with surprise expression; b: failure case on face with happy expression; c: failure case on face with occlusion in mouth region; d: failure case on face with occlusion in eye region.

al. [Dibeklioglu *et al.* 2008] report an accuracy rate of around 99% for nose tips and inner eye, around 90% for the outer eyes and mouth corners within around 19mm precision (3 pixels precision on a reduced face texture with the reduction rate 8:1 in the paper. The 3D distance between pixels are 0.8mm to 1mm in FRGCv1). Compared to this result, our locating technique localize more landmarks (15 instead of 7) in better detection rates with the same precision.

In [Lu & Jain 2006], Lu et al. located seven landmarks, namely nose tips, corners of eyes and mouth on face scans from FRGCv1. The mean errors for these landmarks are around 6.0mm to 10mm, while our technique displays locating errors for 15 landmarks around 2mm to 5mm with much smaller standard derivation. Using the Bosphorus dataset with 3D face scans under occlusion, most mean errors of our landmarks range from 6mm to 10mm with a much lower standard deviation.

In [Koudelka *et al.* 2005], Koudelka et al. located five landmarks, namely inner corners of eyes, Sellion, nose tip and middle mouth with a mean error of 3.57mm over all the five landmarks and 97.22% of all the landmarks are correct detected with a precision of 10mm. In our case, we reach a mean error of 3.43mm over all the 15 landmarks and over 99% of all the landmarks are correct detected with a precision of 10mm.

Compared between our two methods, the average of error mean and std over 15 landmarks are 3.49mm and 2.34mm in the first method and those are 3.43mm and 1.68mm in the second method. All these results have a lower average error and better reliability compared with the curvature analysis based method [Szeptycki *et al.* 2009], 6.15mm and 4.05mm for seven landmarks included by 15 landmarks in our methods.

To the best of our knowledge, there exists only one work in the literature attempting to locate several landmarks on 3D face scans under facial expressions [Nair & Cavallaro 2009]. In their study, a 3D point distribution model is proposed to landmark five landmarks, namely the two inner eye corners, the two outer eye corners and the nose tip. Note that these landmarks are on face regions rather stable to expressions. Trained on 150 unnormalized face scans and tested on 2350 faces from the BU-3DFE dataset, their technique displays respectively a mean error

of 12.11mm, 11.89mm, 20.46mm, 19.38mm and 8.83mm for these five landmarks. Using the same dataset and a comparable quantity of training faces (143 faces), we display respectively a mean error of 4.11mm, 2.93mm, 2.90mm, 4.07mm, 4.04mm for these five landmarks and our technique also located other landmarks from mimic face region on 1157 face scans which produce the two higher levels of expression intensity out of the whole dataset.

We have also studied the reproducibility and the corresponding precision of manual landmarking for two reasons. First, manually labelled landmarks are used as the ground truth of automatic landmarks. Because of subjective variance, it is not necessary that manual landmarks are labelled at the precise location of landmarks. This imprecision may disturb the evaluation of the automatic landmarks. Thus, this study can provide a reference on the errors of manual landmarks used for evaluation. Secondly, this study can also give a reference on the variance of landmarking done by human and plays as a comparison with machine variance. For these purposes, 11 subjects are asked to manually label the 15 landmarks as shown in Table 2.3. The mean error of 15 manual landmarks is 2.49mm with the std at 1.34mm. For comparison, the second landmarking method achieves a mean error of 3.43mm with the corresponding standard deviation at 1.68mm on the same dataset.

2.4.5 Conclusion

We have presented in this section a general learning-based framework for 3D face landmarking which characterizes the configuration relationships between the landmarks as well as their local properties in terms of texture and shape, through a statistical model called SFAM. The fitting algorithm then locates the landmarks through the optimization of an objective function derived from a Bayesian approach. Such a framework is also quite suitable to deal with facial expressions and partial occlusions. Indeed, the consideration of both global and local properties helps to characterize landmarks deformed under expression and partial occlusion. Meanwhile, partial occlusion is taken into account in the objective function provided that occlusion probability around each landmark can be estimated. Based on this evidence, we have also introduced a 3D face occlusion detection and classification

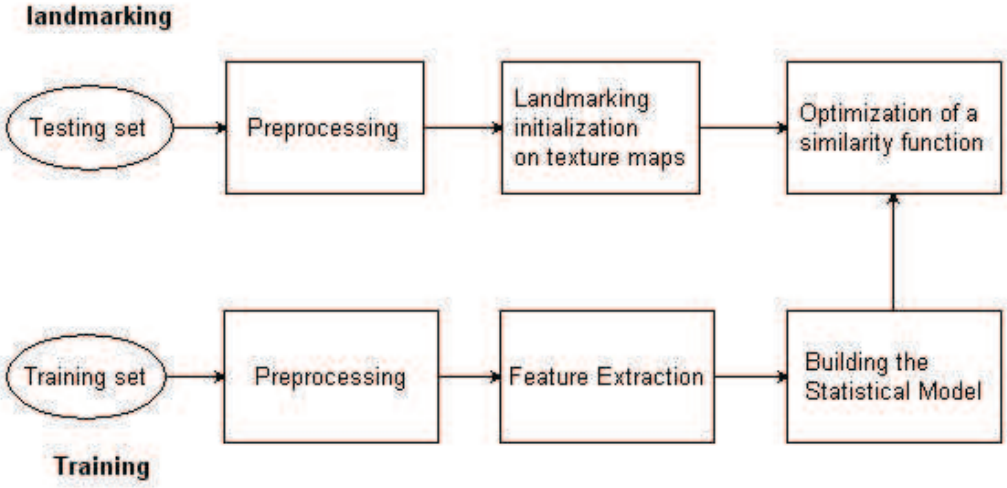


Figure 2.24: Flowchart of the first landmarking method.

algorithm which displays a 93% classification accuracy on the Bosphorus dataset. The detection is based on shape similarity between local range information of an input 3D face scan and the instances synthesized from SFAM. Experimented on FRGC datasets (v1 and v2) and BU-3DFE containing expressions and the Bosphorus dataset containing facial expressions along with partial occlusion, our 3D face locating technique has demonstrated its effectiveness.

2.5 Conclusion on 3D face landmarking

We have presented in this chapter two statistical model based methods for locating landmarks on 3D face scans. Both methods rely on statistical models by learning the variations in global landmark structure as well as local texture and range. However, the major difference between the methods are: firstly, the global landmark configuration in the first method is on 2D texture images while the SFAM is a full 3D statistical model; secondly, the fitting algorithm of the former is similar to active shape model in 2D while the second one introduces correlation meshes in the fitting; thirdly, combined with an occlusion detection algorithm, SFAM is able to perform landmarking on partial occluded faces. Flowcharts of these two methods

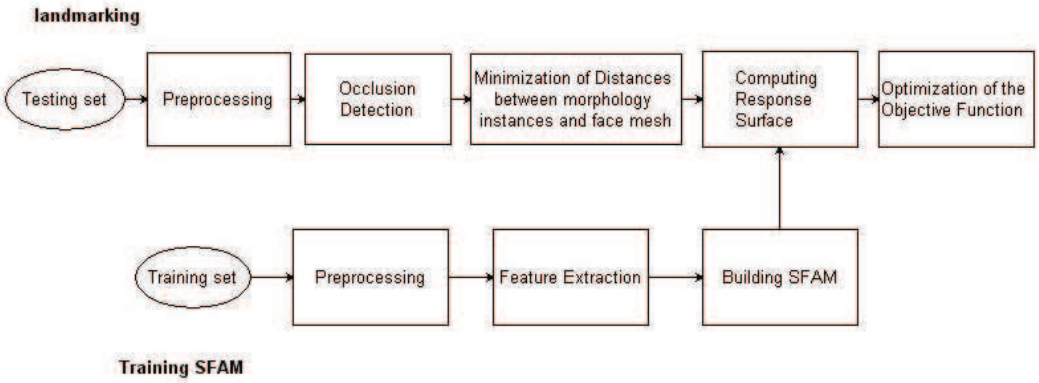


Figure 2.25: Flowchart of the second landmarking method.

are provided in fig 2.24 and fig 2.25 for a clear comparison. Experimental results have demonstrated that by considering both texture and geometry information, our methods is able to locate a set of landmarks beyond those characterized by salient shape with a better accuracy. Thus, SFAM has reached better landmarking accuracy than the previous models proposed in the literature in terms of accuracy and robustness when encountering severe conditions such as expression and occlusion.

In this chapter, only range and texture maps are used as simple descriptors of local shape and texture around landmarks. In the future, the landmark location may be improved by considering other descriptors such as HK curvature, shape index, etc. for shape feature or Local Binary Pattern, Gabor filtering, etc. for texture property.

Chapter 3

3D Facial Expression Recognition

3.1 Introduction

A facial expression communicates information about the characteristics of a person, a message about something internal to the subject, and results from one or more motions or positions of the face muscles. The source of facial expressions includes mental states, physiological activities and interpersonal communication, as shown in fig.3.1. Mental state or affect is one of the main sources, including felt emotions, conviction and cogitation. Physiological states such as pain, tiredness also influence unconscious face muscle activities appearing in forms of expressions. Verbal and non-verbal communications are other causes of facial expressions. [Fasel & Luettin 2003]

Facial expression analysis has interested researchers as early as Darwin in the nineteenth century, who had demonstrated the universality of human facial expressions [Darwin 1872]. In the past facial expression analysis was primarily a research subject for psychologists. In the 70s, few preliminary investigations on automatic facial expression analysis through images were presented [Suwa *et al.* 1978]. From 90s on, automatic facial expression analysis gained much interest due to advancements in related areas such as face detection, face tracking, etc, as well as the availability of more powerful computational facilities.

The recognition of facial expressions has various purposes and applications. It contributes to the development of human-centered human/computer (or robot) interfaces, which have the ability to detect user's affective behaviour and initialize proactive and socially appropriate behaviour during the communication process [Lisetti & Nasoz 2002]. It improves face recognition system by providing a prior knowledge on expression state allowing to overcome the difficulty caused by facial

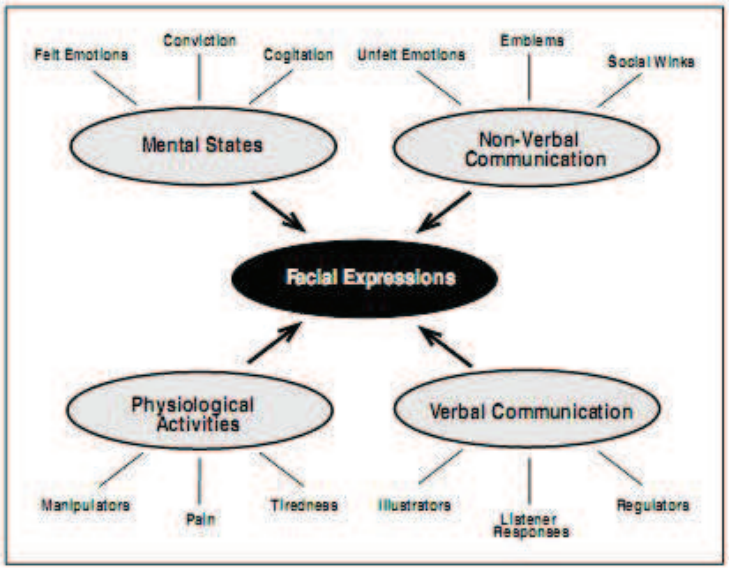


Figure 3.1: Sources of Facial Expressions[Fasel & Luettin 2003].

expressions [Mpiperis *et al.* 2008]. Other applications involve image understanding, psychological studies, tiredness detection, face image compression, face animation, robotics as well as virtual reality etc.

In this chapter, we focus on recognizing facial expression on 3D face scans. Two approaches are proposed to solve this task: first, we have elaborated a facial surface geometry based feature that is extracted and used to feed a Support Vector Machine (SVM) classifier for identifying the face expression; second, we have proposed to fuse the contribution of features from face landmark configuration, texture and geometry representations thanks to a Bayesian Belief Network (BBN) to recognize both universal expressions and facial action units. A fully automatic expression recognition system has also been proposed by combining the BBN with the SFAM presented in the previous chapter.

The reminder of this chapter is organized as follows. In section 3.2 we will firstly introduce the review of the state-of-art techniques dealing with facial expression recognition in both 2D and 3D environment. Then, we will present our first approach using the proposed geometry feature for facial expression recognition in section 3.4. Our BBN will be presented in the following section 3.5. Finally, we draw a conclusion in section 3.6.

3.2 The Problem

3.2.1 Theories of emotion

Most of studies on facial expression recognition aim at analysing human affect. Thus, it is important for designers of these systems to understand the structure and description of affect since it provides information about the affective classes to be detected. However, as there exists for the moment no consensus on the emotion taxonomy, the different theories of human affect have led to different streams of facial expression recognition approaches.

According to psychological theories of emotion, the emotion domain could be characterized by different qualitative states or dimensions. The two traditional theories that have most widely been considered in the past are discrete and dimensional theories of emotion.

Researchers working on the discrete theories propose that there exists a small number of basic or fundamental emotions, shown in table 3.1. Among them, the prototypical (universal) emotion categories and their corresponding facial expression have been proved to be perceived by humans in the same way regardless of culture. As a result, most of existing facial expression recognition approaches focus on recognizing these emotions. However, discrete theories fail to cover the whole range of emotions that people may experience in their everyday lives, and in particular subtle emotions and combination of emotions.

In the dimensional theories of emotion, emotional states are often mapped into a two or three-dimensional space. The two major dimensions consist in a valence dimension (pleasant - unpleasant, agreeable - disagreeable, also presented as appraisal dimension) and an activity dimension (active - passive, also presented as energy dimension or arousal dimension) [Greenwald *et al.* 1989]. If a third dimension is used, it often represents either power or control [Griffith]. Usually, several discrete emotion terms are mapped into the dimensional space according to their relationships to the dimensions, as shown in figure 3.2. Contrary to the discrete theories, a wider range of emotions and their relationships can be defined and described in the dimensional theories. However, it is difficult for facial expression recognition

Table 3.1: Some propositions for the definition of basic emotions [Ortony & Tumer 1990]

Researchers	Definition of basic emotion
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Amold	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman, Friesen and Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Frijda	Desire, happiness, interest, surprise, wonder, sorrow
Gray	Rage, terror, anxiety, joy
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
McDougall	Anger, disgust,elation, fear, subjection, tender, wonder
Mowere	Pain, pleasure
Oatley, Johnson-Laird	Anger, disgust,anxiety, happiness, sadness
Panksepp	Expectancy, fear, rage, panic
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	Fear, love, rage
Weiner and Graham	Happiness, sadness

systems to directly interpret face appearance into the emotional spaces.

3.2.2 Facial expression properties

Emotional states, like many other internal physiological activities (see fig. 3.1) are conveyed by facial expressions, which are generated by facial muscle contractions and result in temporal facial deformations in both facial geometry and/or texture. Facial activities not only cause wrinkles, bulges and other kinds of appearance deformation due to stretch and shrink on facial surface which produce variances on facial texture on captured face data, but also modify the facial geometry. Specifically, facial geometry here includes facial feature locations such as distance between landmarks (nose tip, inner and outer eye corners, mouth corners, ...) or feature point displacement, and geometrical shape of face surface. Because 2D cameras can not capture the 3D face information, face surface shape is seldom investigated by 2D approaches. Although facial muscle activities inherently change the facial appearance for the three face representations, including facial landmark configuration, texture and surface shape, the consequences on them are not necessarily displayed at the same level. For example, blinking eyes causes obvious variance on texture and

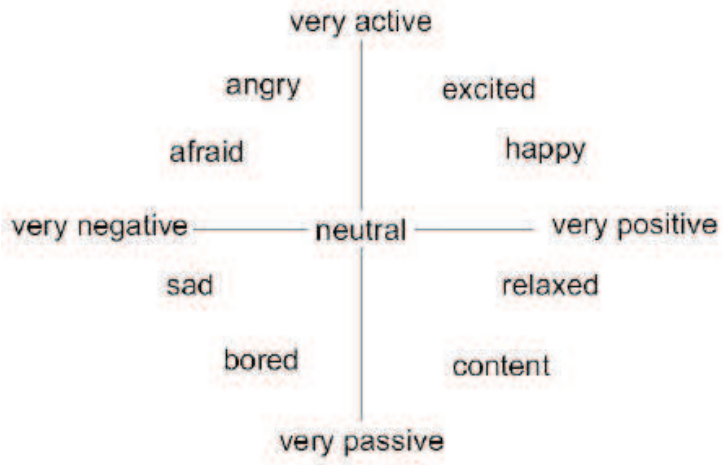
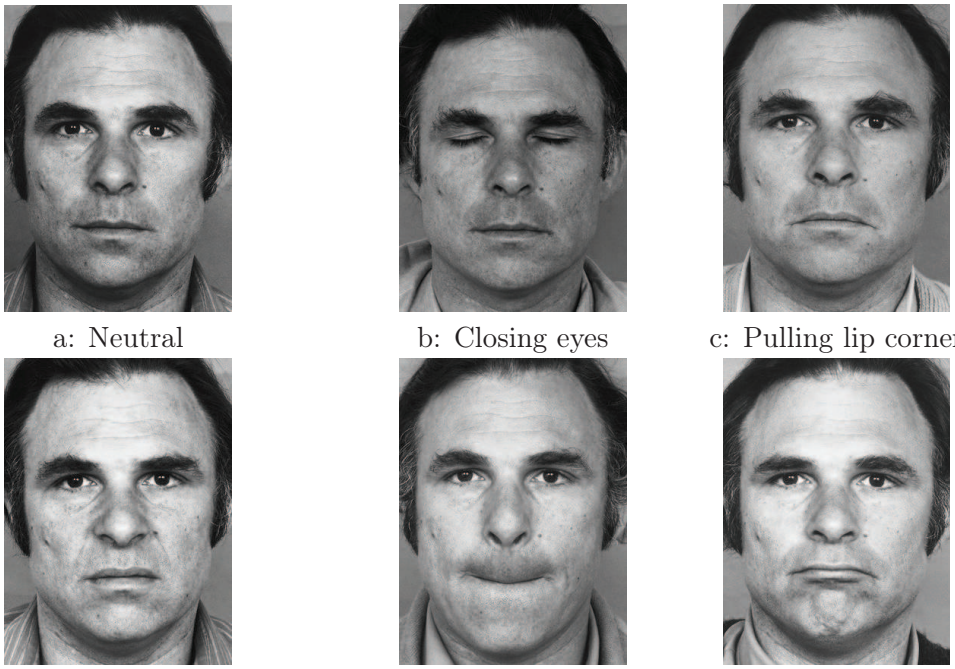


Figure 3.2: Example of emotions plotted into the arousal/valence plane [Wieczorkowska *et al.* 2005].

shape in the eye region without displacing eye corners, as shown in fig.3.3b; pulling a lip corner deforms local shape around mouth corners but causes subtle variance in texture from certain views, as shown in fig.3.3c. Meanwhile, it is difficult and challenging to detect certain facial activities using facial landmark configurations, such as deepening nasolabial furrow, sucking the lips inward, raising chins which are not apparent from movements of facial points but rather noticeable from variations in other two representations, as shown in fig.3.3d,e,f.

Not only the nature of facial deformation carries the message, but also the relative timing and temporal evolution of expression conveys an important meaning. It is suggested that the dynamics of facial expression provide an unique information about emotion that is not available in static images. [Schmidt & Cohn 2001] have shown that spontaneous smiles reach onsets faster than posed smiles and can have multiple rises of the mouth corners. Moreover, they are accompanied by other muscle activities that appear either simultaneously with mouth corner rises or follow them within 1s. Generally an expression process can be segmented into 4 steps: neutral, onset, apex and offset. The duration of typical muscle activities varies from 250ms to 5 seconds. Thus, using facial expression temporal dynamics are of importance for evaluating expression intensity level and categorizing facial expressions or muscle activities.



a: Neutral b: Closing eyes c: Pulling lip corner

d: Deepening nasolabial furrow e: Sucking the lips inward f: Raising chins

Figure 3.3: Facial activity examples [Ekman *et al.* 2002].

3.2.3 Facial expression interpretation

According to the two types of aforementioned emotion theories (discrete and dimensional), we can distinguish two main streams on analysis of the facial expressions: message-based approaches and sign-based approaches.

Message-based approaches are concerned with the message conveyed by facial expressions. They directly associate specific facial patterns with emotions and classify expressions into a predefined number of discrete categories. The most commonly used facial expression categories are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise), proposed by Ekman [Ekman & Friesen 1971].

Sign-based approaches aim at describing face deformation objectively rather than inferring meaning underlying the appearance. Facial muscle activities are hereby abstracted and coded by facial action units and then mapped into a variety of states in emotional space by high-level decision making. To completely describe all possible perceptible changes, the Facial Action Coding System (FACS) has been proposed [Ekman & Friesen 1978]. It is a comprehensive and anatomically based system used

to measure all visually discernible facial movements in terms of atomic facial actions called Action Units (AUs). Over 7000 different AU combinations have been observed and some of these combinations are mapped into basic emotions according to Emotional FACS (EMFACS) rules and various affective states according to FACS Affect Interpretation Database (FACSAID). For example, the combination of AU1, AU2, AU5 and AU26 can be interpreted as the surprise expression while the combination of AU6 and AU 12 is interpreted as happiness. With the assistance of a high-level making process, it is applicable to identify AUs for recognizing spontaneous facial expression in the dimensional space rather than classifying into the universal expressions. A detailed interpretation of AU combinations to emotions can be found in the appendix.

3.3 Related works

Over the past 20 years, facial expression recognition has gained growing interests within the computer vision community. Progress in recently five years has been observed in two main aspects. New methods are proposed to detect facial action units for recognizing more affect states as well as spontaneous expressions besides the six universal expressions. Meanwhile, many studies have begun to consider 3D faces for expression recognition.

3.3.1 Facial expression recognition: 2D vs 3D

Majority of facial expression recognizers have been developed in 2D environment partly because of the data availability. Indeed, most of the databases for facial expression analysis are made up of 2D images with nearly frontal faces displaying expressions. Other researches on spontaneous facial expressions are based on self-captured 2D video sequences. Another reason is that the applications such as HCI and robots are generally equipped with 2D cameras, which limits the type of face data.

Typical features used in 2D approaches are either geometric-based or appearance-based. Geometric features are extracted from contours of face com-

ponents and facial feature points, including shapes and positions of face components, as well as the location of facial feature points [Sohail & Bhattacharya 2007, Tai & Chung 2007, Chang *et al.* 2009a, Ari *et al.* 2008]. In the case of 2D videos, the position and shape of these components and/or landmarks are often detected in the first frame and then tracked throughout the sequence [Obaid *et al.* 2009, Gunes & Piccardi 2009, Brick *et al.* 2009]. The geometric features are easy to extract and quite efficient, however they ignore texture information reflecting facial texture variations and may not have enough discriminative power for identifying subtle expressions and action units. Appearance features such as Gabor wavelets, Haar features and Local Binary Pattern represent facial texture and transient variations due to wrinkles, bulges and furrows [Savran *et al.* 2010, Littlewort *et al.* 2006, Bartlett *et al.* 2006, Koutlas & Fotiadis 2008, Tong *et al.* 2010, Uddin *et al.* 2009, He *et al.* 2009, Yang *et al.* 2007, Zhao & Pietikainen 2007]. These features are very informative but they exclude global configuration of facial components and may be sensitive to illumination variations. Some studies adopt both global facial shape and features extracted from texture. The advantage is the mutually compensation on discrimination power for expression from both representations. Good examples of such a scheme are those in [Park *et al.* 2008, Park & Kim 2008, Mahoor *et al.* 2009] using Active Appearance Model (AAM) to capture the characteristics of the facial texture and the shape of facial expressions.

Generally these 2D recognizers face the challenges of illumination and head pose. The appearance of facial expressions varies with the viewpoint of an observer. Thus, head pose variation on face images includes the in-plane and out-of-plane rotations as well as the face scale. The in-plane rotation occurs around roll axis and can be rectified by face alignment. Face scale is generally normalized by interpolation or subsampling. However, it is difficult to handle the out-of-plane rotations due to the missing data caused by self-occlusion. Illumination has also a great influence on face appearance in images. It has been observed that the face image modifications caused by illumination changes can exceed the differences caused by expression and identity factors. Although some lighting models have been proposed, this problem is not yet completely solved, especially for expressions displayed on faces which are

partly lightened.

3D faces, which contain not only facial texture but also facial surface shape, are reputed to be insensitive to illumination and head pose. Since the geometry property of face can be computed, such as normal of vertices or curvatures, the Phong reflection mode is used to rectify the texture variance caused by different lighting conditions. Head pose can be simply normalized by multiplying the rotation matrix with each vertex and summarizing the translation vector. Because the unit of the 3D face coordinate system is *mm* instead of pixel in 2D, 3D face size does not vary with the distance between 3D scanner and subjects when taking the scan. Thus, recognizing facial expression in 3D offers the ability to handle illumination and head pose problems contrary to 2D-based approaches. Moreover, because subjects can be recorded with less controlled head pose using 3D scanner, spontaneous facial expression can be displayed on faces and analysed by 3D facial expression recognizers. [Savran *et al.* 2010] compares the effectiveness of 2D and 3D modality for detecting 25 AUs and demonstrates 3D modality generally performs better than 2D modality, especially for lower facial AUs and a fusion of both modality achieves the best performance. Specifically, Adaboost feature selection is applied on the Gabor magnitude responses for each AU on both 2D images and 2D conformal maps of 3D faces for comparison.

3.3.2 Facial expression recognition: static vs dynamic

Since a face in a static image can express an emotion, faces necessarily carry static emotion properties. Thus, a majority of studies in the literature dealing with facial expressions considers static images. However, a facial expression also implies a change of a visual pattern over time. This explains why more and more researchers attempt to characterize the dynamic evolution of expressions in order to improve the interpretation of facial activities [Hammal *et al.* 2007]. To do so, features representing the temporal dynamics of facial expression are extracted. The speed of a facial point displacement or the persistence of facial parameters over time can be extracted [Chakraborty *et al.* 2009, Brick *et al.* 2009] either for action phrase segmentation or recognition. In [Tong *et al.* 2007, Tong *et al.* 2010], the dynamic

Table 3.2: Facial Expression Recognition in the 2D environment requiring human intervention
Legend: exp - **S**pontaneous/**P**osed expression; class - **N**number of expressions or **AUs** (Action Units corresponding to AU detection); sub - number of subjects, person **D**ependent / **I**ndependent, JAFFE, CK, FABO, MMI are the database, OD - Other database; ? - missing entry; acc: Im / Vi - Image / Video-based.

References	Facial Feature	Classifier	Performance			
			exp	class	sub	acc (%)
[Vretos *et al.* 2009]	location of the Candide vertices	SVM	P	7	?, D, CK	Vi: 88.7
[Hu *et al.* 2008b]	Feature point displacement	LBNC; QBNC; Parzen; SVM	P	6	100, I, OD	Im: 86
[Tong *et al.* 2007]	Gabor wavelet	Ababoost and DBN	P, S	14 AUs	30, I, CK; 11, I, MMI; OD	Vi: 80.8 (CK)
[Tong *et al.* 2010]	Gabor wavelet	Ababoost and DBN	P,S	14 AUs	30, I, CK; 13, I, OD,	Vi: 85.8(CK)
[Uddin *et al.* 2009]	Texture	HMM	P	7	?,?I, CK	Vi?: 92.2
[Bai *et al.* 2009]	LBP and Gabor wavelet	LDA	P	6	10, I, JAFFE	Im: 92.4
[Zhi *et al.* 2009]	Image Intensity	GSNMF + KNN	P	6	?, I, CK	Im: 93.5
[He *et al.* 2009]	Gabor wavelet	HMM	P	7	10, ?, JAFFE	Im: 96.2
[Li *et al.* 2009]	SIFT, PHOG, Hist of edge	SVM	P	6	97, I, CK	Im: 96.3
[Zhi *et al.* 2008]	Image Intensity	FDP + KNN	P	6	?, D, CK	Im: 96.8

relationships between AUs are proved to be effective to enhance the recognition performance compared with the one directly derived from Gabor Wavelet.

Table 3.2 and 3.3 provide an overview of the main approaches for facial expression recognition from the year 2007 with respect to the facial features, classifiers, and performances. The methods based on static images are tagged by 'Im' in acc(%) columns and those based on videoes are tagged by 'Vi'. A detailed survey for systems before the year 2007 can be found in the Table 2 in [Zeng *et al.* 2009]. Other surveys for 2D facial expression recognition are proposed in [Pantic & Rothkrantz 2000, Fasel & Luettin 2003].

3.3.3 3D facial expression recognition

The number of studies dealing with 3D facial expression recognition has recently significantly increased in particular thanks to the publication of 3D facial expression databases. These databases are interesting since they allow researchers to develop and tune their approach, and then to compare their efficiency with the community. Currently, there exist three public databases which contain 3D face scans for facial expression analysis. The most widely used is the BU3DFE database [Yin *et al.* 2006] which contains face scans from 100 subjects displaying the six universal expressions

Table 3.3: Fully Automatic Facial Expression Recognition in the 2D environment
Legend: same as the legend in table 3.2.

References	Facial Feature	Classifier	Performance			
			exp	class	sub	acc (%)
[Obaid *et al.* 2009]	the deformation of tracked points	Rule-based	P	6	30, I, CK	Im:88.9
[Koutlas & Fotiadis 2008]	Gabor wavelets	NN	P	7	10, I, JAFFE	Im: 90.2
[Tai & Chung 2007]	Geometry property of lines between landmarks	NN	P	7	10, I, JAFFE	Im: 88.2
[Sohail & Bhattacharya 2007]	Distance	SVM	P	7	10, I, JAFFE; 30, I , CK	Im: 89.4; 84.8
[Chang *et al.* 2009a]	Distance	NN	S	4	6, D , OD	Vi: 95.0
[Park *et al.* 2008]	AAM	SVM	P	4	20, I, OD	Vi: 88.1
[Park & Kim 2008]	AAM	SVM	S	4	20, I, OD	Vi: 88.1
[Ari *et al.* 2008]	point displacement	SVM	S	7	11, D, OD	VI:90
[Song *et al.* 2009]	Image Ratio Features	SVM	P	7	?, ?, CK; ?, ?, JAFFE; ?,?, OD	Im: 88.9; 90.1; 87.0
[Mahoor *et al.* 2009]	AAM	SVM	S	2 AU	6, I ,OD	Vi: 82.5
[Martin *et al.* 2008]	AAM	Rule-based; MLP; SVM	S	7	18, D, FEEDTUM	Vi: 92
[Zeng *et al.* 2007]	local deformations of facial features	HMM	P	11	20, I, OD	Vi: 72.4
[Whitehill *et al.* 2008]	Gabor wavelet	SVM	S	12 AUs	8, D, OD	Vi: 75
[Gunes & Piccardi 2009]	Feature point and texture	HMM, rule-based	S	12	10, D, OD	Vi: 78
[Yang *et al.* 2007]	Haar features	Adaboost	P	8 AUs; 6 emotion	96, I, CK	Vi: 77.8(AU) 97.5(ex)
[Orozco *et al.* 2008]	Confidence on AAM parameter	KNN	P	7	30, I, FGnet MMI	Vi: 96.9
[Kim & Bien 2008]	Geometry property of lines between landmarks	LNN	P	7	?, D, CK,	Vi: 91.8
[Zhu *et al.* 2009]	SIFT	Adaboost	S	10 AUs	29, I, OD	Vi: 78.8
[Zhao & Pietikainen 2007]	LBP	K-NN	P	7	97, I, CK	Vi: 96.3
[Chakraborty *et al.* 2009]	Motion vector and fuzzy descriptor	Rule-based	S	6	50, ?, OD	Vi: 96.0
[Asthana *et al.* 2009]	AAM	SVM	P	7	30, I, CK	Vi: 94.3
[Hammal *et al.* 2007]	Geometry property of lines between landmarks	Transferable Belief Model	P	4	21, I, CK, OD	Vi: 77.8%
[Kotsia & Pitas 2007]	Geometric Deformation	SVM	P	8 AUs; 6	?, I, CK	Im: 84.7% (AU); 92.5 (EX)
[Brick *et al.* 2009]	Landmark displacement and velocity	SVM	P	16 AUs	100, I , CK	Im: 90.2
[Martins & Batista 2009]	Landmarks	Laplacian EigenMaps + HMM	P	7	4, I , OD	Vi: 75.7%
[Chang *et al.* 2009b]	Gabor wavelet	**HCRF**	P	15 AUs; 6	97, I , CK	Im: 92.9 (EX); 80.4 (AU)
[Shang & Chan 2009]	Landmark displacement	EM + HMM	P	6	100, I, CMU	Vi: 97.2
[Koelstra & Pantic 2008]	Orient histogram	GentleBoost + HMM	P	27 AUs	15, I, MMI	Vi: 65.1
[Sung & Kim 2008]	AAM	GDA	S	4	20, I , OD	Vi: 91.2
[Li *et al.* 2008]	Face texture	PCA + LDA	P	6	?, I , CK	Vi: 86.0
[Tsalakanidou & Malassiotis 2009]	Face deformation	Rule-based	S	10 AUs; 4	52, ?, OD	Vi: 82.5 (AU); 84.0
[Niese *et al.* 2008]	Geometry property of lines between landmarks	SVM	S	5	?, P, OD	Im: 97.2

as well as the neutral one. Each expression is displayed with 4 intensity levels from onset to apex. The BU4DFE database [Yin *et al.* 2008] contains 606 facial expression sequences in 3D captured from 101 subjects, with a total of approximately 60,600 frame models. For each subject, there are six model sequences showing six prototypic facial expressions respectively. This is the only database which contains 3D video sequences displaying facial expressions. Finally, the Bosphorus database [Savran *et al.* 2008] contains 105 subjects scanned with both the six universal expressions and facial action units. This is the only public database that contains dedicated scans displaying action units in 3D.

Existing approaches on expression recognition based on 3D faces can be divided into two categories: feature-based and model-based facial expression recognition. These approaches are further detailed in next subsections.

3.3.3.1 Feature-based 3D facial expression recognition approaches

Feature-based 3D facial expression recognition approaches rely on the extraction of facial features, which are further used to feed a classifier such as SVM, LDA etc. [Berretti *et al.* 2010, Tang & Huang 2008a, Soyel & Demirel 2008, Tang & Huang 2008b, Wang *et al.* 2006, Hu *et al.* 2008a]. Among them, features extracted from landmarks can discriminate the six universal expressions at a quite high recognition rate, over 94% [Tang & Huang 2008a]. In general, feature-based approaches rely on a set of precisely located landmarks, either for feature extraction [Tang & Huang 2008a, Soyel & Demirel 2008, Tang & Huang 2008b, Venkatesh *et al.* 2009] or for face segmentation [Wang *et al.* 2006, Hu *et al.* 2008a].

Similar to geometry-based features in 2D approaches, 3D geometry information is widely extracted for its easiness and efficiency. In [Venkatesh *et al.* 2009], the 3D location of 68 landmarks have been extracted around eyebrows, eyes, nose and mouth and used for classification. Distances among 3D landmarks are invariant to head pose and illumination and thus enable a rather robust recognition under different conditions. Soyel and Demirel [Soyel & Demirel 2008] have retrieved six distances between facial landmarks, describing the openness of eyes, height of eyebrows, openness of mouth, width of mouth, stretching of lip and openness of jaw. They achieve

a recognition rate of 87.9%. Such distance-like features have been further explored in [Tang & Huang 2008a], where less than 30 'best' features were automatically selected from candidate pool (all distances between 83 feature points). They achieve a recognition rate of 94.7% with a requirement of one neutral scan from each subject. Besides distance feature, Hao and Huang [Tang & Huang 2008b] have also extracted properties (the slope and length) of the line segments connecting 83 feature points, to make up 96 distinguishing features for recognizing the six universal facial expressions. They achieve a recognition rate of 87.1%. Landmark based features are easy to extract and invariant to head pose. However, its robustness to landmark precision has not yet been investigated. The recognition performance may highly rely on the landmark location accuracy which is difficult to achieve by automatic landmarking methods. Moreover, as a face contains information related to both person identity and expression, a normalization process is generally adopted to exclude the identity information that may disturb the expression recognition process.

Another kind of geometrical features is surface shape-based features, which are extracted from 3D face meshes and describe local shape properties. In [Wang *et al.* 2006], principal curvatures, surface principal directions and steepness of the surface have been calculated and further mapped into one of 12 primitive features on each vertex. Histograms of these primitive features from manual defined regions are extracted for classification. They achieve a recognition rate of 83.6%. However, 64 manually labeled landmarks are still required for defining the face regions. In [Savran & Sankur 2009], least squares conformal maps and elastic registration are used to map 3D faces into 2D images and register mapped faces into a reference one automatically. 22 AUs are detected by estimating the deformation between the registered face and the reference. The average of overall correct recognition rate is 91.4%.

3.3.3.2 Model-based 3D facial expression recognition approaches

Instead of directly extracting features, model based approaches make use of a generic face model, generally deformable, as an intermediate [Ramanathan *et al.* 2006, Mpiperis *et al.* 2008, Rosato *et al.* 2008, Venkatesh *et al.* 2009]. The expression

is recognized either from model parameters or features extracted from fitted models. In 2D, the Active Appearance Model (AAM) has been well explored for recognizing expressions. Usually, a set of model parameters are first obtained by fitting AAM on the target face. Then, they are used to extract distances among fitted parameters and training parameters in the parameter space for measuring the degree of similarity [Abboud *et al.* 2004] or to feed classifiers like SVM, MLP etc [Martin & Gross 2008]. In 3D environment, a number of generic face models has been proposed, such as annotated face model (AFM) [Kakadiaris *et al.* 2007], 3D morphable model [Blanz & Vetter 2003], bilinear model [Mpiperis *et al.* 2008]. They are widely applied in 3D face recognition [Kakadiaris *et al.* 2007, Blanz & Vetter 2003], 3D face reconstruction [Hu *et al.* 2004] and 3D facial expression recognition [Rosato *et al.* 2008, Mpiperis *et al.* 2008, Ramanathan *et al.* 2006]. These face models include a prior knowledge on 3D face, such as landmark location, face segmentation and deformation modes. As the transfer of the knowledge from the models to new faces naturally happens during the fitting process, it is promising to base an automatic expression recognition system on this category. Moreover, most of them contain deformation knowledge so that new faces can be registered by fitting the model into them via a model parameter optimization. The optimized parameters are further classified into expression classes [Mpiperis *et al.* 2008, Ramanathan *et al.* 2006]. However, fitting face models introduce errors since the learnt deformation modes can not be so comprehensive to synthesize every face precisely and perfectly without residual. Meanwhile, the model is built on limited features, usually on 3D face mesh [Mpiperis *et al.* 2008] and face texture [Ramanathan *et al.* 2006]. Using the model parameters from these two raw features may not have sufficient discriminant power for classifying various expressions.

Morphable Expression Model (MEM) [Ramanathan *et al.* 2006] is built by applying a Principal Components Analysis (PCA) on both 3D face shape and texture, whose process is similar to AAM. The morphable MEM learns expression variation modes from faces with expression. After fitting MEM into new faces by minimizing an energy function, model parameters are projected as a point into a low-dimensional

space made of the eigen-expressions for recognition. They achieved a recognition rate of 97% for 4 expressions in their own dataset. Similarly, 3D point distribution models [Venkatesh *et al.* 2009] are built for each expression by applying PCA. The difference between the coefficients of test faces and those of each expression from training are computed for recognition. They achieve a recognition rate of 81.7%.

Mpiperis [Mpiperis *et al.* 2008] has done a joint recognition based on decoupling identity and expression components of face appearance by bilinear models. A subdividable base mesh has been used to build point-to-point correspondence for building a symmetric and asymmetric bilinear models. Using these two models, identity and expression parameters for new faces have been optimized and further classified to perform face and expression recognition respectively. They achieve a recognition rate of 90.5%.

Instead of using model parameters, Rosato [Rosato *et al.* 2008] has extracted the distribution of primitive features from a generic 3D face model. The model has been fitted into new faces through 2D planer meshes, which are generated based on a circle pattern-based conformal mapping. They achieve the average recognition rate at 80.1% for recognizing seven prototypic facial expressions.

Because face models carry a prior information on landmark locations, extracting features on fitted model naturally do not need human intervention. Moreover, generic model may be built with a low resolution so that computation burden can be reduced. The limitation of this kind of approach is that a dense correspondence among face is necessary, which normally has high computation complexity.

3.3.4 Discussion

Overall, when comparing the state-of-the-art methods for recognizing 3D facial expression, one can observe that most of approaches:

- aimed at recognizing of a small number of universal expression (i.e., happiness, sadness, anger, fear, surprise, and disgust) using posed, controlled, static 3D face scans;
- are generally not fully automatic since they require human intervention for

locating landmarks or for fitting the initial model, due to the lack of reliable facial landmarking technology in 3D.

- are based on single feature or face model and thus do not contain sufficient information to describe a wide range of expressions or action units which are commonly investigated in 2D environment as a promising alternative solution for spontaneous expression recognition .

As it was mentioned in section 3.2.2, different facial actions deform different face representations to different level/extent. Thus, we are convinced that features from all face representations, including landmark location, facial texture and facial surface shape/geometry, should be extracted and combined in order to characterize a wide variety of facial expressions and action units comprehensively. In section 3.5, we will present a unified probabilistic framework for both expression and AU recognition problems, which aims at fusing the discriminative power of features from different facial representations. Combined with the automatic landmarking methods we proposed in the previous chapter, this framework is able to recognize facial expression efficiently in a fully automatic manner. Before this work, we will first propose in next section a local geometry based feature for describing face shape deformation caused by expressions that can be used to feed a classical classifier such as SVM for identifying the six universal expressions. Most of existing geometry based features are based on curvature computation, which are computation complex and sensitive to surface noise, or based on landmark configuration, which are easy to implement but exclude the rich surface shape information. So we propose here a feature which can be easy computed and effective in expression recognition.

3.4 3D Facial expression recognition based on a local geometry-based feature

Although features such as Gabor wavelet [Tong *et al.* 2010, He *et al.* 2009, Chang *et al.* 2009b] or Local Binary Patterns (LBP) [Zhao & Pietikainen 2007, Bai *et al.* 2009] have been widely used for recognizing facial expression or action

units in 2D environment, they can not carry information related to surface deformation occurring on faces in the real 3D world since they do not integrate shape information, and thus can not accurately reflect complex and authentic facial expressions. Moreover, the assumption of frontal images of faces under good illumination generally required by approaches in 2D is unrealistic in 3D. Therefore, there is a high demand to represent efficiently facial expressions in 3D.

Thus, several 3D geometry-based features have been proposed previously, such as HK curvature [Szeptycki *et al.* 2009], shape index [Dorai & Jain 1997] or primitive surface feature [Wang *et al.* 2006]. They are generally based on curvature computation from 3D face meshes [Szeptycki *et al.* 2009][Dorai & Jain 1997] or a fitted local surface patch [Wang *et al.* 2006]. They have the advantages of being pose free and efficient to describe the local surface property. In this section, we will propose a novel feature that derives directly from point clouds of 3D faces to describe the local shape property to enrich information for 3D face analysis, and particularly 3D facial expression recognition. Compared to the curvature based features, it is easy and fast to extract, but requires of a priori knowledge on head pose which is not needed by aforementioned features.

3.4.1 Brief introduction of popular 3D surface feature

Because of the explicitness of 3D visible surfaces, analysing faces thanks to their geometric shape should be easy and robust to illumination and head pose contrary to the intensity images conventionally used. In order to describe surface properties, principal curvatures are computed for extracting HK curvatures, shape index and primitive surface feature on a vertex. These curvatures (κ_1, κ_2) can be computed as the maximum and the minimum degrees of a surface bending around the vertex (see [Wang *et al.* 2006] for details) or as the extrema of the normal curvature function at the vertex ([Besl & Jain 1986]).

Shape index is a scale value computed as:

$$SI = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2}\right) \tag{3.1}$$

Note that the value of shape index is necessarily between 0 and 1.

H (mean), K (gaussian) curvatures are computed as:

$$H = \frac{(\kappa_1 + \kappa_2)}{2}, K = \kappa_1 \kappa_2 \tag{3.2}$$

The primitive surface feature derives from the principal curvatures κ_1, κ_2, the surface principal directions v_1, v_2 and the $\|\nabla z\|$ representing steepness of the surface around a vertex. Specifically, the local surface around a point is estimated by locally approximating it with a smooth polynomial function, $z(x, y) = \frac{1}{2}Ax^2 + Bxy + \frac{1}{2}Cy^2 + Dx^3 + Ex^2y + Fxy^2 + Gy^3$. The Weingarten matrix for the local surface is $W = \begin{bmatrix} A & B \\ B & C \end{bmatrix} = (\ \tilde{v}_1 \quad \tilde{v}_2\ ) \cdot diag(\ \lambda_1 \quad \lambda_2\ ) \cdot (\ \tilde{v}_1 \quad \tilde{v}_2\ )^T$, where λ_1, λ_2 are eigenvalues and $\tilde{v}_1$, $\tilde{v}_2$ are the orthogonal eigenvectors in local coordinate system. v_1, v_2 are further computed by rotating $\tilde{v}_1$, $\tilde{v}_2$ into the global coordinate system. The gradient magnitude $\|\nabla Z\|$ is computed from the smooth polynomial function. Two thresholds are defined, namely T_G and T_λ.

These features represent local surface characteristics and their categorization allows to label every vertex with one of the basic surface types. Using the value of shape index, five basic surface types can be characterized, as shown in fig. 3.4; the signs of mean and Gaussian curvature yield eight basic surface types, as shown in fig. 3.5; twelve primitive surface can be defined from the value of primitive surface feature using the rule shown in fig. 3.6. In general, if $\|\nabla z\| < T_G$ or there is a zero crossing in the direction of the maximum curvature, one of the non-hillside labels is assigned; otherwise, one of the hill-side labels is assigned using the rule defined in the table.

All of the aforementioned features are obtained by computing principal curvatures either on the discrete mesh of the underlying surface or on a fitted continuous local surface. However, this curvature approximation is sensitive to noise, such as spikes. Generally, such noises can be reduced by surface smoothing techniques or enlarging the size of neighborhood in computing curvatures [Szeptycki *et al.* 2009]. Consequently, either errors on face meshes may be introduced by the smoothing tech-

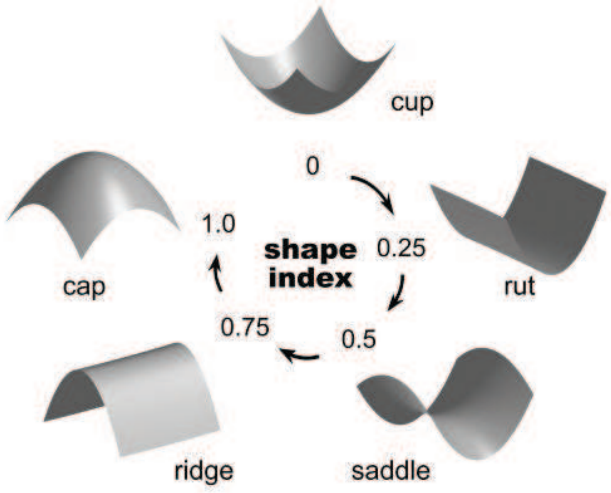


Figure 3.4: Five basic visible-invariant surface types defined by shape index [Yoshida *et al.* 2002].

niques or computational complexity increases. Moreover, the fitting of local surface and calculating principal curvatures are computational complex. In the following, we propose a surface characterizing method which relies on the point clouds instead of face meshes. This feature is relatively easy to implement and quick to compute, but requires to evaluate head pose. Therefore, a 3D pose estimation approach has also been proposed.

3.4.2 SGAND: a new Surface Geometry feAture from poiNt clouD

Instead of using curvature to characterize local surface, we directly sample the peripheral vertices of a vertex and characterize the vertex by comparing their geometrical relationship. We name this feature: Surface Geometry feAture from poiNt clouD (SGAND). From the fig. 3.4 and 3.5, we can observe that the basic surface types can be modeled by the geometrical relationship between the center part and the peripheral parts. For example, the center part of peek surfaces is higher than the peripheral parts while the center part of pit surfaces is lower; the center of saddle ridge surfaces is lower than some peripheral parts and higher than others. Here, we use the investigated vertex p to represent the center and eight clusters of vertices around to represent the peripheral parts. Their relationships are detected

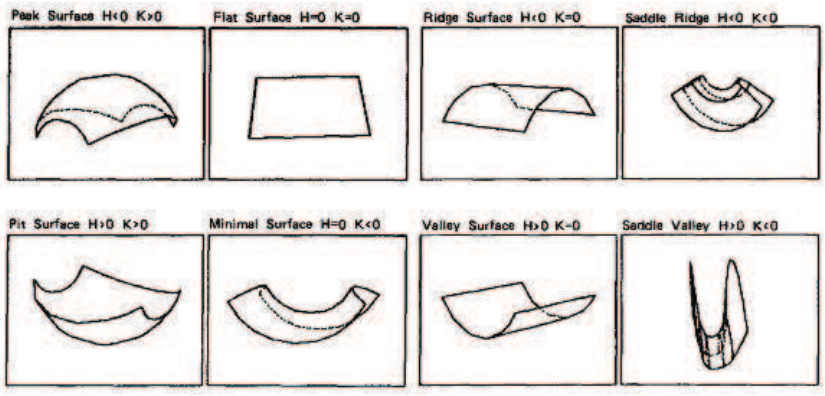


Figure 3.5: Eight basic visible-invariant surface types defined by HK curvatures [Besl & Jain 1986].

as shown in fig. 3.7. Specifically, we first find a plane M which is perpendicular to the z axis of the coordinate system and crosses the vertex p, as the red plane in the figure. Then, eight virtual cylinders C_{1-8} are placed perpendicular to the plane M and uniformly distributed on a circle S centered on p. The plane formed by the axis of the top cylinder C_1 and the axis of the bottom cylinder C_5 is parallel to the yz plane in the coordinate system. To obtain the geometrical relationship, vertices inside all cylinders are sampled and compared with the plane M. Taken C_1 in the picture c of the figure as an example, the sampled vertices inside of C_1 are grouped into two parts, those above the M plane and those below. A binary value is then set to this cylinder following the rule: if the number of vertices above is larger than the vertices below, we set 1; otherwise, we set 0. We repeat this process clockwise for all $C_i, i \in 1, 2, ..., 8$ and concatenate the binary array into a scalar value which is necessarily between 0 to 255 for the vertex p. Therefore, the peak surface can be featured as 0 and the pit surface as 255. Other surfaces as flat and minimal have multiple values as 7, 28, 112, 193 for flat and 34, 136 for minimal surface because of the in-plane rotation. Furthermore, SGAND allows to model not only these basic surface type but also other surfaces since totally 256 slots can be generated.

The radius of the cylinders C and the circle S are respectively defined as 2mm and 7mm in fig. 3.7. We fix the radius of C because local surfaces with a size smaller than 2mm can be considered as a flat surface. The radius of circle S can vary so that

λ_1	λ_2	Hillside Label	Non-Hillside Label
$\lvert\lambda_1\rvert<T_\lambda$	$\lvert\lambda_2\rvert<T_\lambda$	flat	slope hill
$\lambda_1<-T_\lambda$	$\lambda_2<-T_\lambda$	peak	convex hill
$\lambda_1<-T_\lambda$	$\lvert\lambda_2\rvert<T_\lambda$	ridge	convex hill
$\lambda_1<-T_\lambda$	$\lambda_2>T_\lambda$	ridge saddle	convex saddle hill
$\lambda_1>T_\lambda$	$\lambda_2<-T_\lambda$	ravine saddle	concave hill
$\lambda_1>T_\lambda$	$\lvert\lambda_2\rvert<T_\lambda$	ravine	convex hill
$\lambda_1>T_\lambda$	$\lambda_2>T_\lambda$	pit	
$\lambda_1>T_\lambda$	$\lambda_2<-T_\lambda$		concave saddle hill

if $\|\nabla z\| < T_g$ or there is a zero crossing in the direction of the maximum curvature, one of the non-hillside labels is assigned; otherwise, one of the hill-side labels is assigned

Figure 3.6: Classification rule of primitive 3D surface labels [Wang *et al.* 2006].

the vertex P can be featured by different peripheral parts on a surface and thus be more informative. fig. 3.8 illustrates the variation of the radius S. We compute and compare the quantity of sampled vertices within the same cylinder above and below the plane. Since these two set of vertices always have the same density, SGAND is invariant to face scale.

When the investigated vertex p and the main direction (the normal of plane M) are fixed, the feature varies with the radius of the circle S over which the cylinders C are distributed. Different vertices are sampled with varying radii and thus may influence the binary values. The implicit reason for changing the radius is the different geometrical properties of facial surfaces at different distances. For example, the feature for the nose tip is always 0 because it is the highest vertex on the face and this property does not change with the radius. However, the feature value of the inner corner of eyes definitively changes with the radius since the sampled regions move across the nose saddle region and thus cause variations on sampled surface property. fig.3.9 displays our feature extracted from one face scan with various radiuses of the circle S. Each color corresponds to a value in SGAND ranging from 0 to 255. We can see the SGAND distribution on faces and how this distribution is affect by radius of the S.

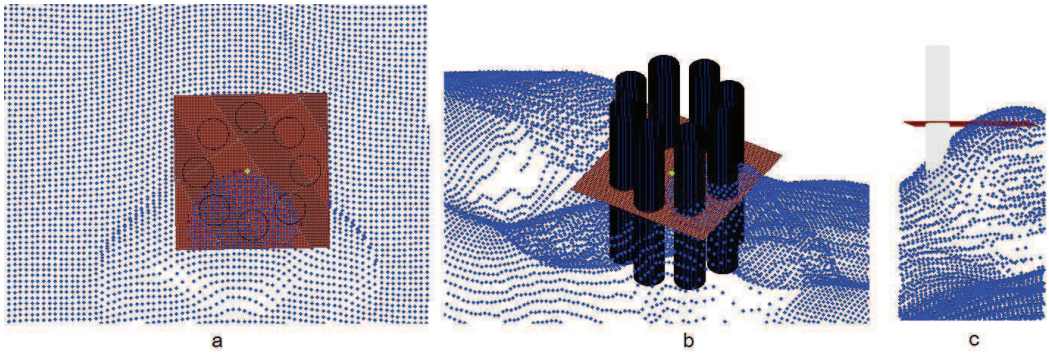


Figure 3.7: Extraction of our proposed feature. a: frontal view, b: side view, c: one cylinder for clearance. The green dot represents the investigated vertex which is located in the nasal region of a face. A plane and eight cylinders are involved as displayed.

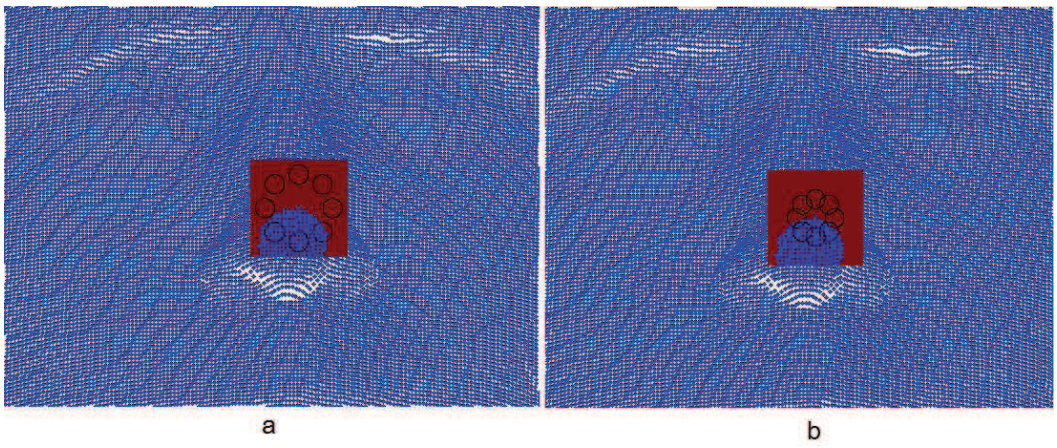


Figure 3.8: The radius variation of circle S. a: 7mm, b: 4mm.

Unlike the primitive surface feature which is extracted using a local coordinate system, we always extract SGAND using the direction vertical to the plane M in the 3D coordinate system. This is more intuitive and matches human perception habit since we look at faces through the gaze direction. Usually, the M plane is formed perpendicular to the z axes of the coordinate system for frontal 3D faces. If a face has an other head pose, we need to define another direction for the M plane instead of the z axes which varies with the head pose and indicates the frontal direction of the face. Thus, we propose an automatic approach to estimate head pose and find the direction in order to form the face planes for our feature extraction. This

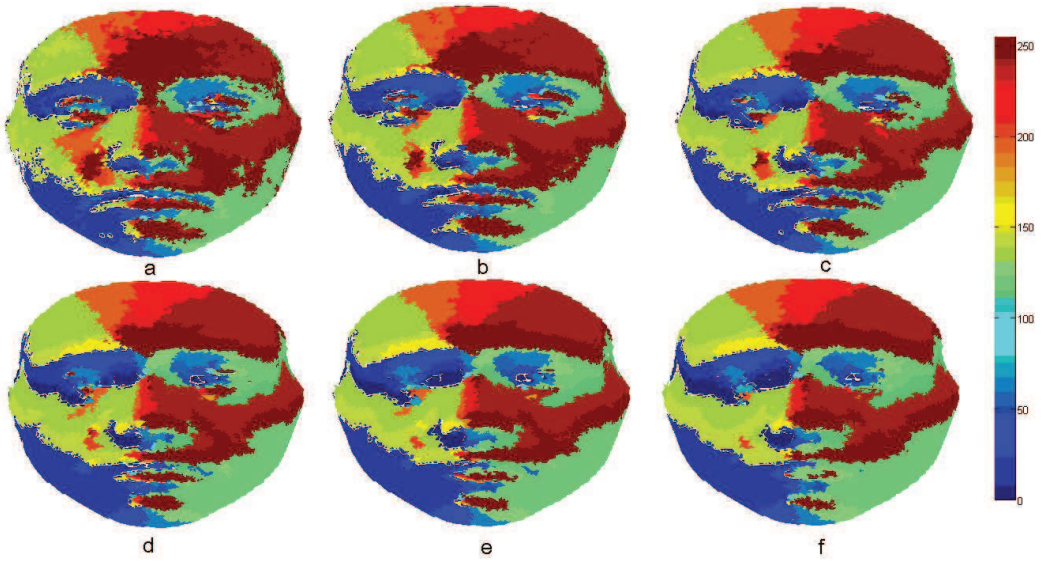


Figure 3.9: Influence of the radius of circle S on our feature extracted from a neutral face: a: 3mm, b: 5mm, c: 7mm, d: 9mm, e: 11mm, f: 13mm.

approach will be presented in subsection 3.4.3.

3.4.3 Pose estimation of 3D faces

Pose estimation of a 3D facial model aims at finding how the 3D face surface is embedded into the 3D coordinate system [Besl & Jain 1986]. A reliable pose estimation plays an important role in face alignment and feature extraction. For example, a plane is required vertical to the face frontal direction when extracting the SGAND. There are some existing methods that have been proposed both in 2D [Bailly *et al.* 2009] and in 3D. These approaches either use range data [Breitenstein *et al.* 2008], which is applied to 2.5D faces, or require a training process for a generic face model [Kinoshita *et al.* 2006]. Thus, we propose in the following a fast and efficient pose estimation approach which is based on face mesh and does not require any training process. Therefore, this method is suitable to be adopted as a preprocessing step in 3D face analysis systems.

The basic idea is to use the vertices on the frontal side of a face to generate a face plane by regression. With the normal of the plane and a direction from top to bottom of faces, head pose in 3D coordinate system can be estimated and thus the

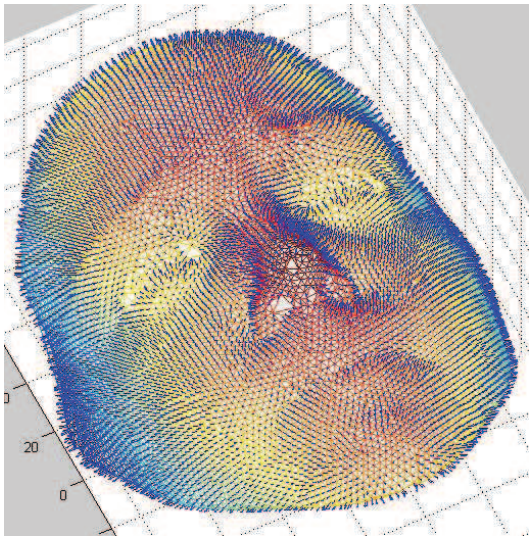


Figure 3.10: A face with vertex normals

roll, yaw and pitch directions can be computed.

In order to find the frontal vertices on a face, we take the normals of vertices into consideration because they are rotation invariant and represent the directions of vertices on facial surface. The normal of a vertex is computed by averaging the normals of surrounding triangle facets. fig. 3.10 illustrates a face with normals on all vertices.

In order to select vertices on the frontal side of a face, we use a clustering method to cluster those with normals pointing to the frontal direction as well as those whose normals point to the left and right direction. This idea is intuitive since a facial part of a skull can be roughly approximated by three planes from frontal, left and right side respectively. Thus, the clustering of normals can perfectly group vertices into three sets. There are several clustering methods conceivable such as Mixture of Gaussians, K-means, etc.. Among them, the K-Means algorithm has been chosen because of its efficiency and easiness to be implemented. Given normals of all vertices $(n_1, n_2, ..., n_N)$, where each normal is a 3D real vector, the K-means clustering partitions the N normals into 3 sets: $S = S_1, S_2, S_3$ by minimizing the within-cluster variance:

$$\arg\min_{s} \sum_{i=1}^{3} \sum_{x_j \in S_i} \|x_j - u_i\|^2 \tag{3.3}$$

where u_i is the mean of S_i. The K-means algorithm is enumerated as follows:

1. Place three points into the space represented by the normal data that are being clustered. These points represent initial centroid groups.

2. Assign each normal to the group that has the closest centroid.

3. When all normals have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids stabilization. This produces a separation of the normals into groups from which the metric to be minimized can be calculated.

We have compared the clustering results between K-means and Mixture of Gaussians, and display one example in fig. 3.11. We can observe that normals separated by K-means are more symmetrical than the results obtained from Mixture of Gaussians.

The clustering process outputs three normals u_1, u_2, u_3 which are the centroids of S_1, S_2, S_3 respectively. In order to distinguish the mean normal which represents the normals pointing to the front, we compute the inner products of each pair of u_1, u_2, u_3. Indeed, the angle between the left mean normals and the right mean normals is the biggest among all 3 angles formed by any pair of u_1, u_2, u_3. Thus, we can find the minimum inner product from the pair and the other centroid represents the frontal vertices.

After the group of frontal vertices have successfully been obtained, a Principal Component Analysis (PCA) is used to fit a linear regression that minimizes the perpendicular distances from the those points to a plane and a line. The process is as follows. The coefficients D_1, D_2 for the first two principal components define vectors that form a basis for the plane. The third principal component is orthogonal to the first two, and its coefficient D_3 defines the normal vector of the plane. The

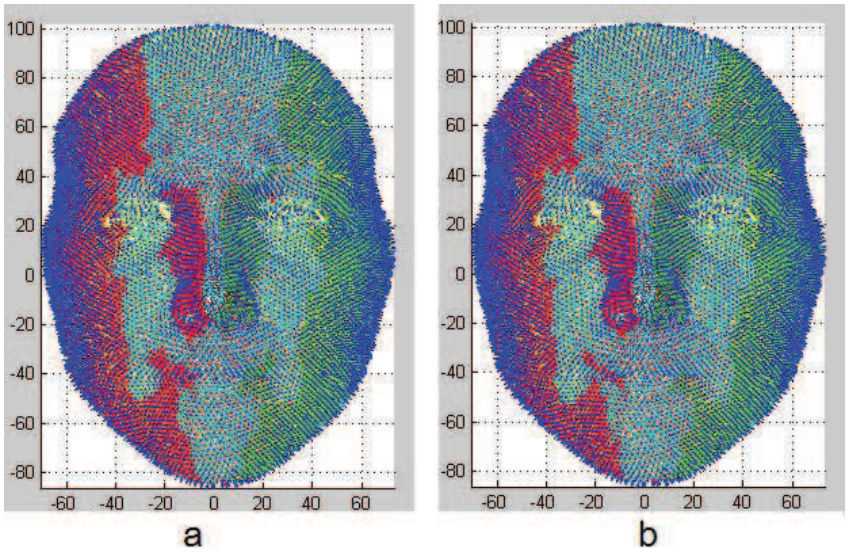


Figure 3.11: Separation of vertices into 3 sets: left (red), frontal (blue), right (green). a: K-means, b: Mixture of Gaussians

plane passes through the mean point P_m of the group. Meanwhile, the coefficient D_1 of the first principal component is the vertical direction of the face. Indeed, the first component explains the most prominent variance in the data which is the vertex location variance along the top-bottom direction as seen in blue points in fig. 3.11. The direction is the best 1-D linear approximation to the data. In summary, D_3 and P_m form the face plane while D_1 and P_m form the line.

The plane can also be fitted by using other methods on the group of frontal vertices, such as the least square method. Overall, the least square method is only able to approximate a face plane whose normal can be used in our feature extraction. Thus, our method offers the advantage of allowing the estimation not only of the face plane but of three head pose directions including yaw, pitch and roll.

3.4.4 3D expression description and classification based on SGAND

Our proposed SGAND has been designed for 3D face analysis including face detection, facial landmarking, face recognition and facial expression recognition. In this subsection, we propose to make use of it for 3D facial expression recognition.

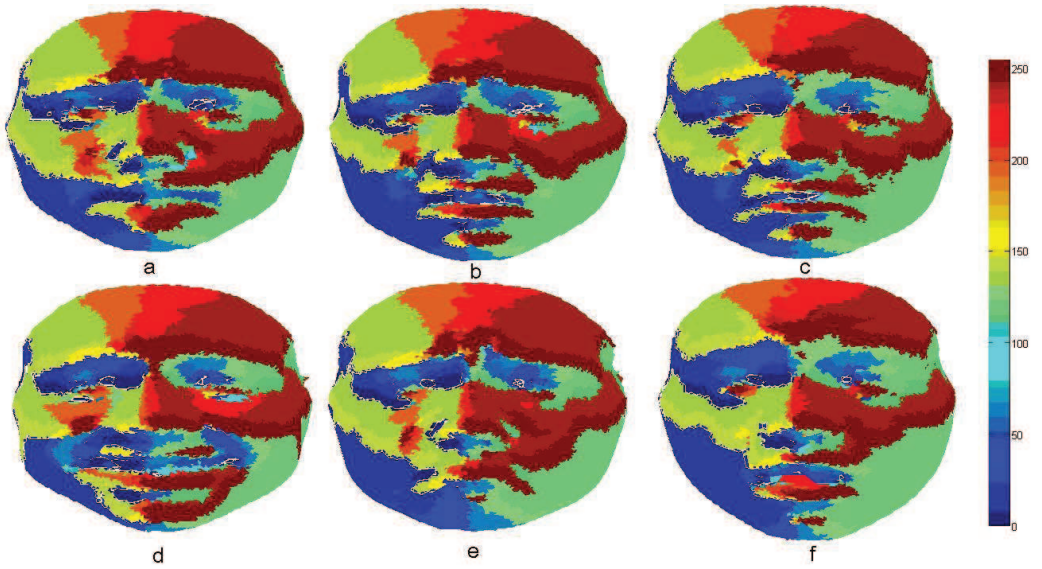


Figure 3.12: Feature extracted from faces with six universal expressions. a: anger, b: disgust, c: fear, d: happiness, e: sadness, f: surprise.

After having extracted the features from a face, facial expressions can be represented by the distribution of the features over the facial region. Indeed, facial expression is the consequence of human emotion and implies facial muscle activation that modifies the facial surface geometry. Such a variation results in the distribution variations of SGAND, as illustrated in fig. 3.12. Thus, one can identify facial expressions by using SGAND.

To find an explicit description of the fundamental structure of facial surface details, we have investigated the statistical distributions of the feature for nine expressive facial regions. As shown in fig. 3.13, 83 manually labelled landmarks are defined on the facial surface, and accordingly, the nine expressive local regions are constructed based on these points. Note that the nose region and interiors of eyes are currently not included in the nine local regions. Nose region is widely accepted as a rigid facial region whose surface shape does not vary with facial expression. Thus, it is useless to include nose region since no expression information provided. Meanwhile, because of the flaw of 3D face scan capture, the interiors of eyes generally contain hole and thus can not accurately record the local shape.

In short, the selected nine local regions cover the most mimic facial areas. In

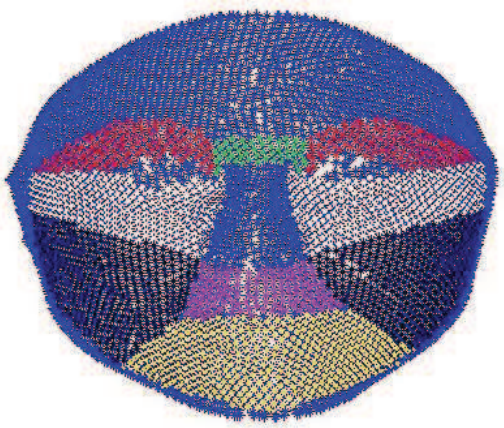


Figure 3.13: Nine selected facial regions labeled by colors other than blue.

each selected region, we model the feature distribution by computing a histogram as follows:

$$L_i = [\frac{n_{i1}}{n_i}, \frac{n_{i2}}{n_i}, ...\frac{n_{im}}{n_i}, ..., \frac{n_{iM}}{n_i}] \tag{3.4}$$

where n_{im} is the number of vertices having a feature value m from 0 to 255, and n_i is the total number of vertices in the ith local region ($n_i = \sum_{m=1}^{M} n_{im}$). $M = 256$ is the number of the slot of our feature value.

The concatenation of nine histogram distributions of the selected entire regions generates an expression descriptor.

$$E = [L_1, ..., L_i, ..., L_K] \tag{3.5}$$

where K is the number of selected regions (K=9 in our experiments).

The descriptor E is computed under multiple radiuses of the circle C in order to precisely represent the local surface around each vertex, and is used to feed Support Vector Machines (SVM) classifiers [Chang & Lin 2001], each one being associated to a radius value. These classifiers allow to obtain, for all expressions from a given face, a set of probabilities following a one-against-one strategy. Specifically, each classifier has been trained with the same number of faces from the six universal expressions. When testing on an unknown face, it is able to output six probabilities

(P_i in eqn. 3.6) corresponding to the six classes. Sets of probabilities from all classifier associated with different radiuses are summarized respectively to obtain the overall probability set, as shown in eq. 3.7.

$$P_i = [P_i^1, P_i^j ..., P_i^E] \tag{3.6}$$

where E is the number of expressions and i represents the index to radius of the circle C. The expression can be recognized by choosing the one with the maximum in the overall probability set. Eq. 3.7 can be tantamount as score fusion for easy understanding.

$$X = \arg\max(\sum_{i}^{S} [P_i^1, P_i^j ..., P_i^E]) \tag{3.7}$$

where $X \in anger; disgust; fear; happinees; sadness; surprise$ and S the number of different radius.

3.4.5 Experimental results

In this section, we present the experimental results obtained for the evaluation of our approaches on pose orientation estimation and on our SGAND for facial expression recognition. The database we have used in the tests is the BU-3DFE dataset.

3.4.5.1 Results on pose estimation

For the evaluation of the pose estimation approach, we have used the neutral faces and faces with the six universal expressions of the two highest level from all subjects so that 1300 face scans have been tested in total.

In fig. 3.14, faces with six universal expressions are displayed with the estimated planes (the black rectangles) and three directions (green, blue, red lines) from PCA.

We have further analysed quantitatively the test results on the 1300 facial models with different number of vertices, poses and expressions. For evaluation, we have manually selected the feature points of inner eye corners and nose corners of each model and derived its orientation as the ground truth of pose orientation. We then

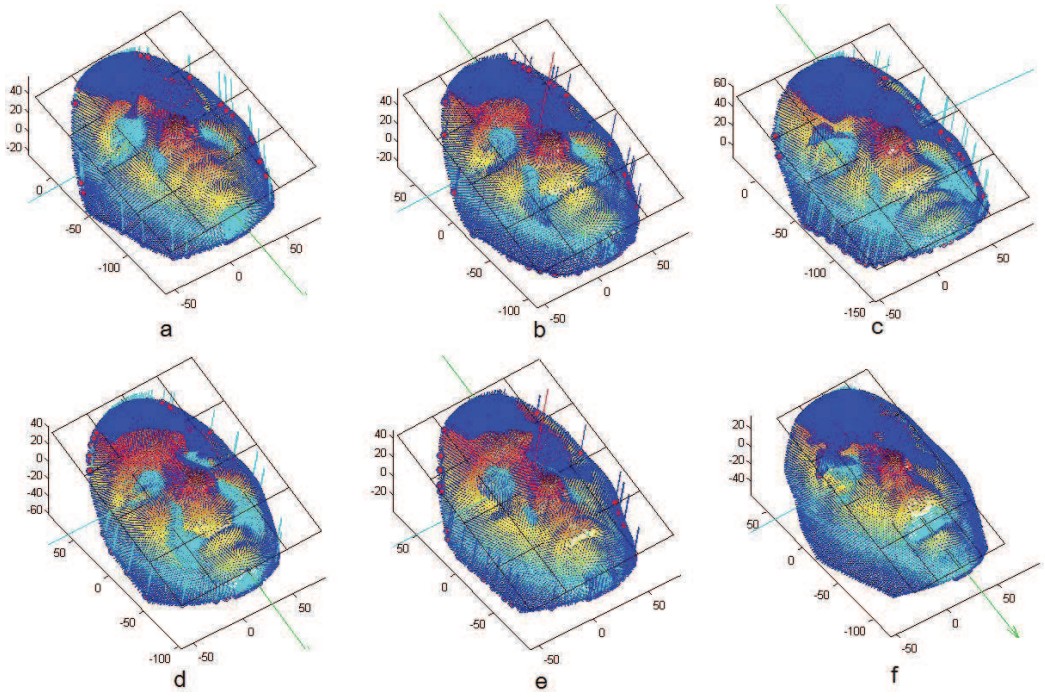


Figure 3.14: Results of pose estimation on faces with the six universal expressions. a: anger, b: disgust, c: fear, d: happiness, f: sadness, e: surprise

compared the estimated pose orientation using our approach with the ground truth orientation, and have considered that the estimated pose is correct if the difference between the estimated and the ground truth pose orientation is less than 10°. The correct pose estimation rates of face models are 94.36% for the normal of the face planes ($D3$), displayed as the red lines in fig. 3.14 , 98.24% for the vertical directions ($D1$), displayed as the green lines and 96.44% for the horizontal directions($D2$) displayed as blue lines. The approach in [Breitenstein *et al.* 2008] achieve a correct rate of 80.8% with the same criterion for correct estimation using their own dataset. In [Seemann *et al.* 2004], a pose success rate of 75.2% for 10° has been achieved. Compared with them, our method appears to be more accurate. However, no direct comparison is possible because different datasets have been used.

3.4.5.2 Results on facial expression recognition

For the evaluation of facial expression recognition, we have used faces with the six universal expressions of the 2 highest level from 60 subjects so that the results can be compared with other works in the literature. Each face has been manually labelled with 83 fiducial points for the face segmentation.

Our facial expression recognition experiments have been conducted in a person-independent way, which is believed to be more challenging than a person-dependent approach. We have followed a ten-fold person-independent cross-validation method, where 60 subjects have been partitioned into two subsets in each round (totally 10 rounds): one with 54 subjects for training and the other with 6 subjects for testing. This experiment setup guarantees that each subject appears once in testing set and 9 times in training set and any subject used for testing does not appear in the training set since the partition is based on the subjects rather than the individual expression.

For these tests, we have set the radius of the circle C to 3, 5, 7, 9, 11, 13mm and have computed the expression descriptor E with these radiuses respectively. Because most of the faces are in rough frontal pose, we have used the Z axis as the main direction and have extracted our features directly on the faces. Then, we have trained an SVM for each E and have fixed its parameter for all rounds of the test. Table 3.4 shows the confusion matrix of the average case for the test. Expressions surprise, happiness, sadness and disgust are well identified with accuracies over 90%, especially 100% recognition rate for the recognition of surprise. However, the anger and fear have quite lower recognition rates. Most of anger expressions are confused with sadness, and fear expressions are more likely to be misclassified to happiness. The average recognition rate for all the six universal expressions is 75.3%.

We mainly compare our results with those in [Wang *et al.* 2006]. The main purpose of the comparison is to show the performance of the proposed feature. The scheme of our method is very similar to theirs so that the efficiency of features can be compared directly and fairly. [Wang *et al.* 2006] extracts another geometry-based feature (primitive features) and compute histogram of features from different

Input \ Output	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	21.7%	13.3%	4.9%	1.7%	56.7%	16.7%
Disgust	0.0%	91.7%	1.7%	4.9%	0.0%	1.7%
Fear	0.0%	3.3%	48.3%	26.7%	15.0%	6.7%
Happiness	0.0 %	1.7%	3.4%	94.9%	0.0%	0.0 %
Sadness	1.7%	0.0%	1.7%	0.0 %	94.9%	1.7%
Surprise	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %

Table 3.4: Confusion Matrix of the person-independent expression recognition.

face regions (segmented by manual landmarks) for facial expression representation. Our average recognition rate (75.3%) is comparable to the average recognition rate (77.8%) using the same classifier (SVM) in their method. Because they have not provided the confusion matrix on recognizing the six universal expressions using the primitive surface feature and a SVM classifier, we can only provide their confusion matrix result in Table 3.5 obtained by a combination of their proposed feature and a LDA classifier. This classifier performs better than SVM, which achieves an average recognition rate of 83.6%. By comparing the two confusion matrix, we can see that our recognition rates for identifying anger and fear are quite lower than theirs. However, the recognition results for other expressions is better, especially for disgust, sadness and surprise. Fig. 3.15 illustrates some failure cases. The distributions of the proposed features (extracted with the radius of 11mm) are quite similar between the left column and the right column. This suggests that the major reason for the significant performance drops for anger and fear compared to [Wang *et al.* 2006] is the lack of discriminant power of the expression description for these two expressions.

Table 3.13 lists several recognition results in the literature using the same database. Among them, [Tang & Huang 2008a] achieves the best average recognition rate of 94.7%, which selects good features from all distances between 83 landmarks by Adaboost algorithm. However, this study requires a neutral face from each subject for distance normalization and thus is subject biased. The recognition rates of other works are reported between 83% and 90% using manual landmarks.

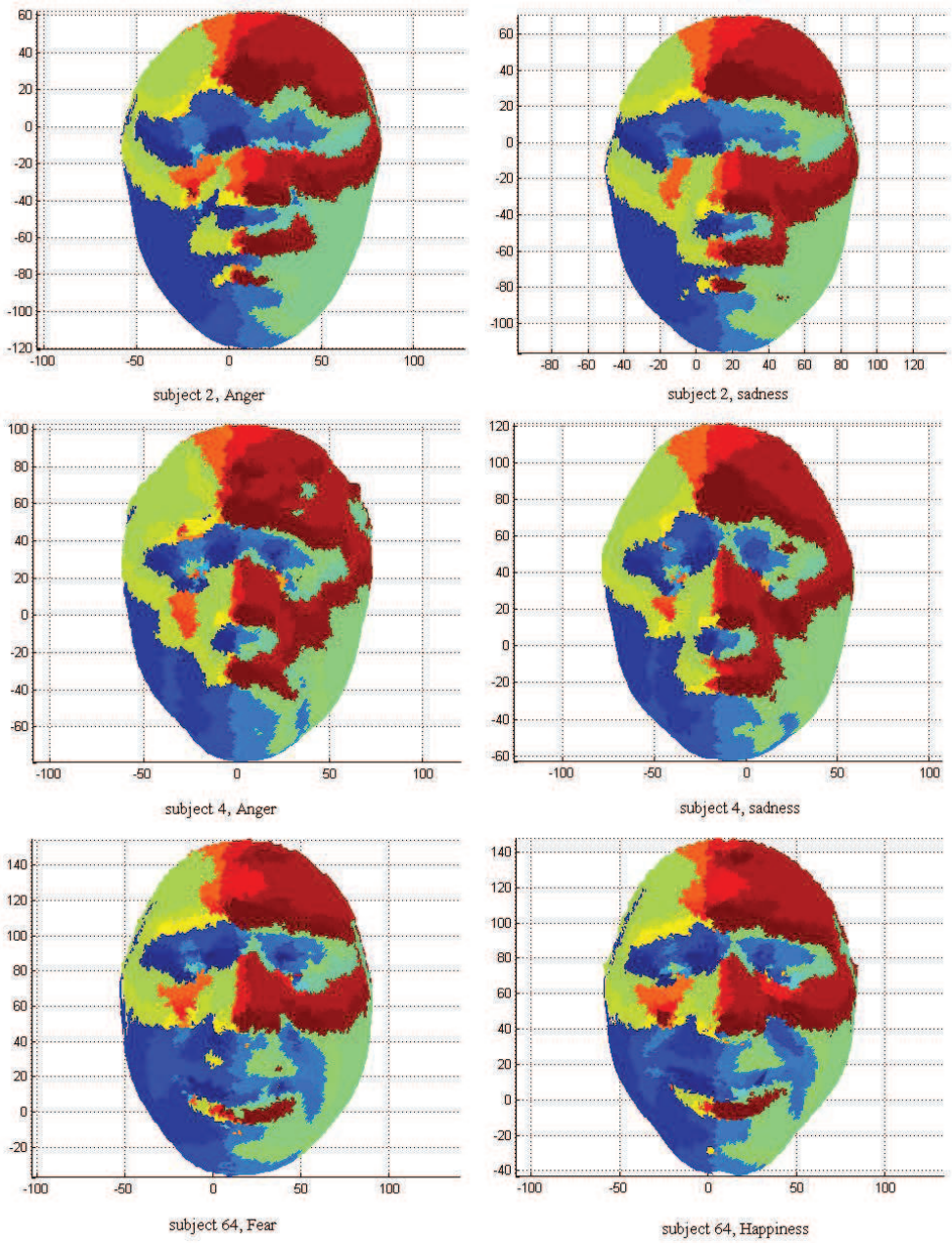


Figure 3.15: Failure cases for expression recognition using the proposed SGAND features. The first and second row show the misclassification of anger into sadness for subject 2 and 4. The third row shows the misclassification of fear into happiness for subject 64. It can be observed that the distribution of extracted features under different expressions are quite similar, which is the main reason for confusion.

Input \ Output	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	80.0%	1.7%	6.3%	0.0%	11.3%	0.8%
Disgust	4.6%	80.4%	4.2%	3.8%	6.7%	0.4%
Fear	0.0%	2.5%	75.0%	12.5%	7.9%	2.1%
Happiness	0.0 %	0.8%	3.8%	95.0%	0.4%	0.0 %
Sadness	8.3%	2.5%	2.9%	0.0 %	80.4%	5.8%
Surprise	1.7 %	0.8 %	1.2 %	0.0 %	5.4 %	90.8 %

Table 3.5: Confusion Matrix of expression recognition in [Wang *et al.* 2006].

3.4.6 Conclusion

In this section, we have discussed and analyzed the popular geometry-based features, including HK curvatures, shape index and primitive surface features. Then, we have proposed our geometry-based feature SGAND, which can be extracted from point clouds of 3D faces and is invariant to face scale, contrary to the other approaches. Indeed, instead of computing the principal curvatures, we describe the local geometry by comparing the number of vertices within the sampled regions above and below a plane defined by center vertex and a frontal face direction. Thus, the extraction of this feature is fast and easy to be implemented.

In order to extract the feature on faces under various pose, we have proposed a pose estimation approach which estimates the frontal, vertical and horizontal orientations of 3D faces. This approach first clusters normals of vertices on a face to detect those vertices on the frontal side. Then, the directions are estimated by a PCA-based regression. Thanks to this approach, our geometry-based feature can be extracted under various head poses.

In order to apply our feature to facial expression recognition, we have used manual landmarks to segment faces into 9 regions, and extract the histogram of the SGAND as the expression descriptors. We have then used SVM to classify the descriptors computed from SGAND under multiple conditions and obtain the final results via score-level fusion.

The experimental results on pose orientation estimation have demonstrated the efficiency and robustness of our approach on faces with various expressions and poses. The experimental results on facial expression are comparable to other works

which use the similar scheme and experimental setup. However, more expressions are better recognized by our approach which is computationally more efficient. Thus, experiments have brought to the fore the ability of our proposed feature to describe efficiently facial local geometry.

However, the local geometry may not carry sufficient information to represent all kinds of deformations caused by various expressions and more information about the local shape property may be necessary, such as texture for characterizing for example bulges and furrows. Thus, in order to identify expressions and action units with high precision, features from different facial representations should be considered. In the next section, we will present our approach based on a Bayesian Belief Net for fusing the features extracted from different face representations.

3.5 3D expression and Action Unit recognition based on a Bayesian Belief Network

Existing 3D facial expression recognition systems mostly aim at identifying the six universal expressions, using geometry-based features extracted from the face surface. Line properties between landmarks, such as angles and distances, are often used and can achieve rather good results. [Tang & Huang 2008a, Soyel & Demirel 2008, Tang & Huang 2008b, Wang *et al.* 2006, Hu *et al.* 2008a]

However, as we discussed previously, expressions are created by facial muscle contractions and result in the variations of landmark locations as well as texture and surface shape in mimic facial parts. By means of these variations, a wide range of expressions other than the universal ones can be exhibited on a face as well as all the 44 basic facial action units (AUs). The geometry based approaches exclude information on other face representations and thus do not make use of the comprehensive characteristics of facial appearance. Although experiments have proved their good performance in recognizing the universal expressions, the geometry based features may be not rich enough to discriminate other subtle expressions or facial action units.

Moreover, feature based approaches generally rely on a large number of precisely

located landmarks, either for feature extraction or for face segmentation. Thus, their performance highly depends on the landmark precision, which can not be achieved by automatically located landmarks. Thus human intervention is generally required in these approaches.

On the other hand, morphable facial models are built by learning the deformation modes on texture and geometry representations, and use the deformation parameters as features for recognition [Ramanathan *et al.* 2006, Mpiperis *et al.* 2008, Rosato *et al.* 2008, Venkatesh *et al.* 2009]. Two major problems arise: firstly, the deformation modes learnt from whole faces describe the major variations globally and thus can not properly reflect local deformation patterns caused by AUs. Secondly, the learnt variation modes are not necessarily consistent with the variations among AUs and expressions, thus may not synthesize expressions accurately. In other words, AUs or expressions can only be approximated by combining a set of variation modes, rather than being modeled by one specific mode in the models. For certain expressions such as happiness or surprise, or action units such as AU27 (mouth opening), the deformation is prominent and thus can be approximated modelled well enough to be distinguishable. However, for some of other moderate expressions and AUs, small variances in parameters yield different expressions.

Therefore, in order to characterize the facial deformations comprehensively, features from all three face representations (facial landmark location or global geometry, texture and local geometry) should be considered. This raises the problem on how to fuse the contribution from each feature efficiently. In this section, we propose to use a Bayesian Belief Network to solve this problem. Beliefs on the expression node for different expressions or AU states are inferred from network parameters of neighboring nodes. Statistical feature models (SFM) are learnt for estimating these parameters on those nodes corresponding to the subject and the facial features.

A distance-like feature is extracted to describe the global geometry relationships of face components. Meanwhile, local information is also extracted not only from the raw facial texture and shape but also from other features such as shape index, LBP so that subtle local deformations can be well characterized and different kind of expressions or AUs can be more distinguishable. Thus, SFMs are learnt for each

type of feature and the parameters are estimated following an uniform process in our BBN. This leads to a flexible system where any new feature can be modeled by a SFM, and the corresponding knowledge directly "plugged" into the BBN.

Moreover, the BBN can be further combined with our SFAM proposed in the previous chapter to realize a fully automatic expression and AU recognition system. Indeed, the adopted features are extracted from local regions on important facial parts. Our SFAM is able to locate landmarks in those regions automatically. We thus use the SFAM as the first part of the automatic system to locate feature points and then extract features around those landmarks for recognition. Because we consider features from three face representations, our system is more robust to landmarking errors than state-of-the-art approaches as it has been proved by the evaluation of the system.

Graphical models have already been used in facial expression analysis in 2D. A Dynamic Bayesian Network is developed in [Tong *et al.* 2007] to model the dynamic and semantic relationships among facial action units. The network has been extended to a more sophisticated one in [Tong *et al.* 2010] which coherently represents head pose and action units. A Bayesian Belief Net aiming at describing the relationship between expression and facial action units is developed in [Datcu & Rothkrantz 2004] for expression recognition.

However, the BBN we propose differs from them in three aspects:

1. The purpose of our graphical modal is to recognizing expressions and AUs within one framework based on data fusion whereas the one in [Tong *et al.* 2007] aims at enhancing the system performance of AUs or facial expression recognized by other classifiers and [Datcu & Rothkrantz 2004] use BBN to interpret the AUs according to the six universal expressions.

2. The structure of our BBN is different. In [Tong *et al.* 2007], the learnt structure of the Bayesian Network explores the dynamic relationship among AUs. In [Datcu & Rothkrantz 2004], the structure of BBN describes the relationship between AUs and the six universal expressions. However, our BBN concentrates on describing the causal relationship among subject, expression and

facial features.

3. Because the objects and the structure of the graphic models are not same, the computation of their parameters is consequently different.

In the following sections, we will present our Bayesian Belief Network as well as feature extractions adopted in this network.

3.5.1 A bayesian belief network for 3D facial expression recognition

In this subsection, we first introduce some background knowledge on BBN and then specify its usage for facial expression and AU recognition. The belief computation in BBN is then presented. Since the BBN structure is elaborated in a unified way for recognizing both facial expressions and AUs with the same procedure, we will use the term 'facial activity' to represent the six universal expressions as well as AUs.

3.5.1.1 Overview of BBN

A Bayesian Belief Network [Duda *et al.* 2000] is a probabilistic graphical model with the topology of a directed acyclic graph (DAG), shown as fig. 3.16. It is made up of a collection of nodes and directed edges, but without directed cycles, as shown in fig. 3.17. Nodes represent a set of random variables and directed edges represent their conditional dependencies.

In fig. 3.17, the 'belief' of a variable on a node X ($X = (x_1, x_2, ..., x_n)$) describes the probability of its states in condition of knowing evidences e (observations) on its connected neighbor nodes. These nodes can be divided into parents (nodes pointed directly to X via an edge) and children (those nodes pointed directly from X via an edge) to compute the belief as:

$$P(X|e) \propto P(e^c|X)P(X|e^p) \tag{3.8}$$

where e^p is evidence on all parents and e^c is evidence on all children.

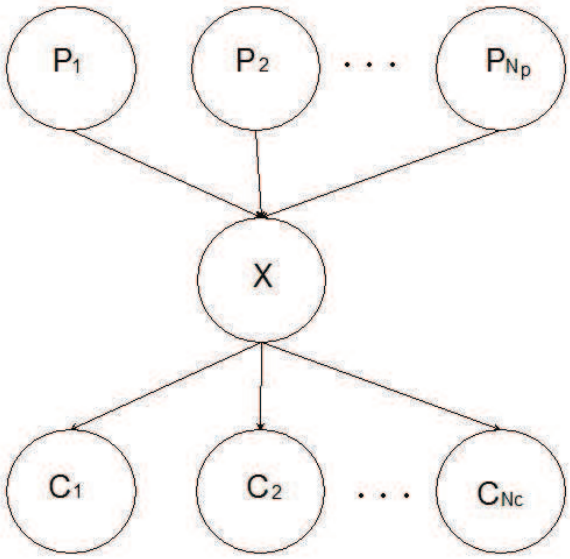


Figure 3.16: An example of Bayesian Belief Network.

The factor about parent nodes in eq. 3.8 is calculated as the conditional probabilities of X under all combinations of all 'parents' states as well as their probabilities given evidences, as in eq. 3.9.

$$\begin{aligned} &P(X|e^p) = P(X|e_1^p, e_2^p ..., e_{Np}^p) \\ &= \sum_{i,j,..,k}^{I,J,...,K} P(X|p_1^i, p_2^j, ..., p_{Np}^k) P(p_1^i|e_1^p) P(p_2^j|e_2^p) ... P(p_{Np}^k|e_{Np}^p) \end{aligned} \tag{3.9}$$

where e^p is the evidence of parents, p_1^i means the ith state of the first parent, p_2^j means the jth state of the second parent, etc. $P(p_1^i|e_1^p)$ is the probability of the ith state (I states in total) of the first parent given its evidence e_1^p. $P(p_2^j|e_2^p)$ is the probabilitiy of the jth state (J states in total) of the second parent, etc. The $P(X|e^p)$ is a sum of totally $I * J * ... * K$ factors.

The factor about children nodes can be rewritten as:

$$P(e^c|X) = P(e_1^c, e_2^c, ..., e_{Nc}^c|X) = \prod_{l=1}^{Nc} P(e_l^c|X) \tag{3.10}$$

where e_l^c is the evidence or observation of the lth child node, N_c is the number of children, $P(e_l^c|X)$ is the probability of evidence knowing the X state.

3.5.1.2 BBN for expression & AU recognition

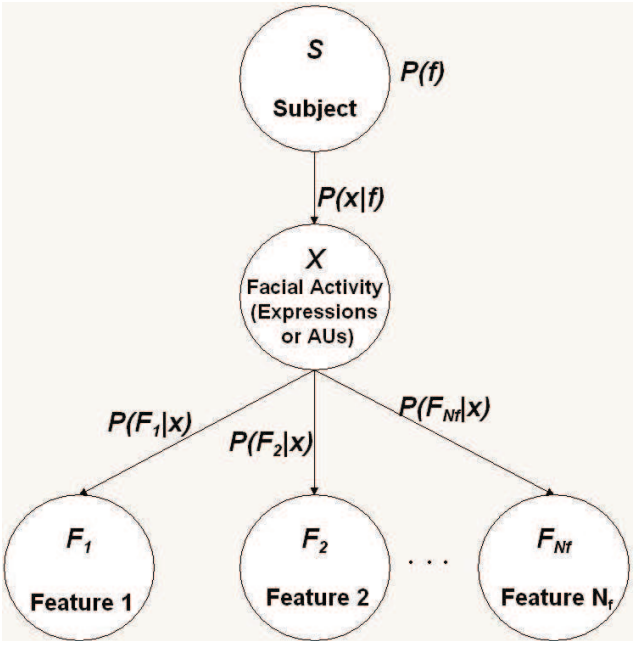


Figure 3.17: The proposed Bayesian Belief Network. We infer belief of states in node X, which represents facial activity (expression or AU), from its parent node S, which represents 3D face scans and its children nodes $F1$,$F2$,..., F_{N_f} which represent facial features (landmark displacement, raw local texture and range around landmarks, etc...)

The structure of our BBN is illustrated in fig. 3.17. The node X represents the facial activity variable and has as many states as the kinds of facial expressions or AUs that are to be recognized, such as six states for the six universal expressions or 16 states for the 16 AUs mentioned later in this section. The node S is X's parent, representing human subjects that we explore. It has as many states as the number of subjects. X's children F_1 ,F_2, ... ,F_{N_f} represent the facial features that are extracted to carry face information.

Since there is only one parent for the node X, the factor $P(X|e^p)$ in eq. 3.8 can be expressed as:

$$P(X|e^p) = \sum_{i}^{N_f} P(X|p_S^i)P(p_S^i|e_S^p) \tag{3.11}$$

where N_f is the total number of subjects that we explore, $P(p_S^i|e_S^p)$ is the prior

probability of the ith subject and $P(X|p_S^i)$ is the conditional probability of X given the state of the ith subject. When all tested subjects perform the same number of expressions (as it is the case for the available face databases), $P(X|p_S^i)$ and $P(p_S^i|e_S^p)$ follow an uniform distribution. Thus, $P(X|e^p)$ also follows an uniform distribution. In other cases, the computation of $P(p_S^i|e_S^p)$ can be based on face recognition approaches while $P(X|p_S^i)$ can be computed either from expression probability distribution in databases, or in a realistic situation, from the frequency of each expression appearing on subjects' face in a period of time in daily life.

Therefore, for a given face κ, eq. 3.8 can be rewritten as follows:

$$P(X|e_\kappa) \propto \prod_{l=1}^{Nc} P(e_l^c|X) \tag{3.12}$$

where e_κ refers to observations from the face κ. Thus, the belief for each expression state is computed from e_κ and the state holding the highest belief is considered as the most probable expression (or AU) of the face κ, as in eq. 3.13.

$$X = \arg\max_X \mathrm{P(X|e_\kappa)} \tag{3.13}$$

Our BBN is derived from the Bayesian Belief Network in [Duda *et al.* 2000], a general example of which is presented in section 3.5.1.1. However, the method to obtain $P(e_l^c|X)$ has to be designed for our specific problem. In our case, we propose to use a statistical feature model (SFM) to estimate $P(e_l^c|X)$ as in the following section.

3.5.1.3 Belief computation for BBN

To know the beliefs for the X node, we need to estimate $P(e_l^c|X)$ for each child node, which is computed based on a statistical feature model (SFM) method. SFMs are built for all features in an uniform manner. Specifically, given a training set for the feature F_l, we divide it into Ne (number of expressions or AUs) subsets containing the corresponding faces. For each subset i_x, Principle Component Analysis (PCA) is applied to learn the variation modes of the feature under the i_x expression, where

95% of major components are preserved.

$$F_l^{i_x} = \bar{F}_l^{i_x} + P_l^{i_x} b_l^{i_x} \tag{3.14}$$

where $\bar{F}_l^{i_x}$ is the feature mean, $P_l^{i_x}$ is the set of eigenvectors resulting from PCA, and $b_l^{i_x}$ is a set of parameters which are supposed to follow Gaussian distributions with a zero mean and a standard deviation $\sigma_{lj}^{i_x}$ where j refers to each parameter of $b_l^{i_x}$. The feature instances $\hat{F}_{l\kappa}^{i_x}$ can be generated from the above equation using feature $F_{l\kappa}$ to estimate the best parameter $b_l^{i_x}$:

$$b_l^{i_x} = {P_l^{i_x}}^T (F_{l\kappa} - \bar{F}_l^{i_x}) \tag{3.15}$$

We set a boundary ($\pm 0.5\sigma_{lj}^{i_x}$) for the corresponding parameter in $b_l^{i_x}$ to form $\hat{b}_l^{i_x}$ in order to constrain the instance deformations and thus to increase their separability. $\hat{F}_{l\kappa}^{i_x}$ is computed by inputting $\hat{b}_l^{i_x}$ in eq.3.14.

The probability $P(e_l^c|X)$ can be considered as the probability of matching the feature $F_{l\kappa}$ with its instances $\hat{F}_{l\kappa}^{i_x}$ knowing the expression state X, which follows a Gibbs distribution.

$$P(e_l^c|X) \propto e^{A_l Q_l} \tag{3.16}$$

Q_l is the matching quality, computed as the normalized cross-correlation between evidence $F_{l\kappa}$ and its instance $\hat{F}_{l\kappa}^{i_x}$, and A_l is a normalization constant.

Inserting the Gibbs distribution into eq. 3.12 and taking logarithm gives:

$$\log p(X|e_\kappa) = \log(\prod_{l=1}^{Nc} P(e_l^c|X)) + c = \sum_{l=1}^{N_c} A_l Q_l + c \tag{3.17}$$

Through the above process, eq. 3.13 can be computed by taking 3.17. A block diagram illustrating the recognition process using the BBN is demonstrated in fig. 3.18.

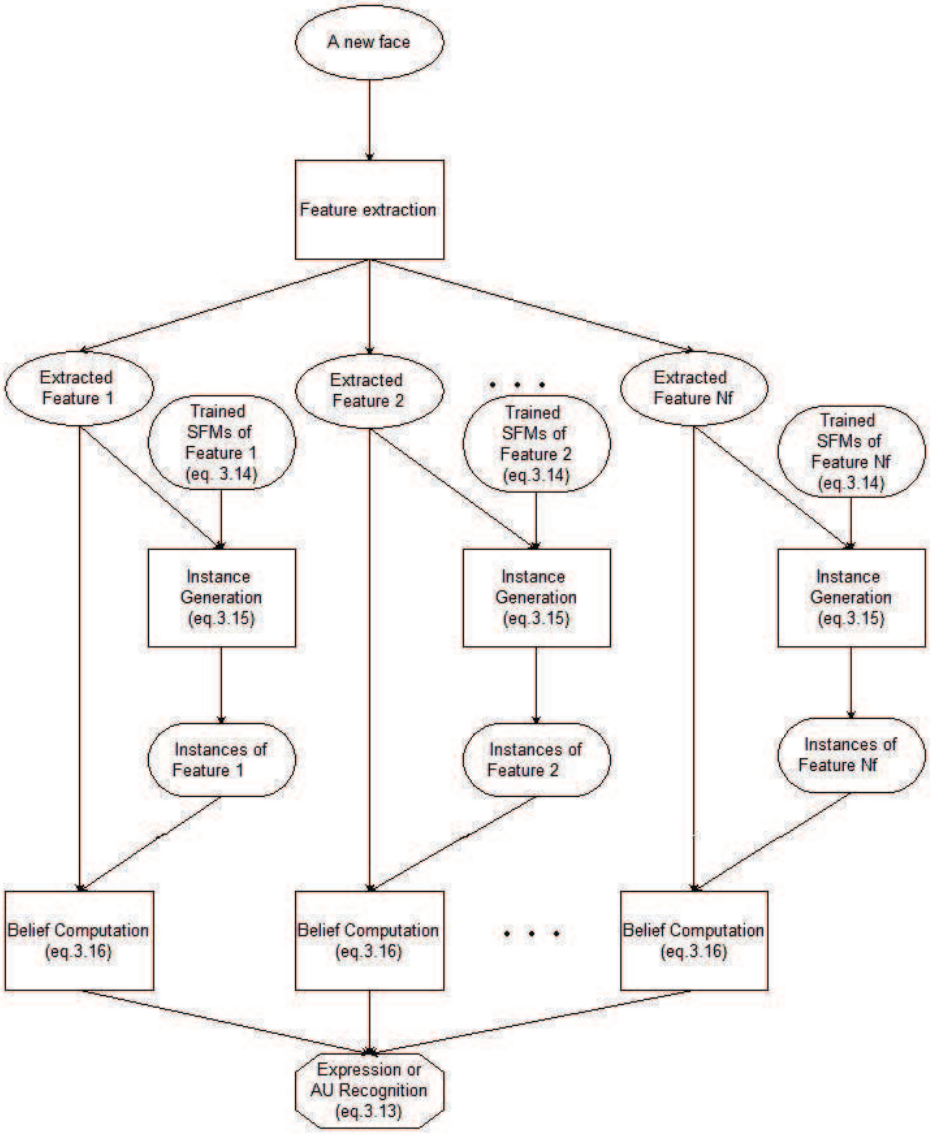


Figure 3.18: Block diagram of the BBN for expression and AU recognition.

3.5.2 Characterization of facial deformations

Two strategies for recognizing facial expressions can be drawn: detection of affects (emotions) and detection of facial muscle actions (AUs). The first one infers what underlies a displayed face, such as the six universal expressions, while the second one aims at describing objectively the facial appearance mostly by FACS. Both rely on the representation and analysis of facial deformations. Our approach for this

purpose is detailed in the following subsections.

3.5.2.1 Facial deformation analysis

Facial activities including expressions and AUs are both consequences of facial muscle activities and the difference between them lies on the muscles involved and the intensity of their contraction. AUs describe facial deformation locally at a low level manner while facial expressions can be considered as a combination of AUs at a high-level manner over the whole face. Some combinations of AUs correspond to basic expressions according to decision making rules. For instance, the combination of AU4, AU5, AU7 and AU24 corresponds to anger. Thus, we are convinced that a good characterization on AUs at a low level can also be effective to represent the six universal expression at a higher level. In the following paragraphs, facial representations are drawn mainly by analysing AUs. However, this representation also applies to facial expressions.

Totally, 16 facial AUs are analyzed in this work which are chosen based on the 3D data availability. They are AU2, AU4, AU7, AU9, AU10, AU12, AU14, AU17, AU18, AU22, AU24, AU26, AU27, AU28, AU34, AU43, illustrated in fig. 3.19. More details on AUs and their combination rules for recognizing emotions can be found in the appendix part.

From fig. 3.19, we can observe that the variations of facial appearance occur in three face representations: facial morphology, facial texture and facial geometry. Specifically, facial morphology consists in a set of reproducible landmarks located on different facial parts. Facial texture contains the unique lines, patterns, and spots apparent in a face skin whereas facial geometry contains facial surface shape information delivered by a face surface mesh. Facial variations caused by AU or expression have an influence on these representations to different extents. For example, AU7 and AU43 change the texture in the eye region significantly without moving corners of the eyes. AU24 changes the local geometry and texture in mouth region mostly while having less influence on landmark location. However, most of AUs influence all three face representations simultaneously and notably, such as AU4, AU10, AU22, AU26, AU27, etc. AUs normally occur locally and change the

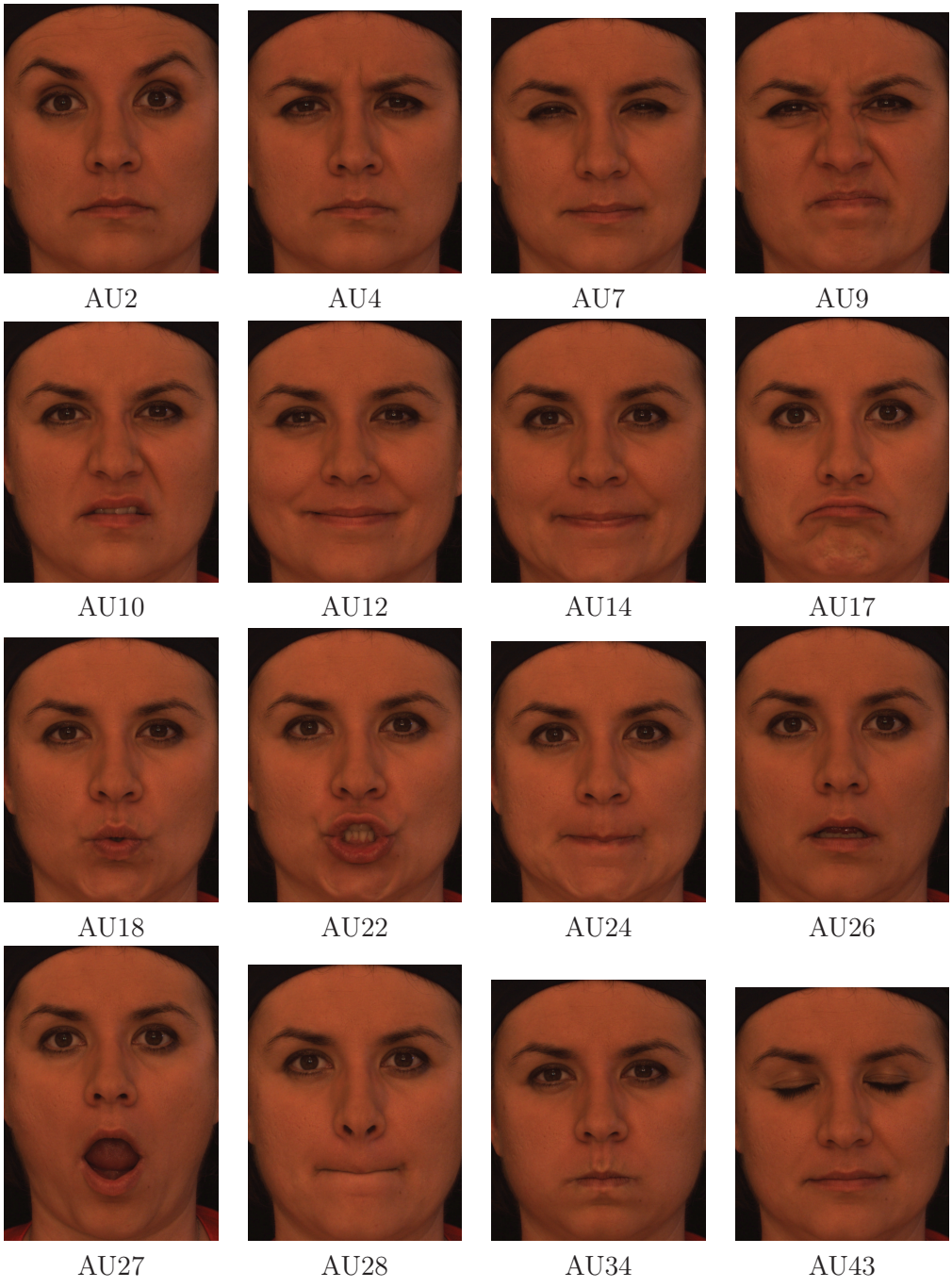


Figure 3.19: Examples of Facial AUs.

appearance in the regions where the corresponding muscles are located. However, some AUs can influence appearance in other regions besides where they happen. For instance, AU10 raises the upper lip while deepens the nasolabial furrow between the

nose and the eyes. Thus, the description scheme we use is based on local regions that are distributed on the important facial parts where most of AUs occur, including the eyebrows, the eye, the nose and the mouth, as it is detailed in the next subsection.

3.5.2.2 Feature extraction

In order to describe facial deformations simultaneously according to the three representations, 19 landmarks are first located manually or automatically and then raw texture and range data are extracted from the local regions around them. Based on these local informations, we further compute other features for enriching face representation to better characterize the morphology described by the relationships between landmarks as well as local texture and geometry.

After a manual labeling of landmarks or an automatic fitting of SFAM on a face, the corresponding configuration is represented by a vector S made up of the concatenation of the landmarks 3D coordinates. Texture feature G and shape feature Z are also extracted by concatenating intensity and range values on remeshed local grids centered at landmarks as in the previous chapter. Local patches in fig. 3.20 correspond to the remeshed grids formed by local shape and rendered by local texture.

$$S = (x_1, y_1, z_1, x_2, y_2, z_2, ..., x_N, y_N, z_N)^T \tag{3.18}$$

$$G = (g_1, g_2, ..., g_m)^T \tag{3.19}$$

$$Z = (z_1, z_2, ..., z_m)^T \tag{3.20}$$

where N is the number of landmarks and m is the number of vertex in all local regions.

For representing the morphology representation, S is used to compute a distance feature L and a point displacement feature D. 11 distances between the involved landmarks are computed and then concatenated into feature vector L. The distances are pictorially shown as green lines in the fig. 3.20 and their textual descriptions are

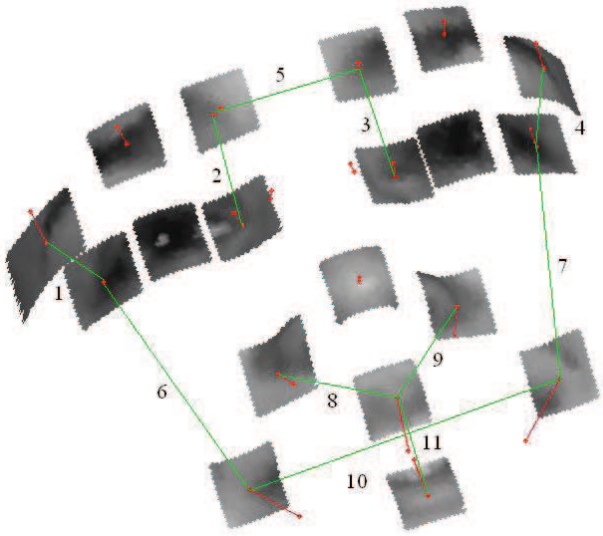


Figure 3.20: Feature extraction

Index	Textual description
1	distance between left outer eyebrow and left outer eye corner
2	distance between left inner eyebrow and left inner eye corner
3	distance between right inner eyebrow and right inner eye corner
4	distance between right outer eyebrow and right outer eye corner
5	distance between left and right eyebrows
6	distance between left outer eye corner and left outer mouth corner
7	distance between right outer eye corner and right outer mouth corner
8	distance between left nose corner and upper mid lip
9	distance between right nose corner and upper mid lip
10	width of mouth
11	height of mouth

Table 3.6: Distances between some strategical facial landmarks on the 3D facial expression model. Distance index refers to the fig. 3.20.

given in Table 3.6.

Feature point displacements represent a change of landmark locations when an expression appears from a neutral face. It is very informative since it represents the shape difference between the face with expression and the neutral one. However, it imposes the constraint that one neutral face from a subject is available and therefore is subject biased. To loosen this constraint, we use a mean landmark set computed from all training neutral faces instead of using landmarks on neutral face of every subject. Thus, D is computed by subtracting the mean of S for training neutral faces ($\bar{S}_{neutral}$) from S (eq. 3.21), represented as red lines in fig. 3.20. This solution avoid

the need for providing a neutral face in conjunction with the face to be recognized with expression, which is unrealistic in a real application.

$$D = S - \bar{S}_{neutral} \tag{3.21}$$

The LBP operator, a powerful texture measure used widely in 2D face analysis, extracts information which is invariant to local gray-scale variations of the image with low computational complexity. Multi-Scale LBP [Shan & Gritti 2008] is an improved facial representation compared to standard LBP (eq. 3.22). We have adopted multi-scale LBP features for three reasons: first, LBP describes local property of images, which is consistent with the local deformations that correspond to AUs; second, the variance in the apparent AU magnitude is large since some are quite notable while some are subtle, thus it is necessary to analyze them under different scales; third, LBP is efficient and easy to compute.

$$LBP_{P,R}(x,y) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{3.22}$$

where $s(x) = 1\ if\ x \geq 0;\ s(x) = 0\ if\ x < 0$, g_c is the value of current pixel and g_p is the value of the neighbors, R is the radius of neighborhood circular and P is number of pixels in the neighborhood. Examples of LBP operator are shown in fig. 3.21. A subset of these 2^P binary patterns, called uniform patterns, can be used to represent spot, flat area, edge and corner [Chan *et al.* 2007].

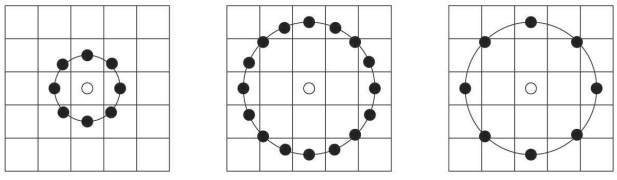


Figure 3.21: LBP Operator. The circular (8,1), (16,2), and (8,2) neighborhoods. The pixel values are bilinearly interpolated whenever the sampling point is not in the center of a pixel.

In our case, LBP are computed and extracted from scale 1 to 5 respectively for all points on the local grids on both texture $LBP^{U^2}_{(16,1)}t$, $LBP^{U^2}_{(16,2)}t$,..., $LBP^{U^2}_{(16,5)}t$

and range maps $LBP^{U2}_{(16,1)}r$, $LBP^{U2}_{(16,2)}r$, ..., $LBP^{U2}_{(16,5)}r$. Superscript $U2$ indicates that the definition relates to uniform patterns with a U value of at most 2 (refer to [Chan *et al.* 2007] for details). fig.3.22 illustrates the extraction of LBP feature at different scales and on both local texture and range maps. Finally, the values for each (P,R) pair on local grids are concatenated into a vector to build 10 LBP feature vectors: $(LBP_t1-5,\ LBP_r1-5)$.

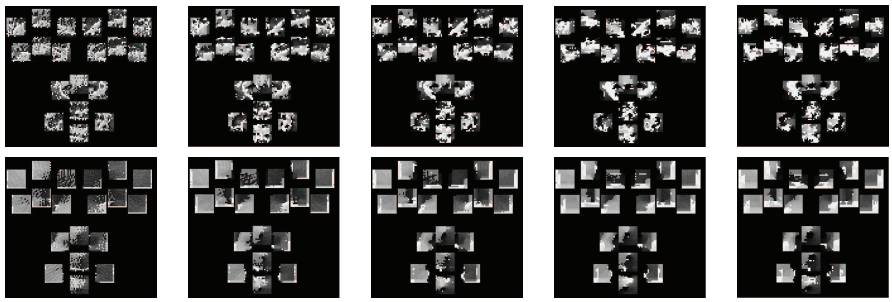


Figure 3.22: Multi-Scale LBP extracted from local texture and range map on a 3D face scan. In the first row are LBP features extracted from texture and in the second row are LBP features extracted from range. In the third row are the (P,R) values of the corresponding columns.

To describe local surface curvature information, we compute shape index of all vertices on the local grids and concatenate them into vector SI. We choose shape index because it has been proven to be an efficient feature to describe local curvature information and is independent of the coordinate system. The computation and more details about shape index can be found in section 3.4.1. The shape index is computed on each vertex on local grids as illustrated in fig. 3.23.

To summarize, 15 types of features are extracted to represent the knowledge used in the BBN as children nodes : D, G, Z, L, SI, LBP_t1-5, LBP_r1-5. Therefore, N_f is equal to 15. These features are summarized in Table 3.7.

3.5.3 Fully automatic expression recognition system

By combining the BBN with SFAM, a fully automatic 3D facial expression recognition system can be realized. It consists of 4 main stages, as shown in fig. 4.3: offline SFAM construction, offline BBN training, online landmarking and feature extrac-

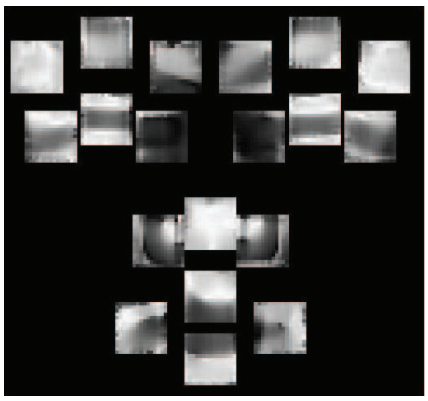


Figure 3.23: Shape index computed on local grids of a face

Symbol	Textual description	Dimension
D	Person independent point displacement of 19 landmarks	57
Z	Range values extracted from 19 local patches	4275
G	Intensity values extracted from 19 local patches	4275
L	Distances extracted from landmarks	11
SI	Shape index extracted from 19 local patches	4275
LBP_t1	LBP feature extracted at scale 1 from 19 local texture maps	4275
LBP_t2	LBP feature extracted at scale 2 from 19 local texture maps	4275
LBP_t3	LBP feature extracted at scale 3 from 19 local texture maps	4275
LBP_t4	LBP feature extracted at scale 4 from 19 local texture maps	4275
LBP_t5	LBP feature extracted at scale 5 from 19 local texture maps	4275
LBP_r1	LBP feature extracted at scale 1 from 19 local range maps	4275
LBP_r2	LBP feature extracted at scale 2 from 19 local range maps	4275
LBP_r3	LBP feature extracted at scale 3 from 19 local range maps	4275
LBP_r4	LBP feature extracted at scale 4 from 19 local range maps	4275
LBP_r5	LBP feature extracted at scale 5 from 19 local range maps	4275

Table 3.7: 15 adopted features and their textual description.

tion, and finally online facial expression/AU recognition. Thus, SFAM is trained using a small set of faces with all kinds of expressions or AUs. A set of statistical feature models are also trained corresponding to these classes and for each feature respectively. During online recognition, faces are first landmarked by SFAM, then a variety of features are extracted and used as evidence by the BBN for computing belief of states for the facial activity node X. Specifically, feature instances are generated corresponding to trained feature models and further used to compute the post-probability of each extracted feature. The output of the system is the type of expression whose corresponding state has the highest belief among different expression or AU states, which are computed from probabilities on both parents and children nodes. Of course, this system is also applicable with manual landmarks. In this case, the landmarking process is skipped for input faces where features are directly extracted based on manual landmarks.

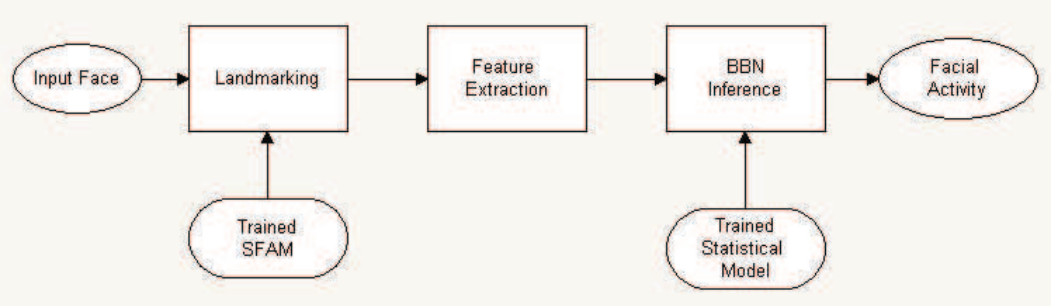


Figure 3.24: Flow chart of the automatic facial expression/AU recognition system

3.5.4 Experimental results

We present in this section our experiments driven in order to evaluate the performance and efficiency of our facial expression/AUs recognition approach based on statistical feature models merged by a BBN. To do so, we have compared the performance of our BBN against other popular classifiers, i.e. Support Vector Machine (SVM) and Sparse Representation Classifier (SRC) on identifying the six universal expressions. Then, in order to prove BBN flexibility and robustness, we have experimented the recognition of 16 AUs. Finally, we have tested the expression

recognition scheme which combines the SFAM and the BBN in order to recognize the six universal expression in a fully automatic manner.

3.5.4.1 Database and experimental setup

In the experiments for facial AU recognition, face scans displaying 16 AUs have been used from 60 subjects in the Bosphorus database[Savran *et al.* 2008], which are AU2, AU4, AU7, AU9, AU10, AU12, AU14, AU17, AU18, AU22, AU24, AU26, AU27, AU28, AU34 and AU43. Thus, 60*16=960 3D face scans have been involved in this tests. Noting that these acted AUs are not FACS coded and singly occurring AUs. The FACS coded version of the database will soon be available.

In the experiments for facial expression recognition, face scans of two high-intensity from each expression have been used from each subject in BU3DFE database [Yin *et al.* 2006]. For both tests using manual landmarks and automatic landmarks, we have used the data of 60 subjects. A part of subjects are different between the tests using manual landmark and those using automatic landmarks, because face scans from a group of subjects are consumed to build the SFAM, which is used to obtain the automatic landmarks on the left face scans. In fact, SFAM has been trained using the data of 11 subjects with scans displaying the six universal expressions at two high-intensity level and neutral. The trained SFAM has then been used to locate 19 landmarks for scans of other 89 subjects.

All tests in AU and facial expression recognition have followed a 10-fold person-independent cross-validation process. Thus, 60 subjects have been partitioned into two subsets in each round (totally 10 rounds): one with 54 subjects for training and the other with 6 subjects for testing. This experiment setup guarantees that each subject appears once in testing set and 9 times in training set and any subject used for testing does not appear in the training set because the partition is based on the subjects rather than the individual expressions.

3.5.4.2 Results for 3D AU recognition

In the test for AU recognition, we have defined the states of the X in the BBN corresponding to the aforementioned 16 AUs.

	AU2	AU4	AU7	AU9	AU10	AU12	AU14	AU17
PR	90.0%	75.0%	78.3%	81.7%	95.0 %	85.0 %	75.0 %	80.0%
FAR	3.6%	26.2%	13.0%	5.8%	10.9%	19.0%	23.7%	7.7%
	AU18	AU22	AU24	AU26	AU27	AU28	AU34	AU43
PR	91.7%	90.0%	76.7%	91.7%	91.7%	81.7%	88.3%	98.3%
FAR	14.1%	3.6%	40.3%	12.7%	3.5%	7.5%	20.9%	4.8%

Table 3.8: Average positive rates (PR) and Average false-alarm rates (FAR) of AUs.

Real \ Predicted	AU_i	non-AU_i
AU_i	True Positive (TP)	False Negative (FN)
non-AU_i	False Positive (FP)	True Negative (TN)

Table 3.9: Explanation of TP and FAR definition.

The results are given in Table 3.8 in terms of average positive rates and average false-alarm rates for all AUs. Indeed, recognizing each AU_i can be considered as a two-class classification according to the AU_i and the non-AU_i. The positive rate is defined as $PR = \frac{TP}{TP+FN}$ and the false-alarm rate is $FAR = \frac{FP}{TP+FP}$ where TP stands for "True Positive", FN for "False negative" and FP for "False Positive" (see Table 3.9 for details) .

Among the 16 AUs, 7 of them (AU10 , AU18, AU22, AU26, AU27, AU2, AU43) have an average PR over 90%, while 4 of them (AU14, AU24, AU7, AU4) have average PR below 80%. Meanwhile, AU24 has the highest FAR, which suggests that it is easily confused with other AUs, which is also the case for AU34 and AU4 having a FAR above 20%. On the contrary, AU43, AU27, AU22 having a FAR below 5% are relatively clearly identified. Globally, our BBN achieves an overall average PR for all 16 AUs of 85.6% with an overall average FAR of 13.6%.

To further demonstrate the performance of our system, we have performed a ROC analysis for each AU. In order to obtain these ROC curves, we have first normalized the set of scores for all 16 AUs of a face to the range from 0 to 1, where 0 corresponds to the minimum and 1 corresponds to the maximum of the individual AU recognition score. Then, for a given AU, all normalized scores from all faces (960) have been computed and used for computing its ROC curve. The decision threshold has been changed from 0 to 1 and the ROC curves are obtained

by plotting the true-positive rates against the false-positive rates. Notice that the values on the left end of ROC curves correspond to the positive rates in table 3.8 because our decision threshold in use is 1 after score normalization. Specifically, the highest score of an AU is always transformed into 1. Actually we choose the state which have this score as the predicted AU. fig. 3.25 and fig. 3.26 are the ROC curves for 16 AUs respectively. The ROC curves which have a greater area below indicates a better recognition. Thus, we can see AU43, AU27 and AU10 are among those best recognized, which correspond to the results in table 3.8.

In [Savran & Sankur 2009], 22 AUs are detected automatically by estimating the deformation between the registered face and the reference. Based on the same dataset, they achieve an average PR of 91.1% . In [Sun *et al.* 2008], 7 AUs are considered and a AU combination on their own database is performed allowing to achieve a PR of 89.1%. In [Tong *et al.* 2010], authors use a Dynamic Bayesian Net to learn the relationship between AUs on 2D Cohn-Kanade database in order to enhance the recognition performance using gabor features and Ababoost classifier. They achieve an 85.8% PR on 14 AUs. Our approach achieves an average PR of 85.6% for 16 AUs, which achieves a consistent result with the highly optimized 2D method [Tong *et al.* 2010].

3.5.4.3 Results for 3D facial expression recognition

In order to evaluate the performance of our BBN, we have compared it with two other classifiers, the Support Vector Machine (SVM) [Chang & Lin 2001] and the Sparse Representation Classifier (SRC) [Wright *et al.* 2009]. All tests have followed a 10-fold cross validation process. The face scans in level 3 and level 4 are tested separately and the final recognition rate is obtained by averaging the results from two intensity levels for all three approaches.

For classification tests using SVM, a multi-class SVM has been trained respectively for each feature extracted from each level of expression (30 SVMs in total). Parameters have been empirically tuned to gain the best performance for each of them. The output of the SVMs is a set of probabilities describing how likely the face belongs to each expression class according to the testing feature. These prob-

abilities (15 in total per level) have been added together and the testing faces have been labeled according to the maximum probability score.

For classification tests using SRC, 30 SRCs have been trained respectively following the principle of the approach proposed in [Wright *et al.* 2009], with a $l^1 - norm$ minimization via orthogonal matching pursuit. Parameters have also been set empirically to obtain the best performance. The SRC output is a set of distances between the testing feature and its six approximations which are generated from a set of coefficients associated with each class. These distances (15 in total per expression intensity level) have been added together and the testing faces have been labelled according to the minimum distance.

	SVM	SRC	BBN
M	83.6% (4.4%)	61.7% (7.9%)	76.9% (8.7%)
T	76.9% (6.8%)	74.7% (6.4%)	75.8% (7.5%)
Ge	84.3% (5.6%)	81.3% (6.7%)	82.9% (5.9%)
M+T	83.1% (6.1%)	78.3% (8.2%)	84.9% (5.8%)
M+Ge	86.4% (5.7%)	83.1% (5.6%)	86.5% (5.1%)
T+Ge	87.2% (4.3%)	83.5% (7.0%)	86.1% (4.5%)
M+T+Ge	88.1% (4.1%)	85.3% (6.8%)	89.2% (3.6%)

Table 3.10: Average recognition rates for the six universal expressions with different features configurations (Morphology, Texture and Geometry) and different classifiers using manual landmarks. The standard deviations over 10 fold tests are the values in the brackets.

Table 3.10 shows the performance in terms of average recognition rates for BBN with different setups on children nodes, as well as the comparison with other classifiers. The first row contains the results where the BBN only has two children nodes for inference, i.e. L, D features extracted from the morphology representation M. The second and third rows contain the results where the BBN adopt features from texture T and geometry representation Ge respectively, i.e. $G, LBP_t 1-5$ and $Z, LBP_r 1-5, SI$. The following rows contain the results with different combinations of these features. We can see that SVM performs better in the tests on each of the single representation, named M, T, Ge. However, BBN is comparable with it in tests on two representations for manual landmarks and finally outperforms SVM on

all three representations with an average recognition rate of 89.2% and least std of 3.6%. Therefore, proved by the tests, BBN is more effective than score-level fusion strategy with SVM and SRC when adopting features from all three representations. Moreover, BBN uses an uniform non-parameter-tuning process for building SFM and estimate parameters, which avoids the trouble for manually tuning parameters to optimize the performances in SVM and SRC.

We have also evaluated the influence of the local grids size and the number of local grids in the feature extraction process. Using the same data as in the previous test, we have first extracted the feature from the same 19 local grids which has a 25mm*25mm size, as shown in fig.3.27a; then we have extracted the feature from selected 32 local grids which has a 15mm*15mm size same as the one in the previous test, as shown in fig.3.27b. The average recognition rate for the test on faces sampled on 19 grids with a size of 25mm*25mm is 90.3% and the average recognition rate for the test on faces sampled on 32 grids with a size of 15mm*15mm is 89.3%. These results (90.3% vs 89.3% vs 89.2%) suggest that it is sufficient and has a lower computation burden to use the grids on the 19 locations with the size of 15mm*15mm to extract features.

3.5.4.4 Results for the fully automatic facial expression recognition

For the fully automatic 3D facial expression recognition, the SFAM has first be used to locate 19 landmarks automatically and then features have been extracted around these landmarks.

The results are given in Table 3.11 with different child node setup for the BBN similar to those in Table 3.10. We can see that the recognition rate increases with the number of child nodes, and finally achieved 84.9% when adopting all children nodes, corresponding to the 15 features. Unlike the results based on manual landmarks, we can not observe an notable contribution of M in the recognition rates (0% vs 3.1%) in the last row which may be due to the inaccuracy of automatic landmarks. Besides, when using only M, an obvious decrease on the recognition rate is observed from using manual landmarks to using automatic ones. This confirms the claim that landmark-based features have a high reliance on locating accuracy and thus

	BBN (m)	BBN (a)
M	76.9%	51.1%
T	75.8%	67.8%
Ge	82.9%	77.3%
M+T	84.9%	67.8%
M+Ge	86.5%	77.5%
T+Ge	86.1%	84.9%
M+T+Ge	89.2%	84.9%

Table 3.11: Recognition rates for 6 universal expressions with different features configurations (Morphology, Texture and Geometry) using both manual and automatic landmarks. The left column is results based on manual landmarks (m) and the right column is results based on automatic landmarks (a).

are sensitive to landmarking errors.

Input \ Output	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	86.7 / 83.3%	2.5 / 3.3%	1.7/ 1.7%	0.0 / 0.0%	9.1 / 11.7%	0.0 / 0.0%
Disgust	3.3 / 3.3%	89.3 / 86.7%	3.3 / 6.7%	0.8 / 0.0%	3.3 /0.0%	0.0/ 3.3%
Fear	1.7 / 5.1%	6.7 / 8.5%	79.1 / 67.8%	6.7/ 8.5%	5.0 / 1.7%	0.8 / 8.5%
Happiness	0.0 / 0.0%	0.0 / 1.7%	5.8 / 3.3%	94.2 / 93.3%	0.0 / 0.0%	0.0 / 1.7%
Sadness	6.7 / 13.3%	0.8/ 0.0%	2.5 / 1.7%	0.0 / 0.0%	90.0 / 83.3%	0.0 / 1.7%
Surprise	0.0 / 1.8%	1.7 / 1.8%	2.5 / 1.8%	0.0 / 0.0%	0.0 / 0.0%	95.8/ 94.6%

Table 3.12: Confusion Matrix of the expression recognition. Left value on each cell is the result based on manual landmarks and right value is the result based on automatic landmarks.

Table 3.12 contains the average recognition rates for the six universal expressions based on manual landmarks (first value in each cell) and by the fully automatic approach (second value in each cell), using the combination of all features (M+T+Ge). The average recognition rate is 89.2% based on manual landmarks and 84.9% for automatic ones. The decrease is mainly due to localization errors for automatic landmarks. Most of the expressions are indeed identified with high accuracy in both tests, while anger and fear have comparatively lower recognition rates. Anger is

classified more likely into sadness because their confusion, even for humans, is much larger than for other expressions. Faces with sadness are more easily to be misclassified into anger in the tests on automatic landmarks. However, the case of fear is different. The motions of this expression are moderate compared to happiness or surprise for example, and thus more difficult to discriminate.

3.5.4.5 Discussion

Table 3.13 presents a comparison with typical results of the literature. While most of other works are dedicated to the recognition of the six universal expressions in 3D, our classification scheme based on BBN and statistical feature models performs the recognition of both expressions and AUs with an uniform structure. It is also found that the proposed BBN outperforms most of the other methods while it requires no parameter tuning and less constraints, such as a large number of landmarks and the neutral face from each subject. Indeed, our approach has achieved the second rank in the literature, the first one having been obtained by [Tang & Huang 2008a]. However, their method requires a neutral face from each subject for distance normalization, which introduce subject bias.

Concerning the fully automatic expression recognition, our results are also of good quality since the second rank in the literature has been reached. Compared with [Mpiperis *et al.* 2008], our approach has two advantages. Firstly, the building and fitting of the SFAM can be easily implemented. Secondly, the recognition by BBN is not only efficient according to the accuracy but also in terms of computational cost. The normalized cross-correlations are computed between each feature and its instances within 0.24s for each child node in average on a desktop PC with Intel Core2 E4400@2.00GHz CPU.

3.5.5 Conclusion

We have proposed a new 3D facial expression and AU recognition approach based on a BBN associated with statistical feature models. We have further combined it with a morphable SFAM to realize a fully automatic recognition of facial expression.

Method	Methodology	Express -ions	Manual Landmarks	Results
[Soyel & Demirel 2008]	Neural network	7	Yes (23)	87.9 %
[Tang & Huang 2008a]	SVM	6	Yes (83)	94.7%
[Wang *et al.* 2006]	LDA	6	Yes (64)	83.6%
[Tang & Huang 2008b]	AdaBoost	6	Yes (83)	87.1%
Our approach	BBN	6	Yes(19)	89.2%
[Mpiperis *et al.* 2008]	Bilinear model	6	No	90.5%
[Venkatesh *et al.* 2009]	Modified PCA	6	No	81.7%
Our approach	BBN & SFAM	6	No	84.9%

Table 3.13: Comparison of the results from different facial expression recognition methods.

Different from graphical models built for 2D facial expression analysis, the proposed BBN has a flexible topology allowing to integrate knowledge carried on new features by adding new children nodes of the X node. By defining the states in the X node, we can change the facial expressions or AUs that need to be recognized. Furthermore, we have proposed a novel parameter estimation method for the BBN which evaluates the similarity between features and their instances generated from statistical feature models. Our experiments have proved that the BBN is more effective than score-level fusion approaches using SVM, SRC when employing features from all three representations. Meanwhile, it is easy to apply BBN in fusing information from a group of features for recognizing since it does not require any parameter tuning procedure. In general, our approach has achieved an average positive rates of 85.6% for 16 AUs and 89.2% for the six universal expressions. Furthermore, thanks to using the feature extracted from three representations, it is robust to the landmarking errors, which allows it to be implemented as an automatic FER approaches. The recognition rate of 84.9% has been achieved for recognizing the six universal expressions automatically. Compared to other existing 3D FER approaches, our method offers the advantages of good performance and implementation simplicity with the ability to be fully automatic.

3.6 Conclusion on 3D expression and Action Unit recognition

In this chapter, we have proposed two approaches for analysing 3D facial expression. In the first approach, a new feature named SGAND, has been proposed to describe

local facial geometry property by comparing the number of sampled peripheral vertices above and below a face plane around a vertex. A head pose estimation method has been elaborated in conjunction with the feature so that it can be extracted under various head poses. SGAND has been evaluated for the purpose of recognizing the six universal expressions.

The results demonstrate the efficiency of SGAND when classifying disgust, happiness, sadness and surprise. However the other two universal expressions are not classified satisfyingly. There are two conceivable directions to improve this approach:

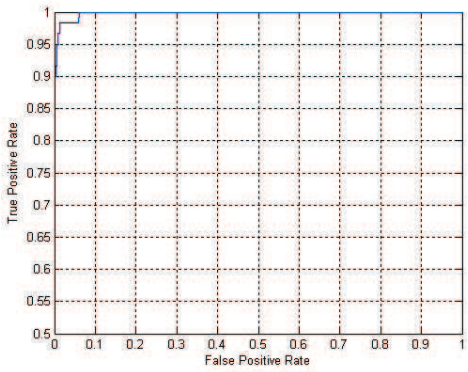
- Feature extraction process: currently, we use a binary value obtained from the numbers of local sampled points on the two sides of the plane to describe the local surface. It is enough to describe the bending trend of the local surface, which is more like a qualitative analysis. However, this binary value is not sufficient to analyse the surface bending quantitatively. Thus, in order to represent the local surface characteristic more precisely, more values will be set according to the distances of sampled vertices to the plane. Moreover, a lookup table will be created to map the value arrays to the typical surface types.

- Expression representation and classification process: We currently use SGAND histograms from face segments to represent expressions. However, this global strategy is reputed to be less effective than the local based face representation. So, we will consider other local-based face representations or a hybrid way to represent facial expressions. Moreover, for the classifiers, besides using SVM classifier, we will evaluate other classifiers such as SRC, LDA, KNN, etc.

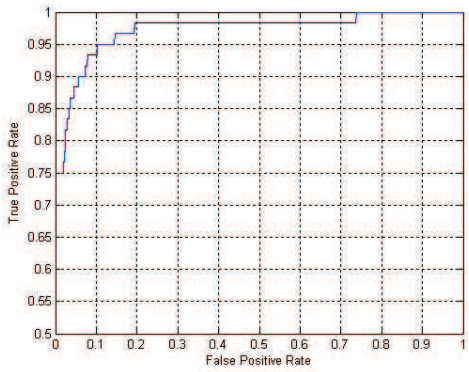
In the second approach a Bayesian Belief Network associated with statistical feature models has been proposed to recognize the six universal expressions as well as facial AUs. The BBN can be further combined with the SFAM proposed in the previous chapter to build a fully automatic facial expression recognition system.

The results demonstrate the efficiency of BBN compared with SVM and SRC to fuse features from different face representations. Using an uniform structure,

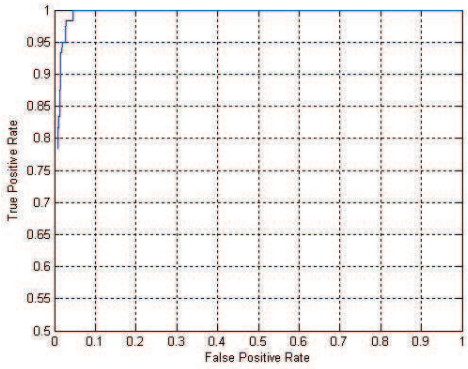
the BBN achieves good results for recognizing both expressions (second rank in the literature) and facial AUs. Tested on automatically located landmarks, the BBN shows its robustness to landmark locating errors. In the future, we envisage to build a probabilistic latent semantic space of AUs and recognize spontaneous expressions based on this space.



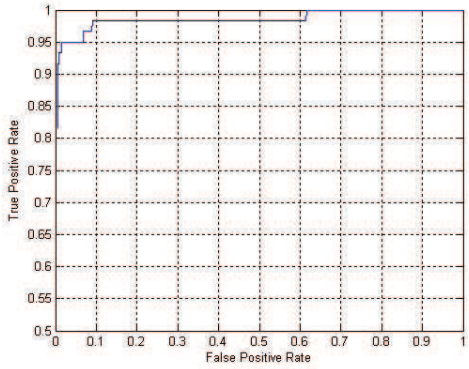
AU2 (99.6)



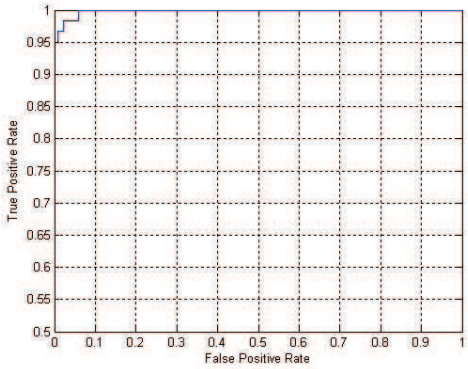
AU4 (96.0)



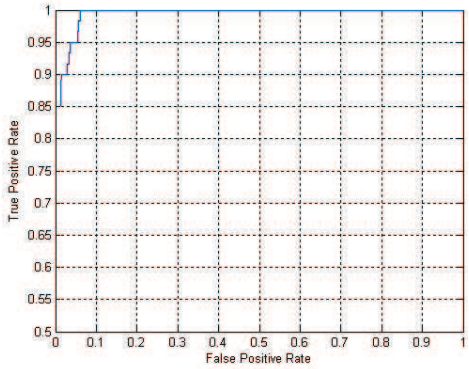
AU7 (99.0)



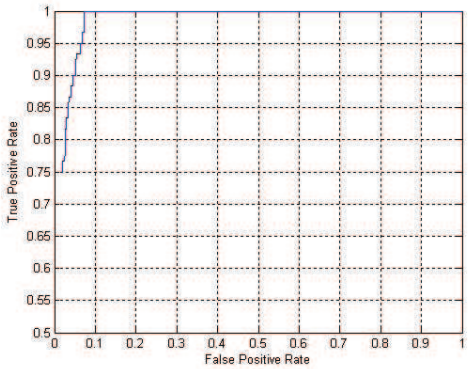
AU9 (98.3)



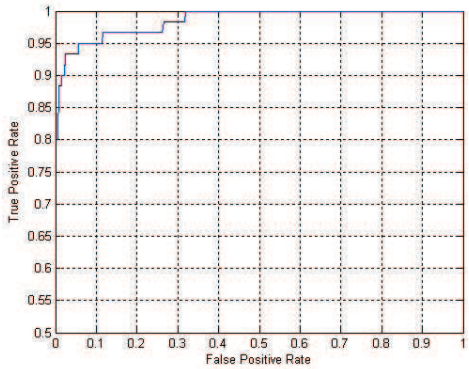
AU10 (99.1)



AU12 (98.4)

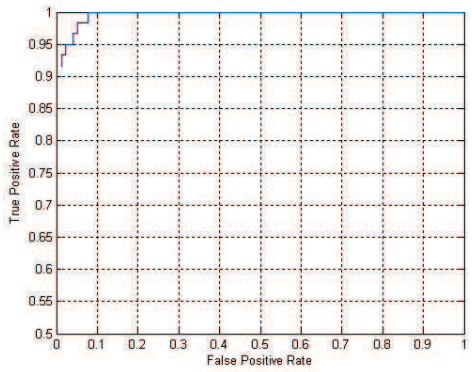


AU14 (97.7)

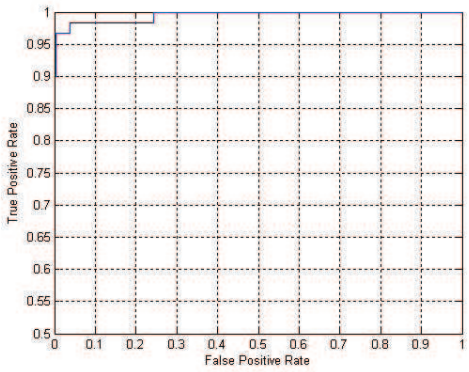


AU17 (98.2)

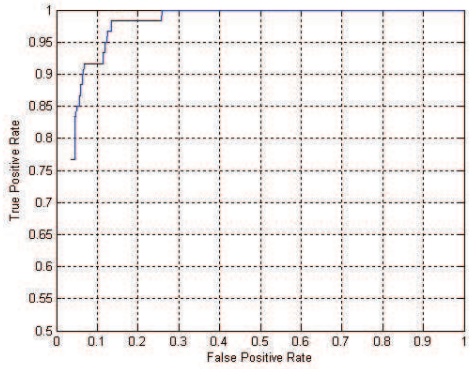
Figure 3.25: ROC curves for the 16 AUs on the Borphorus database. The area under ROC curve is in the bracket. (Part 1)



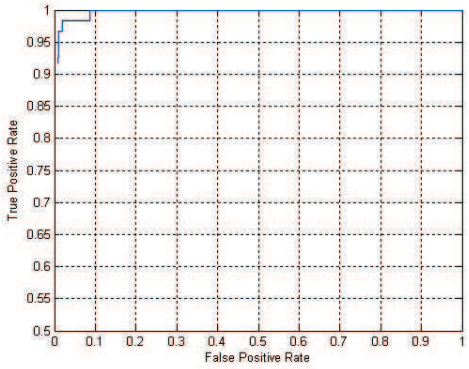
AU18 (98.7)



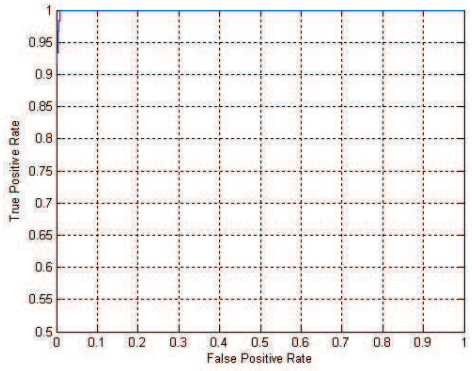
AU22 (99.3)



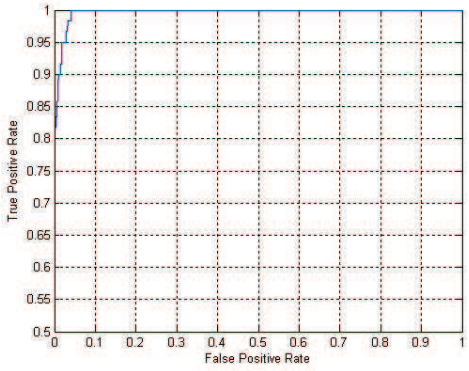
AU24 (95.3)



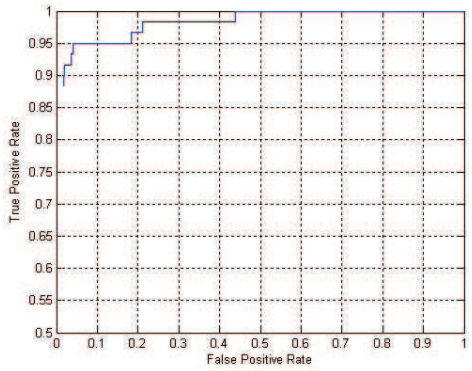
AU26 (98.9)



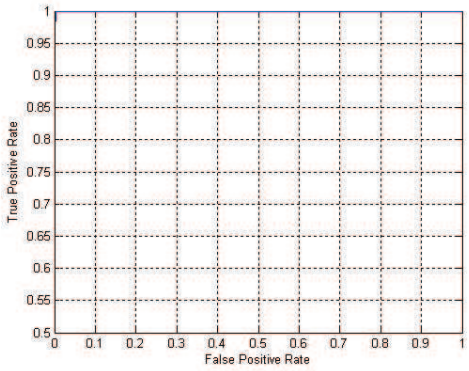
AU27 (99.8)



AU28 (99.3)



AU34 (97.1)



AU43 (99.7)

Figure 3.26: ROC curves for the 16 AUs on the Borphorus database. (Part 2)

a b

Figure 3.27: Two examples of local grid configuration (number and size).

Chapter 4

A minor contribution: People Counting based on Face Tracking

4.1 Introduction

People counting systems aim at automatically estimating the number of people in open or close places. They have wide potential applications including public transportation management system and video surveillance. Several technologies could be envisaged to elaborate such systems. The vision-based approaches are more promising because they can take advantage of the widely used video surveillance systems.

4.1.1 Related work

Traditional counting systems are generally based on infrared or pressure sensors. They are low cost but not easy to integrate with video surveillance system. Vision-based people counting systems become more popular these years for different scenes, like in buildings, streets and hot spots. In the literature, authors developed approaches relying on two main strategies: non-tracking based approaches and tracking based approaches. In the first case, authors try to discriminate foreground from background and count interesting targets. Mehta et al [Mehta & Stonham 1996] made use of classifiers such as neural networks trained to recognize the background in order to facilitate the location and counting of objects in the scene. Moreover, Schlögl et al [Schlögl *et al.* 2003] used motion features to classify each pixel as moving, stationary or background, and then grouped similar pixels together into blobs. They were compared later with the average human size varying with positions in the

scene to estimate people number inside. Chao et al [Chan *et al.* 2008] segmented crowd by motion model and extracted features from each segmentation. The correspondence between features and number of people were learned by Gaussian Process regression. [Dalal & Triggs 2005] proved that locally normalized Histograms of Oriented Gradient (HOG) in a dense overlapping grid can be applied as a successful feature in a pedestrian detector. The speed of HOG based pedestrian detector has been increased significantly in [Cui *et al.* 2008], which make the detector applicable in practical application. However, HOG based pedestrian detector is not applicable to our study because the full body of pedestrian are not always presented in our collected datasets. In the second case, authors either count tracked people at a defined counting line or count people trajectories from tracking. In [Kim *et al.* 2002], a tracking region was partitioned off from the scene with counting line on the edge. people were tracked by motion prediction combined with background subtraction and counted at the line. Another approach consists in getting feature trajectories in the scene by Kanade-Lucas-Tomasi (KLT) tracker, and then clustered trajectories with similar movement together for representing one moving object [Rabaud & Belongie 2006]. This kind of methods are generally able to count a large number of people in a homogeneous crowd. From the state of the art, it appears that most of people counting approaches rely on the assumption that any moving objects in scenes are humans and suffer the miscount of other moving objects. In [Schlögl *et al.* 2003], a model of humans is defined based on average people size. In [Harasse *et al.* 2005], a skin color model is used to detect human. These are among the first tentatives to elaborate more accurate people counting systems but still lack accuracy. In order to avoid this kind of miscounting, the basic idea of our approach is to use the most discriminant human feature: their face.

One fundamental problem of this approach is face detection. Hundreds of approaches have been address on this problem, among which the study by Viola and John [Viola & Jones 2002] has made face detection applicable in real world. [Zhang & Zhang 2010] presents the recent advances in face detection. Earlier studies (before 2004) have been comprehensively surveyed in [Tsishkou *et al.* 2004] and [Yang *et al.* 2002].

4.1.2 Our approach

In this work, we address the problem of counting people moving toward the camera in a close space such as the entrance of a supermarket, bank or bus, where lighting conditions are relatively stable and people are generally facing the camera. Based on these scenes, we propose an approach that presents several improvements compared to the literature. The first improvement is the use of the face detector to ensure that counted objects are people. Second, in order to deal with drastic changes of face scales in our scene, a scale-invariant Kalman filter is proposed. It is further combined with a kernel-based object tracking algorithm to handle face occlusions. Finally, we propose a strategy to count people by automatically classify face trajectories, which are characterized by an angle histogram of neighboring points. Two Earth Mover's Distance based classifiers are used to discriminate true trajectories and false trajectories. The advantages are twofold. On the one hand, a filtering of the trajectories can be realized in order to reject false trajectories caused by false face detection and thus to improve counting accuracy. On the other hand, the automatic classification of the trajectories allows to avoid the manual and empirical elaboration of rules for counting people in a given scene.

4.2 System framework

Fig. 4.1 shows the framework of this system. It combines a face detection module, a face tracking module and a counting module. Synchronizing periodically with the face tracker (every 5 frames), the face detector can initialize tracking for new faces as soon as it detect them, and verify faces being tracked. Moreover, the synchronization results can reveal the events that new faces appear in the video, faces disappear temporarily caused by occlusion and faces leave the scene. After they leave, the face trajectories are sent to the counter for further analysis.

In our work, we use the face detector of [Tsishkou *et al.* 2004] which is based on Viola's one [Viola & Jones 2002]. The overall form of the detection process is that of a degenerate decision tree, what is called a "cascade". Because overwhelming majority of sub-windows is negative for face detection, the cascade attempts to

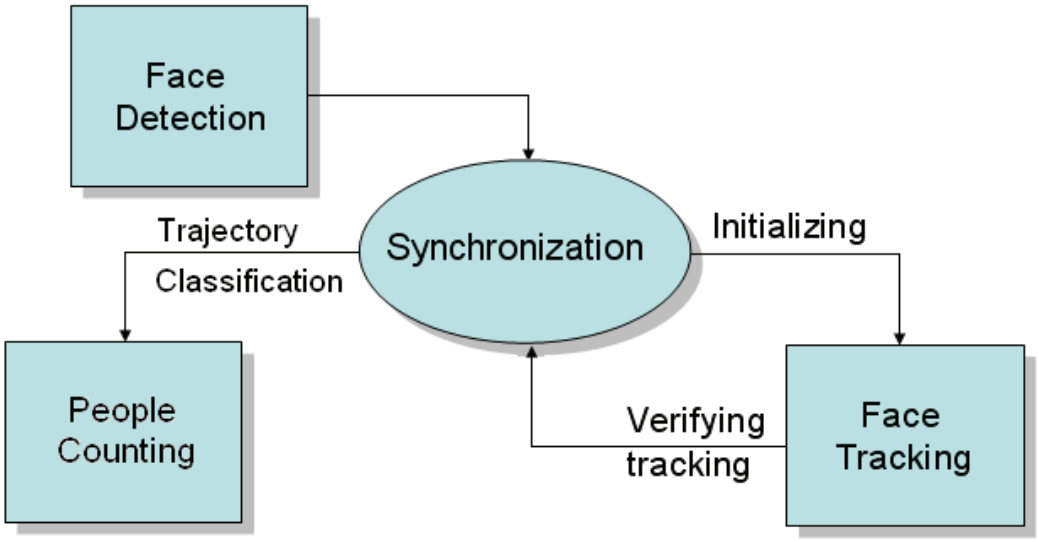


Figure 4.1: System framework

reject as many negatives as possible at the earliest stages. Subsequent classifiers are trained using those examples which pass through all the previous stages. The architecture is extremely efficient in fast and accurate face detection.

After faces have been detected in a frame, rectangle shaped face regions are sent to face tracking. Tracking algorithm adopts linear Kalman Filter for modelling the tracking process, and use a kernel based mean-shift algorithm for evaluating the prediction of Kalman Filter. When faces are severely occluded, mean-shift procedure can hardly find the proper face location. With the prediction from the Kalman filter as an alternative for potential face locations, even if correspond faces are not found in the current frame, it survives the chance to track them in the next few frames till they appear again.

4.3 Face tracking

Existing trackers usually track objects without large scale changes. However, our system faces the difficulties of drastic face scale changes in the scene. Thus, we make an improvement on the original Kalman filter to track objects more accurately under this situation. Face occlusion is another problem we aim to solve. By the prediction of face position from Kalman filter, we can continue tracking the occluded faces

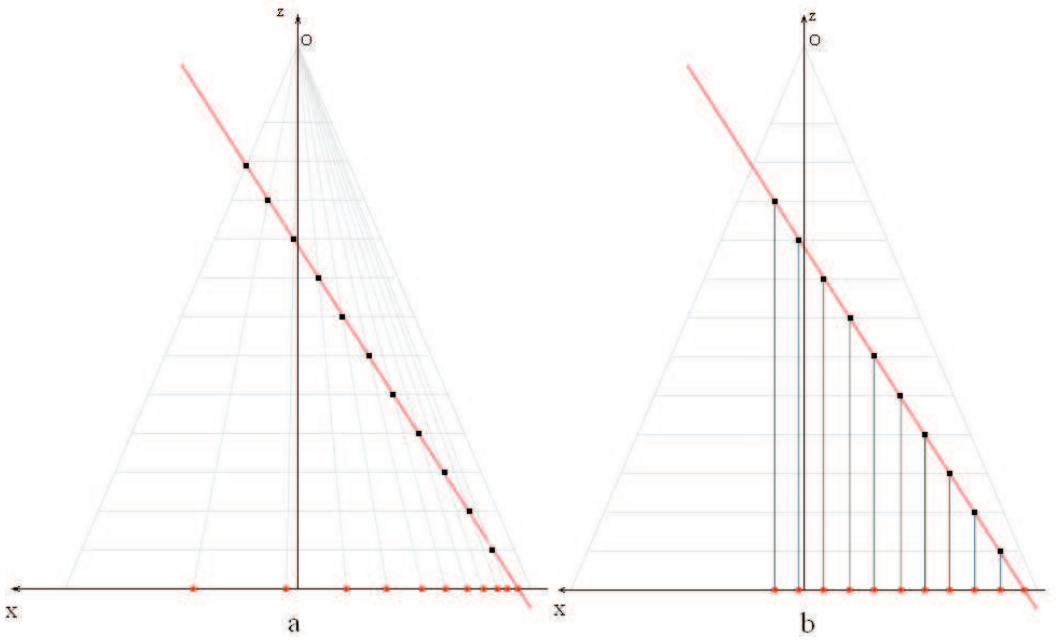


Figure 4.2: Two ways of project face's motion into X plane

until they appear again within a period of several frames.

4.3.1 Scale invariant Kalman filter

In the scenes where faces move towards a camera, an expansion on the face scale is inevitable. As a consequence, faces seem to move faster when they are near a camera in a video. This phenomenon may change the evaluation of movements and introduce process noises into Kalman filter. As shown in fig. 4.2a, the red line is a face's trajectory moving towards the camera O, and the red points are its positions in image sequence with the same time interval. After projected into camera coordinate system, the movement changes from uniform motion (points along the trajectory) to variable motion (points on X axis). This variable motion requires a more complicated movement model than the linear one commonly used in Kalman filter, which is hard to develop.

However, the complexity can be reduced as follows. We consider that face movements with scale changing in a video is "2.5D" movements through image planes, planes vertical to the camera optical axis. Face scales imply some information on

the distance between faces and camera. Based on [Azarbayejani & Pentland 1995] which presented a 3D central projection model to recover 3D positions of tracking objects, we propose a scale-invariant Kalman filter as (4.1) and (4.2), taking the advantage that a face has the constant size in real world but different sizes in different image planes. In our Kalman filter, face movements are projected into a fixed image plane using on face scales and thus "2.5D" tracking problem can be simplified into "2D" tracking problem, like shown in fig. 4.2b.

$$T^k G^k = T^{k-1} A G^{k-1} + W^{k-1} \tag{4.1}$$

where $G^k = [X^k, v_x^k, a_x^k, Y^k, v_y^k, a_y^k]^T, p(W) \sim N(0, Q)$

$$A = \begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix}, m = \begin{bmatrix} 1 & T & 0.5T^2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix}$$

$$T^k = \begin{bmatrix} t & 0 \\ 0 & t \end{bmatrix}, t = \begin{bmatrix} S_x/S^k & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$J^k M^k = J^k H G^K + V^k \tag{4.2}$$

where $J^k = \begin{bmatrix} S_x/S^k & 0 \\ 0 & S_x/S^k \end{bmatrix}, p(V) \sim N(0, R)$

$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, M^k = \begin{bmatrix} Z_x^k & Z_y^k \end{bmatrix}^T$ (X, Y) is the face location, and (v_x, v_y), (a_x, a_y) are the velocity and acceleration of the face movement. A and H are process model and measurement model for Kalman filter. T is the time interval between two continuous frames, k is the index of frames. W^k is the process noise, white Gaussian noise with diagonal variance Q. M_k is the measurement of face location. V^k is the measurement noise, white Gaussian noise with diagonal variance R. S is the face scale, S_x is the face scale in the fixed image plane, like the plane x in fig. 4.2b. In our implementation, S_x is set to 20 pixels, which is the lower boundary of face scale for our face detector.

4.3.2 Face representation and tracking

Each tracked faces are assigned with a Kalman filter and kernal based tracker. For each frame, Kalman filter first predicts the face position for tracker. Then, a coarse-to-fine tracking process is performed by the tracker which handles back a measured face position and scale to Kalman filter for measurement update. In cases of tracking failure due to occlusion or pose variance, the predicted face position and previous scale are give back to Kalman filter. This process is illustrated in fig. 4.3.

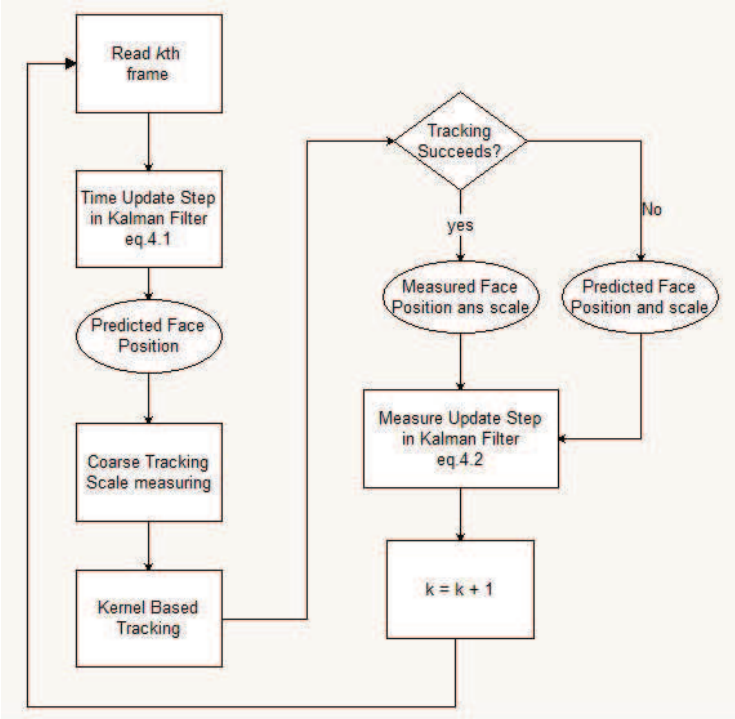


Figure 4.3: Flowchart of the tracking process

Color-based features can be used for tracking non-rigid objects and can keep consistency when face scales change. They also tolerate more changes in pose than edge and texture features. Thus, we use chromatic colors defined in (4.3) to reduce the influence from lighting changes.

$$r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B} \tag{4.3}$$

Detected faces are represented by a kernel based 2-D color histogram, which

consists of 200 bins in each axis. The value of each bin u is calculated as in 4.4.

$$q_u = C \sum_{i=1}^{n} k(\|x_i^*\|^2)\delta\left[b\left(x_i^*\right) - u\right], C = \frac{1}{\sum_{i=1}^{n} k(\|x_i^*\|^2)} \tag{4.4}$$

where δ is the Kronecker Delta function, k is the Epanechnikov kernel function and b is the function projecting pixel (x_i^*) into color feature (rg) space.

The distance among target model and candidates are evaluated by the similarity function ρ in eq 4.5. The maximum value of the function indicates the potential position y^* of the target.

$$\rho = \sum_{u=1}^{n} \sqrt{p_u q_u(y^*)} \tag{4.5}$$

where p, q is the histogram of face model and face candidate respectively with a dimension of $400 * 1$; n is the number of bins in the color histogram. The p of face model is initially computed when face is first time detected and updated when a face is detected close to tracked position. Face detector scans every five frames for a balance of accuracy and efficiency.

For each tracked face, the predicted position from Kalman filter is used to locate the center of sub-image whose size varies dynamically with the face scale. A coarse scan procedure is first processed in this sub-image to approximate the face position. The location with the maximum similarity ρ is used to initialize a fine tracking procedure.

Then, the kernel-based tracking algorithm is used to move the face location iteratively to reach the maximum of similarity between the face model and the face candidate. A dynamic threshold ε_k is set and updated every frame, as shown in (4.6). If the maximum of ρ is above this threshold ε, we consider that the face has been measured at the position obtaining the maximum of ρ .

$$\varepsilon_k = (1 - a)\varepsilon_{k-2} + a\varepsilon_{k-1} \tag{4.6}$$

where a is a weight factor, preset as 0.7.

In the cases where the maximum of ρ does not reach the threshold, we consider that face occlusion happens or the face being tracked has left the scenario. In the

algorithm, the face positions are always predicted no matter the face is occluded or not. If the face is really occluded, we assume the face is at the prediction position. The prediction and the assumption are always made until the face appears again or the face has not been detected for 20 frames consecutively.

4.4 Trajectory analysis and people counting

We have based people counting on classifying potential face trajectories. When trajectories are sent from the tracking module, the counting module is activated. Trajectories caused by false face detection or tracking fragments, meaning that a single face trajectory has been divided into two separated trajectories, are also sent to this module. It is necessary to distinguish them from true trajectories. Since true ones reflect real face movements in the scene, we take the advantage of face moving pattern which is decided by the scene context. An angle histogram featuring the moving direction of a trajectory is used. For one trajectory, its directions in each step are calculated to build an angle histogram, which consists of 36 bins with a 10 degree span for each bin. The value of each bin is calculated for the trajectory T as (4.7).

$$
\begin{aligned}
J_u &= C \sum_{i=1}^{n} \delta[b(\kappa) - u], \\
\kappa &= \begin{cases} \theta, & if \quad x_i > x_{i-1} \\ \pi + \theta, & if \quad x_i < x_{i-1} \& y_i > y_{i-1} \\ -\pi + \theta, & if \quad x_i < x_{i-1} \& y_i < y_{i-1} \end{cases} \\
\theta &= \arctan \tfrac{y_i - y_{i-1}}{x_i - x_{i-1}}, (x_i, y_i) \in T
\end{aligned}
\tag{4.7}
$$

where, δ is the Kronecker Delta function, b is the function projecting degree κ into direction feature space and C is the factor for normalizing the sum of all J_u to 1, $u \in [1, 2..., 36]$. n is the number of bins.

In order to measure the similarity between two groups of J_u from two trajectories, the Earth Mover's Distance (EMD) [Rubner *et al.* 1998] is used. It is more efficient than other distances because it evaluates the similarity of histogram shapes rather than the similarity between their corresponding bins.

Based on this representation of trajectories by angle histograms and EMD, we make use of a classifier trained to recognize true trajectories and count them. Two types of classifiers based on EMD are considered: a K-Nearest Neighbors classifier and a mean-trajectory classifier. For the K-NN method, we use the same amount of true trajectories and false trajectories for training, and find k nearest neighbors for new trajectories by EMD. In mean-trajectory method, we calculate a general direction histogram by averaging histograms from several true trajectories as training procedure. A threshold is set for EMDs between this general histogram and new coming trajectories for discriminating two classes. Trajectories have excessive distances between two frames, which obviously can not be motions from faces. Some trajectories with an extent too smaller are considered as a trajectory fragment. So we pre-filter out those trajectories with these two unreasonable features.

4.5 Experimental results

In this section, we present some experimental results of our scale invariant Kalman filter, face detection and tracking algorithm, and people counting application. They are carried out on different video sequences with multiple faces appearing at different scales.

4.5.1 Scale invariant Kalman filter implementation

We compared our Kalman filter with the original Kalman filter in 3 videos, where a single face moves towards the camera and its scale increases. The frontal face is always showed in the video and we manually measure the nose tips for ground truth position of faces, as in fig. 4.4. Each video was divided into two parts according to face scales. Face scales in first parts varies from the minimum face scale in the whole sequence to around 0.6 of the maximum face scale, and face scales in second parts are from around 0.6 of the maximum to the maximum. ω is a ratio of our Kalman filter's error to the original Kalman filter's error, defined as (4.8):

Figure 4.4: Testing video and face annotation

$$\omega = \frac{E^s}{E^o}, E = \sum_{i \in I} \sqrt{(X_i - X_{GTi})^2 + (Y_i - Y_{GTi})^2} \tag{4.8}$$

where I is different parts of test videos, (X,Y) is the location states of face in Kalman filter(X_{GT}, Y_{GT}) is the ground truth of face locations.

Table 4.1 shows the comparison between two Kalman filters. We can see that compared to the original one, when face scale increases, the error of our Kalman filter decreases more. In other words, our Kalman filter works more accurately when face scales increase.

Index	Video Part 1		Video Part 2	
	Face Scale Range	ω	Face Scale Range	ω
1	45∼89	1	89∼117	0.67
2	41∼69	0.83	69∼88	0.5
3	33∼65	0.92	65∼92	0.67

Table 4.1: Comparison between two Kalman filters

Figure 4.5: Testing video for Kalman filter

4.5.2 Face tracking performance

We evaluated the robustness of the framework when multiple faces appear and occlusion happens. fig. 4.5 shows the results of tracking multiple faces. Three faces moved together in this video and have been detected and tracked separately. Because of the partial face occlusion, the trajectory of first detected face is not smooth in the last frame.

fig 4.6 shows the results when the tracked face experience a totally occlusion. The tracked face had been detected and was tracked for several frames before it was totally occluded by another face. Our tracking algorithm overcame the occlusion and continued to track the face until it appeared again.

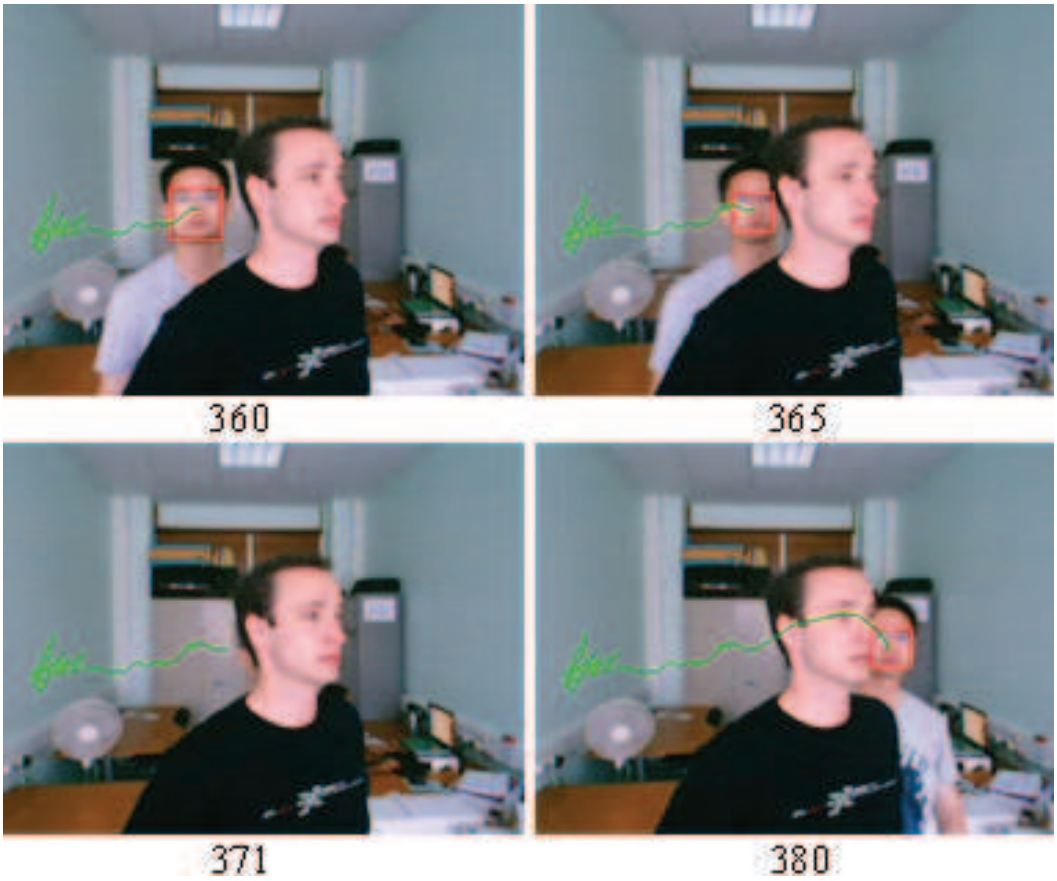


Figure 4.6: Face tracking with occlusion

4.5.3 Trajectory analysis and people counting

We tested the people counting application on our database, which contains 5 videos (6345 frames in total) recorded by the cameras installed at the corridor of our building and the entrance of our conference room, as scenes in fig. 4.4 and 4.5. In these video, people passed either individually or in group more than 100 times. All videos were processed at the size 320*240 pixels. Different databases, like CAVIAR can not readily be used since we require frontal faces detectable by the face detector.

In order to train and test the K-NN classifier and the mean-trajectory classifier, we process our dataset to obtain trajectories. To get more false trajectories to balance the two classes, we tune the detector to have more false detection. Thus 105 true trajectories and 56 false trajectories are obtained. For K-NN classifier, we

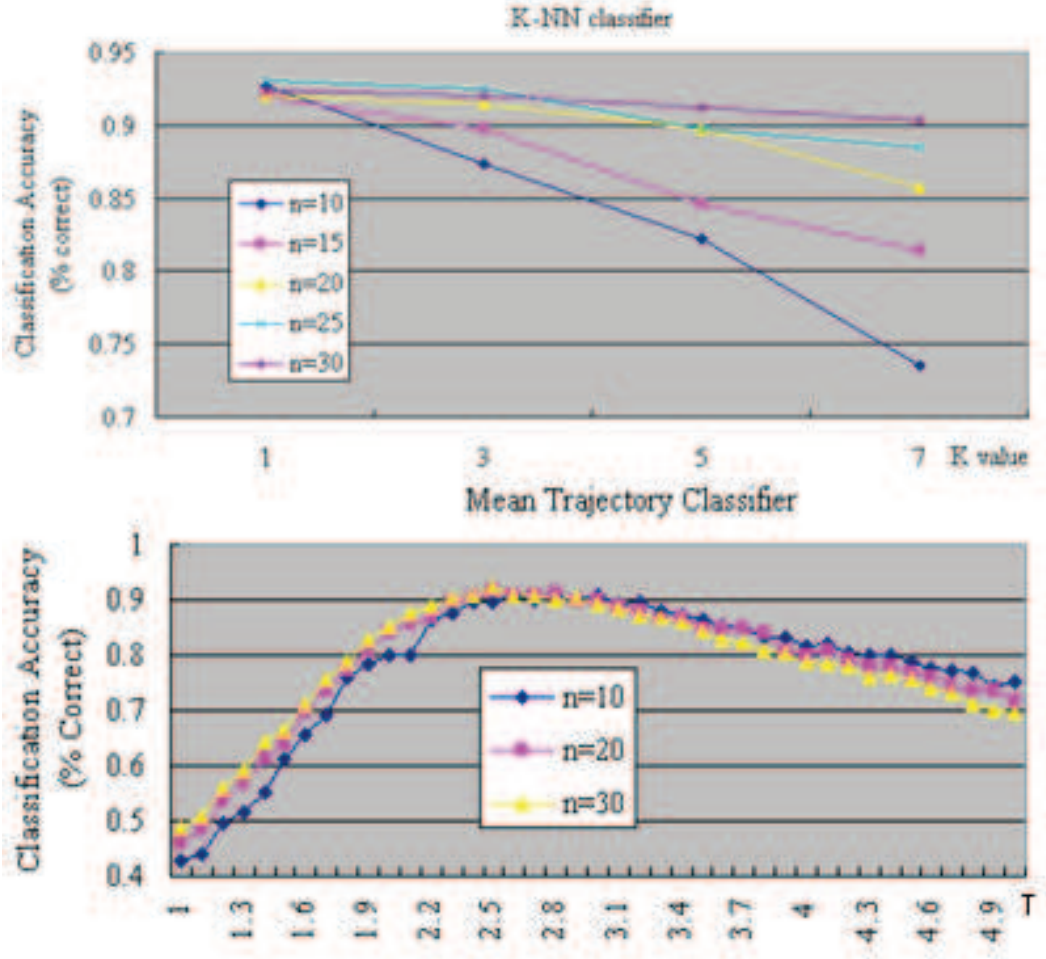


Figure 4.7: Accuracy of Two trajectory classifiers

randomly choose n true trajectories and n number of false trajectories for training and use other trajectories for testing. For mean trajectory classifier, we also choose n trajectories for training and use other trajectories for testing. For each pair of K and n in first classifier and each pair of n and threshold T in second classifier, we test the classification rate for 20 times and choose 10 continuous results with higher scores for evaluation. Results are shown in fig. 4.7. The best counting accuracy we reached 93% by 1-nearest neighbor classification algorithm.

4.6 Conclusion

We have presented in this chapter a novel video-based people counting system that integrates several improvements as compared to the literature. The detection of faces allows to validate that counted objects are human. Then, scale-invariant Kalman filter is proposed to deal with drastic changes in face scales. Moreover, a combination of it and a kernel based object tracking algorithm enhance the robustness of tracking faces with head pose variations and face occlusions. Finally, we have proposed a strategy for counting people based on the automatic classification of potential face trajectories. They are characterized by an angle histogram and the similarities between histograms are evaluated by the Earth Mover's Distance. Thus, not only bad trajectories can be filtered out to enhance the system's counting accuracy but also the automatic classification can avoid the manual and empiric elaboration of rules for counting in a given scene. Our approach has been validated by our experimental results which have demonstrated a good performance on these different aspects and finally a people counting accuracy of around 93%.

In our future work, we envisage to extend our system to other more complex contexts, such as outdoor where illuminations changes drastically. Moreover, the classification of face trajectories will be improved to better fit different contexts by online learning and the adaptation of a more robust classifier.

Chapter 5

Conclusion and Future Works

5.1 Contributions

This research work mainly addresses the problem of 3D face analysis, including facial landmarking and facial expression recognition. The approaches we have proposed for these purposes can also be combined to build a fully automatic facial expression recognition system.

The contributions in this thesis are discussed as follows.

5.1.1 Landmarking on 3D faces

Most of 3D landmarking approaches have a limited capacity for locating non-shape salient feature points since they rely on the landmark geometry. Thus, the possible landmarks that can be located are very limited. Moreover, these methods hardly handle face deformations caused in particular by expressions and occlusions. We think that this limitation can be solved to some extent by characterizing landmarks using both texture and geometry knowledge. We also believe that local properties of landmarks organized by global shape constraints perform better than directly extracting local features. Thus, we have proposed in this thesis to build statistical face models which learn variations of texture and geometry on local regions and variations of global shape configuring those local regions. Two approaches have been proposed for this purpose. In the first method, we have use the global shape on 2D texture map of 3D faces to locate landmarks by varying parameters of shape model to find the best match between the query face and our model. This method is dedicated to 2.5D faces scans. Thus, we have provided a second approach making use of the full 3D information when it is available. This method relies on a 3D morphable

partial face model (SFAM) which learns variations in 3D shape as well as local texture and local geometry. The fitting is performed thanks to the minimization of an objective function describing the similarity between a query face and SFAM with consideration of partial occlusion, thus enabling landmarking on partial occluded faces. The optimization of the objective function is accelerated by pre-computing correlation meshes. Moreover, an occlusion detection method has been proposed to detect the local regions occluded and give a set of occlusion parameters for the objective function.

Experimental results have demonstrated that by considering both texture and geometry information, our methods is able to locate a set of landmarks beyond those characterized by salient shape with a better accuracy. Thus, SFAM has reached a better landmarking ability than the previous models proposed in the literature in terms of accuracy and robustness when encountering severe conditions such as expression and occlusion.

5.1.2 3D facial expression recognition

Most of works dedicated to expression recognition on 3D faces use holistic features or deformable model, such as line property between landmarks, histogram of primitive surface feature, and morphable face model. Either they require a high precision on landmark locations for feature extraction and face segmentation or they are limited to use raw texture and range contained in the model for computing parameters. However, expressions or action units are consequences of facial muscle contraction reflecting in both facial texture and geometry. Moreover AUs generally appear locally and subtly and thus it is hard to distinguish them by raw face texture and geometry feature, such as color, intensity or range data. Thus, we have proposed in this thesis to extract features from multiple face representations, including face morphology, texture and geometry. In order to combine the contribution of all these features, we have proposed a graphical model which is a Bayesian Belief Network with a structure organizing all features as children nodes of the expression node. Contrary to previous proposed BBN for 2D facial expression, our BBN has an uniform structure which describes the casual relationship among subjects, expression

or AUs appearing on faces and features extracted from face appearance. Moreover, it has a flexible topology allowing to integrate knowledge carried on new features and to express facial activity as expression or AUS. Thus, it can be applied on both expression recognition and AU recognition problems. By combining BBN with SFAM, a fully automatic facial expression recognition system is elaborated.

The experimental results have demonstrated the efficiency of BBN compared with SVM and SRC to fuse features from different face representations. Using a uniform structure, the BBN achieves good results for recognizing both expressions (second rank in the literature) and facial AUs. Tested on automatically located landmarks by SFAM, the BBN shows its robustness to landmark locating errors.

Moreover, in order to enrich information used for 3D face analysis, we have also proposed in this Ph.D work a new 3D facial feature, named SGAND, to characterize the face geometry properties. Indeed, pose-invariant features for 3D faces can be a shortcut for face analysis because using this kind of features avoids the procedure of face alignment. However, most of pose-invariant features, such as shape index, HK curvature, are sensitive to face scale because they are extracted from face meshes which varies with face scale. On the contrary, the SGAND feature we have elaborated to characterize surface properties only relies on the point clouds instead of face meshes. Thus, this feature is insensitive to scale, easy to implement and quick to compute. It relies on the comparison of numbers of vertices above and below face planes with a preset direction in sampled local regions of 3D faces. In order to compute this direction, a head pose estimation method has been developed in conjunction with the feature so that the feature can be extracted under various head poses. As experiments have shown, SGAND feature has been applied successfully to the recognition of the six universal expressions.

5.1.3 People counting based on face tracking

Finally, our last contribution concerning face analysis deals with people counting based on face tracking. Existing people counting systems rely on the assumption that detected or tracked objects are humans. Some of them have a preliminary people model to verify this assumption, such as the ratio of height and width of ob-

jects. Unfortunately, these methods suffer from inaccuracy when validating humans. Thus, we have proposed a method that makes use of the face, the most discriminative feature of human to accomplish this validation. This approach is composed of a face detector and a face tracker that collaborate to detect and track faces in 2D videos. The face detector is cascade Adaboost classifiers and the tracker is a combination of Kalman filter and the kernel based object tracking algorithm. The tracker is improved to be used in the senarios when people move towards the camera. In this case, face scale varies drastically so that it introduces errors to the tracking process using traditional Kalman filter. Thus, we have designed a scale-invariant Kalman filter which tracks faces in an image plane where the face trajectories are projected, so that 2.5D face movements (face movement with scale changes) can be normalized to 2D face movements. Face trajectories from the tracker are featured by histogram of moving directions and classified using a K-NN classifier. By doing this, bad trajectories caused by false face detection and tracking fragment can be filter out so that only correct face trajectories are counted for people counting. Our approach has been validated by our experimental results which have demonstrated a good performance on these different aspects and finally a people counting accuracy of around 93% as been reached.

5.2 Perspectives for future work

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

5.2.1 Further investigations on 3D landmarking

In this thesis, local range and texture maps have been used as simple features to represent local shape and texture around a landmark. In the future, the landmark location may be improved by extracting other features such as our proposed SGAND feature, HK curvature, shape index, etc. for shape feature, and Local Binary Pattern, Gabor filtering, etc. for texture property within our statistical landmarking framework.

Another improvement may concern the constraints applied to instances gener-

ated by SFAM during the fitting process. Indeed, SFAM parameters (b_i) are empirically limited to constrain possible deformations. We plan to add a process to set the boundaries of their variation range according to the face properties available in the training data.

5.2.2 Further investigations on 3D facial expression recognition

The sign-based AU recognition allows to interpret facial muscle activities as expressions or affect states thanks to high-level decision making rules. Approaches using the mapping rules require that the prior knowledge on the relevant AUs is available in the training so that they can recognize these AUs on a testing face and thereafter identify expressions by applying rules. For example, the combination of AU1, AU2, AU5 and AU27 corresponds to the surprise in FACS rules. However, in no face displaying AU5 is present in the training set, this AU can not be recognized on the testing faces even if the face displays a combination of AU1, AU2, AU5 and AU27. Therefore, the recognition of surprise fails because of the absence of AU5 from AU recognizer. Thus, instead of using existing high-level decision rules, we envisage to build a latent AU space with the basis of available AUs that can be recognized by our BBN. Then, for a given face, the set of its AU beliefs can be projected in this space and the corresponding position can be used for expression recognition.

Concerning the universal expression recognition approach relying on our SGAND feature, we envisage to enhance the performance as follows. Fuzzy neighborhood relationships between some expressions or emotional states, for instance between anger and sadness, lead to unnecessary confusion between them when a single global classifier is applied. In the future we will propose a multi-stage classification method dealing with the expression classification in several stages. The basic idea will be that affect states can first be categorized into some broad and rough emotional classes according to the dimensional emotion model in one of the dimensions, such as arousal dimension, and then each broad emotional class can then be further classified into final emotional states according to other dimensions, such as appraisal dimension.

The learning methods used in this thesis is mainly based on PCA, which can only learn linear face models. The learnt models contains mixture variations on identi-

fication, expression and illumination. More recently TensorFace has been proposed for a multi-linear analysis to model explicitly the multiple modes of variations in these factors and their inter-relationships [Jia & Gong 2005]. Thus, in the future, we will investigate 3D TensorFace for a joint recognition of facial expression recognition and face recognition.

Chapter 6

Appendix: FACS and used Action Units

FACS describes all visually distinguishable muscular activities that produce momentary changes in facial appearance on the basis of 44 unique AUs, as well as several categories of head and eye positions and movements. Each AU has a numeric code. They are sorted into three categories: upper face AU, lower face AU and Miscellaneous AU. The first category includes AUs named Inner Brow Raiser (1), Outer Brow Raiser (2), Brow lowerer (4), Upper Lid Raiser (5), Cheek Raiser (6), Lid Tightener (7), Eyes Closure (43), Blink (45), Wink (46). The second category includes AUs named Nose Wrinkler (9), Upper Lip Raiser (10), Nasolabial Fold Deepener (11), Lip Corner Puller (12), Cheek Puffer (13), Sharp Lip Puller, Dimpler (14), Lip Corner Depressor (15), Lower Lip Depressor (16), Chin Raiser (17), Lip Puckerer (18), Lip Stretcher (20), Lip Funneler (22), Lip Tightener (23), Lip Presser (24), Lips Part (25), Jaw Drop (26), Mouth Stretch(27) and Lip Suck (28). The third category includes Lips Toward Each Other (8), Tongue Show (19), Neck Tightener (21), Jaw Thrust (29), Jaw Sideways (30), Jaw Clencher (31), Lip Bite (32), Blow (33), Puff (34), Cheek suck (35), Tongue Bulge (36), Lip Wipe (37), Nostril Dilator (38), Nostril Compressor (39). It is crucial to note that while FACS is anatomically based, there is not a one-to-one correspondence between muscle groups and AUs, since a given muscle may act in different ways and thus produce different appearance.

6.1 AU Examples

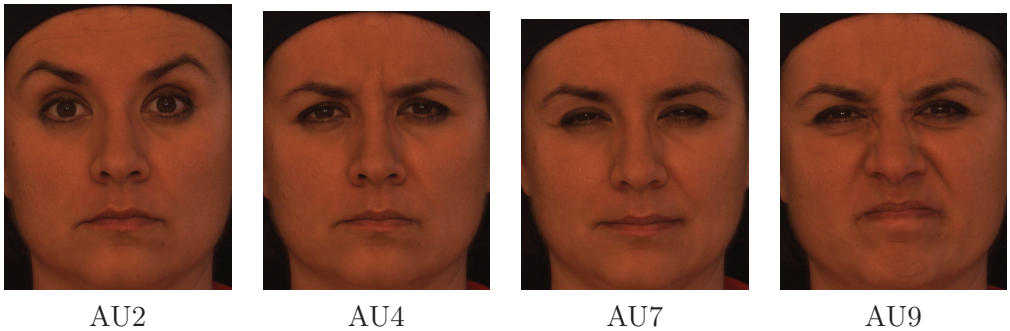
Totally 16 facial AUs are analyzed in the chapter 3. They are AU2, AU4, AU7, AU9, AU10, AU12, AU14, AU17, AU18, AU22, AU24, AU26, AU27, AU28, AU34, AU43 respectively. In order to analyze their characteristics, we demonstrate these AUs here and give explanations on them.

AU2 Outer Brow Raiser: The muscle that underlies AU2 originates in the forehead and is attached to the skin in the area around the brows. In AU2 the action is upwards, pulling the eyebrows and the adjacent skin in the outer portion of the forehead upwards towards the hairline. It produces an arched shape to the eyebrows and causes the lateral portion of the eye cover fold to be stretched upwards.

AU4 Brow Lowerer: Three muscle strands that underlie AU4. One strand runs obliquely in the forehead. Another strand emerges from the root of the nose. A third strand runs from the glabella to the medial corner of the eyebrow. It lowers the eyebrow and pushes the eye cover fold downwards and may narrow the eye aperture. Meanwhile, it pulls the eyebrows closer together and produces vertical wrinkles between the eyebrows as well as an oblique wrinkle or bulge running from the middle of the forehead down to the inner corner of the brow.

AU7 Lid Tightener: The muscle that circles the eye orbit is the basis for AU7. This muscle runs in and near the eyelids. When it is contracted, AU7 pulls both upper and lower eyelids and some adjacent skin below the eye together and towards the inner eye corner. It tightens eyelids and narrows eye aperture. It raises the lower lid so it covers more of the eyeball than is usually covered. Meanwhile, the raising of the skin below the lower eyelid causes a bulge to appear in the lower lid.

AU9 Nose Wrinkler: The muscle underlying AU9 reaches from the area near the root of the nose downward to a point adjacent to the nostril wings. When contracted, this muscle pulls skin from the area below the nostril wings upwards towards the root of the nose. It pulls the skin along the sides of the nose upwards towards the root of the nose causing wrinkles to appear. It also lowers the medial portion of the eyebrows and pulls the center of the upper lip upwards as well as narrows the eye aperture.

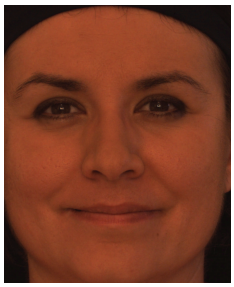

AU2

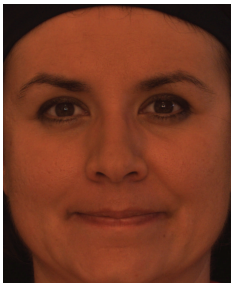
AU4

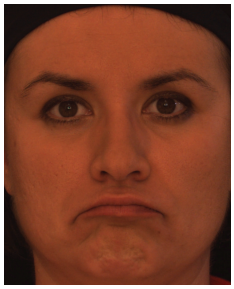
AU7

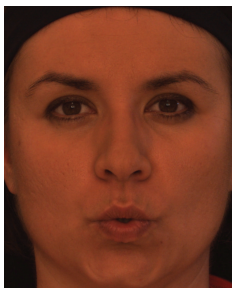
AU9

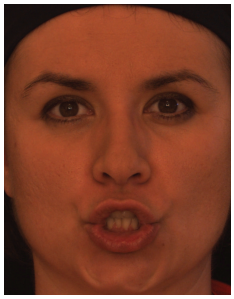

AU10

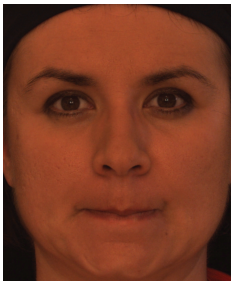

AU12

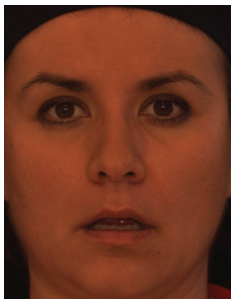

AU14

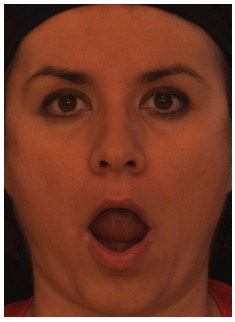

AU17

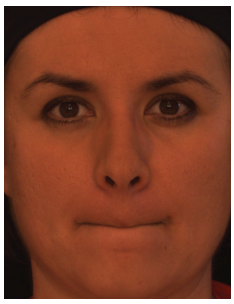

AU18


AU22


AU24


AU26


AU27


AU28


AU34


AU43

Figure 6.1: Examples of Facial AUs.

AU10 Upper Lip Raiser: The muscle underlying AU10 emerges from the center of the infraorbital triangle and attaches in the area of the nasolabial furrow. In AU10 the skin above the upper lip is pulled upwards and towards the cheek, pulling the upper lip up. It raises the upper lip, where center of upper lip is drawn straight up and the outer portions of upper lip are drawn up but not as high as the center. It pushes the infraorbital triangle up, widens the nostril wings and deepens the nasolabial furrow.

AU12 Lip Corner Puller: The muscle underlying AU 12 emerges high up in the lower face by the cheek bones and attaches at the corner of the lips. In AU12, the direction of the action is to pull the lip corners up towards the cheek bone in an oblique direction. It pulls the corners of the lips back and upward and deepens the nasolabial furrow by pulling it laterally and up. In a strong action, it bags the skin below the lower eyelid, narrows the eye aperture and produces crow's feet at eye corners.

AU14 Dimpler: The muscle underlying AU 14 emerges far back in the cheek bones and attaches in the center portion of the lips. In AU14 the skin beyond the lip corners is pulled inwards towards the lip corners, which are themselves drawn somewhat towards the ears. It tightens the corners of the mouth, pulling the corners somewhat inwards, and narrowing the lip corners. It also produces wrinkles and/or a bulge at the lip corner and pulls the skin below the lip corners and the chin boss up towards the lip corners, flattening and stretching the chin boss skin.

AU17 Chin Raiser: The muscle underlying AU 17 emerges from an area below the lower lip and attaches far down the chin. In AU 17 the skin of the chin is pushed upwards, pushing up the lower lip. It pushes the chin boss and the low lip upward and may cause wrinkles to appear on the chin boss. It causes shape of mouth to appear an inverted - U shape.

AU18 Lip Puckerer: The muscle relevant to AU18 is located above and below the upper and lower lips. AU 18 draws the lips medially, pursing or puckering them, causing the lips to protrude. It pushes the lips of the mouth forward and pulls medially and de-elongates the mouth opening, making the mouth opening smaller and rounder, and the lips appear tight. It makes short wrinkles on the skin above

the upper lip and also may cause wrinkles on the skin below the lower lip, and wrinkles in the lips themselves.

AU22 Lip Funneler: It is based on the outer strands of the muscle that runs around the mouth. It pulls in medially on the lip corners and makes lips funnel outwards taking on the shape as though the person were saying the word flirt. It exposes the teeth, gums and more of the red parts of the lips.

AU24 Lip Presser: It is based on the inner portion of the muscle orbiting the mouth within the lips. The lips are pulled in medially and pressed together. It lowers the upper lip and raises the lower lip to a small extent, without pushing up the chin boss. It tightens and narrows the lips and may cause small lines or wrinkles to appear on the upper lip and a bulging of the skin above and/or below the lips.

AU26 Jaw Drop: It describes the limited opening of the oral cavity (i.e., teeth parting) that can be produced by relaxing the muscle that closes the jaw. In AU26, the mandible is lowered by relaxation so that separation of the teeth can at least be inferred. Mouth appears as if jaw has dropped or fallen with no sign of the jaw being pulled open or stretching of the lips due to opening the jaw wide.

AU27 Mouth Stretch: AU 27 measures the forced opening and stretching of the mouth by muscles that act in opposition to muscles that close the jaw. It pull down the mandible and open the mouth quite far, changing the shape of the mouth opening from an oval with the long axis in the horizontal plane to one in the vertical direction. It flattens and stretched cheeks and changes shape of skin on the chin boss and the appearance under the chin.

AU28 Lips Suck: It involves the orbital muscles surrounding the mouth and lips. In AU28 the lips are pulled into the mouth. This movement can involve only the upper or lower lip. It sucks the red parts of the lips causing the red parts to disappear and adjacent skin into the mouth, covering the teeth. It stretches the skin above and below the lips and flattens the chin boss.

AU34 Puff: The cheeks puff out as air is forced into the mouth, but the lips remain closed keeping the air in.

AU43 Eye Closure: The same muscle, which when contracted raises the upper eyelid and when partially relaxed lets it droop, allows the eye to close when totally

relaxed. In AU43, the eyelid droops down reducing the eye aperture and more surface of the upper eyelid is exposed than usual.

6.2 Translating AU Scores Into Emotion Terms

EMOTION	PROTOTYPES	MAJOR VARIANTS
Surprise	1+2+5B+26 1+2+5B+27	1+2+5B 1+2+26 1+2+27 5B+26 5B+27
Fear	1+2+4+5*+20*+25, 26, or 27 1+2+4+5*+25, 26, or 27	1+2+4+5*+L or R20*+25, 26, or 27 1+2+4+5* 1+2+5Z, with or without 25, 26, 27 5*+20* with or without 25, 26, 27
Happy	6+12* 12C/D	
Sadness	1+4+11+15B with or without 54+64 1+4+15* with or without 54+64 6+15* with or without 54+64	1+4+11 with or without 54+64 1+4+15B with or without 54+64 1+4+15B+17 with or without 54+64 11+15B with or without 54+64 11+17
	25 or 26 may occur with all prototypes or major variants	
Disgust	9 9+16+15, 26 9+17 10* 10*+16+25, 26 10+17	
Anger	4+5*+7+10*+22+23+25,26 4+5*+7+10*+23+25,26 4+5*+7+23+25, 26 4+5*+7+17+23 4+5*+7+17+24 4+5*+7+23 4+5*+7+24	Any of the prototypes without any one of the following AUs: 4, 5, 7, or 10.

Table note: * means in this combination the AU may be at any level of intensity.

Figure 6.2: Emotion predictions based on AUs [Ekman *et al.* 2002].

Some of AU combinations can be converted into emotions using high level decision making rules. Fig. 6.2 cites the Table 10-1 in the FACS Investigator's Guide [Ekman *et al.* 2002] demonstrating some prototypes and major variants of AU combinations corresponding to the six universal emotions.

Excluded from the table are dozens of minor variants for each of the emotions, AU combinations for variations in the intensity of each emotion, and AU combinations for blends of two or more emotions [Ekman *et al.* 2002].

Publications

The results obtained during my PhD study have been the subject of five publications in international conferences and one in a national conference. Moreover, two journal paper have been submitted.

International Conferences:

1. X. Zhao, E. Dellandréa, L. Chen: A People Counting System based on Face Detection and Tracking in a Video, 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Genova, pp. 67-72, ISBN 978-1-4244-4755-8, 2009;

2. X. Zhao, E. Dellandréa, L. Chen: A 3D statistical facial feature model and its application on locating facial landmarks, Advanced Concepts for Intelligent Vision Systems (ACIVS 2009), Bordeaux, pp. 686-697, ISBN 978-3-642-04696-4, ISSN 0302-9743, 2009;

3. X. Zhao, P. Szeptycki, E. Dellandréa, L. Chen: Precise 2.5D Facial Landmarking via an Analysis by Synthesis approach, 2009 IEEE Workshop on Applications of Computer Vision (WACV 2009), Snowbird, Utah, pp. 1-7, ISBN 978-1-4244-5497-6, ISSN 1550-5790, 2009;

4. X. Zhao, D. Huang, E. Dellandréa, L. Chen: Automatic 3D facial expression recognition based on a Bayesian Belief Net and a Statistical Facial Feature Model, International Conference on Pattern Recognition, 2010.

5. X. Zhao, E. Dellandréa, L. Chen, D. Samaras: AU Recognition on 3D Faces Based On An Extended Statistical Facial Feature Model, IEEE Fourth International Conference on Biometrics: Theory, Applications and Systems, to appear, 2010.

National Conferences:

1. X. Zhao, E. Dellandréa, L. Chen: Multiple Face Tracking for People Counting, CORESA, Toulouse, pp. 192-196, 2009.

Submissions to International Journals:

1. X. Zhao, E. Dellandréa, L. Chen: Precise landmarking on 3D faces with expression and occlusion based on a 3D statistical facial feature model, submitted to IEEE Transaction on SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS.

2. X. Zhao, E. Dellandréa, L. Chen: An Unified Probabilistic Framework for Automatic 3D Facial Expression Analysis based on a Bayesian Belief Inference and Statistical Feature Models, submitted to Pattern Recognition.

Bibliography

[Abboud *et al.* 2004] B. Abboud, F. Davoine and M. Dang. *Facial expression recognition and synthesis based on an appearance model.* Signal Processing: Image Communication, vol. 19, no. 8, pages 723–740, 2004. 86

[Akakin *et al.* 2006] H. I. Akakin, A. A. Salah, L. Akarun and B. Sankur. *2D/3D Facial Feature Extraction.* Proceedings of the SPIE, vol. 6064, pages 441–452, 2006. 22

[Akakin *et al.* 2007] H. C. Akakin, L. Akarunb and B. Sankur. *Robust 2D/3D Face Landmarking.* Proceedings of 3DTV Conference, pages 1–4, 2007. 12, 13, 15

[Ari *et al.* 2008] I. Ari, A. Uyar and L. Akarun. *Facial feature tracking and expression recognition for sign language.* Computer and Information Sciences, 2008. ISCIS '08. 23rd International Symposium on, pages 1 –6, oct. 2008. 80, 83

[Asbach *et al.* 2008] M. Asbach, P. Hosten and M. Unger. *An Evaluation of Local Features for Face Detection and Localization.* Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on, pages 32 –35, may 2008. 12, 14

[Asthana *et al.* 2009] A. Asthana, J. Saragih, M. Wagner and R. Goecke. *Evaluating AAM fitting methods for facial expression recognition.* In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pages 1 –8, 10-12 2009. 83

[Azarbayejani & Pentland 1995] A. Azarbayejani and A. P. Pentland. *Recursive Estimation of Motion, Structure, and Focal Length.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 6, pages 562–575, 1995. 142

[Bai *et al.* 2009] Gang Bai, Wanhong Jia and Yang Jin. *Facial Expression Recognition Based on Fusion Features of LBP and Gabor with LDA.* In Image and Signal Processing, 2009. CISP '09. 2nd International Congress on, pages 1 –5, 17-19 2009. 82, 88

[Bailly *et al.* 2009] Kevin Bailly, Maurice Milgram and Philippe Phothisane. *Head Pose Estimation by a Stepwise Nonlinear Regression.* International Conference on Computer Analysis of Images and Patterns, pages 25–32, 2009. 95

[Bartlett *et al.* 2006] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel and J. R. Movellan. *Automatic recognition of facial actions in spontaneous expressions.* Journal of Multimedia, vol. 1, no. 6, pages 22–35, 2006. 80

[Berretti *et al.* 2010] S. Berretti, A. Del Bimbo, P. Pala amd B. Ben Amor and M. Daoudi. *A Set of Selected SIFT Features for 3D Facial Expression Recognition.* International Conference on Pattern Recognition (ICPR), 2010. 84

[Besl & Jain 1986] Paul J. Besl and Ramesh C. Jain. *Invariant surface characteristics for 3D object recognition in range images.* Computer Vision, Graphics, and Image Processing, vol. 33, no. 1, pages 33–80, 1986. xi, 89, 92, 95

[Beumer *et al.* 2006] G.M. Beumer, Q. Tao, A.M. Bazen and R.N.J. Veldhuis. *A landmark paper in face recognition.* Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, pages 69 –78, april 2006. 12, 13, 14, 15

[Bevilacqua *et al.* 2008] V. Bevilacqua, P. Casorio and G. Mastronardi. *Extending Hough Transform to a Points Cloud for 3D-Face Nose-Tip Detection.* Proceedings of International Conference of Advanced Intelligent Computing Theories and Applications, pages 1200–1209, 2008. 19, 23

[Blanz & Vetter 2003] V. Blanz and T. Vetter. *Face Recognition Based on Fitting a 3D Morphable Model.* IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pages 1063–1074, 2003. 86

[Boehnen & Russ 2004] C. Boehnen and T. Russ. *A fast multi-modal approach to facial feature detection.* Proceedings of Workshop on Applications of Computer Vision, 2004. 21, 24

[Bowyer *et al.* 2006] K.W. Bowyer, K. Chang and P. Flynn. *A Survey of Approaches and Challenges in 3D and Multi-Modal 3D+2D Face Recognition.* Journal of Computer Vision and Image Understanding, vol. 101, no. 1, pages 1–15, 2006. 9

[Breitenstein *et al.* 2008] M. D. Breitenstein, D. Kuettel, T. Weise, L. V. Gool and H. Pfister. *Real-time face pose estimation from single range images.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008. 95, 102

[Brick *et al.* 2009] T.R. Brick, M.D. Hunter and J.F. Cohn. *Get the FACS fast: Automated FACS face analysis benefits from the addition of velocity.* In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pages 1 –7, 10-12 2009. 80, 81, 83

[Celiktutan *et al.* 2008] O. Celiktutan, H.C. Akakin and B. Sankur. *Multi-attribute robust facial feature localization.* Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –6, sept. 2008. 12, 14, 15, 18

[Chakraborty *et al.* 2009] A. Chakraborty, A. Konar, U.K. Chakraborty and A. Chatterjee. *Emotion Recognition From Facial Expressions and Its Control Using Fuzzy Logic.* Systems, Man and Cybernetics, Part A: Systems and

Humans, IEEE Transactions on, vol. 39, no. 4, pages 726 –743, july 2009. 81, 83

[Chan *et al.* 2007] C. Chan, J. Kittler and K. Messer. *Multi-scale Local Binary Pattern Histograms for Face Recognition.* Proceedings of International Conference on Advances in Biometrics, pages 809–818, 2007. 120, 121

[Chan *et al.* 2008] A. B. Chan, Z. S. J. Liang and N. Vasconcelos. *Privacy Preserving Crowd Monitoring: Counting People without People Models or Tracking.* International Conference on Computer Vision and Pattern Recognition, pages 1–7, 2008. 138

[Chang & Lin 2001] C. Chang and C. Lin. *LIBSVM: a library for support vector machines.* Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001. 100, 126

[Chang *et al.* 2006] K. I. Chang, K. W. Bowyer and P. J. Flynn. *Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, pages 1695–1700, 2006. 19

[Chang *et al.* 2009a] Chuan-Yu Chang, Jeng-Shiun Tsai, Chi-Jane Wang and Pau-Choo Chung. *Emotion recognition with consideration of facial expression and physiological signals.* Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB '09. IEEE Symposium on, pages 278 –283, 30 2009-april 2 2009. 80, 83

[Chang *et al.* 2009b] Kai-Yueh Chang, Tyng-Luh Liu and Shang-Hong Lai. *Learning partially-observed hidden conditional random fields for facial expression recognition.* In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 533 –540, 20-25 2009. 83, 88

[Colbry & Stockman 2007] D. Colbry and G. Stockman. *Canonical Face Depth Map: A Robust 3D Representation for Face Verification.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1–7, 2007. 20

[Colbry *et al.* 2005] D. Colbry, G. Stockman and J. Anil. *Detection of Anchor Points for 3D Face Verification.* Proceedings of Computer Vision and Pattern Recognition - Workshops, pages 118–118, 2005. 20, 23, 41, 66

[Colombo *et al.* 2006] A. Colombo, C. Cusano and R. Schettini. *3D face detection using curvature analysis.* Journal of Pattern Recognition, vol. 39, no. 3, pages 444–455, 2006. 20, 23

[Conde & Serrano 2005] C. Conde and A. Serrano. *3D Facial Normalization with Spin Images and Influence of Range Data Calculation over Face Verification.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 20–26, 2005. 19

[Cootes & C.J.Taylor 2004] T.F. Cootes and C.J.Taylor. *Statistical Models of Appearance for Computer Vision.* Technical Report: University of Manchester, 2004. 31

[Cootes *et al.* 1995] T. F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham. *Active shape models - their training and application.* Computer Vision and Image Understanding, vol. 61, pages 38–59, 1995. 15, 18, 29, 47

[Cootes *et al.* 1998] T.F. Cootes, G.J. Edwards and C.J. Taylor. *Active appearance models.* European Conference on Computer Vision, vol. 2, 1998. 16, 18

[Cootes *et al.* 2001] T. F. Cootes, G. J. Edwards and C. J. Taylor. *Active appearance models.* IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pages 681–685, 2001. 23

[Cristinacce & Cootes 2008] D. Cristinacce and T.F. Cootes. *Automatic Feature Localisation with Constrained Local Models.* journal of Pattern Recognition, vol. 41, no. 10, pages 3054–3067, 2008. 17, 18, 23, 46

[Cui *et al.* 2008] Y. Cui, L. Sun and S. Yang. *Pedestrian detection using improved Histogram of Oriented Gradients.* 5th International Conference on Visual Information Engineering, pages 388 –392, jul. 2008. 138

[Dalal & Triggs 2005] N. Dalal and B. Triggs. *Histograms of Oriented Gradients for Human Detection.* International Conference on Computer Vision and Pattern Recognition, pages 886–893, 2005. 138

[Darwin 1872] C. Darwin. *The expression of the emotions in man and animals.* J. Murray, London, 1872. 73

[Database a] MIT Face Database. ftp://whitechapel.media.mit.edu/pub/images/. 12

[Database b] Yale Face Database. http://cvc.yale.edu/projects/yalefaces/yalefaces.html. 12

[Datcu & Rothkrantz 2004] D. Datcu and L.J.M. Rothkrantz. *Automatic recognition of facial expressions using Bayesian belief networks.* International Conference on Systems, Man and Cybernetics, 2004. 109

[D'House *et al.* 2007] J. D'House, J. Colineau, C. Bichon and B. Dorizzi. *Precise localization of landmarks on 3d faces using gabor wavelets.* International Conference on Biometrices: Theory, Applications, and Systems, pages 1–6, 2007. 20, 23, 41, 66

[Dibeklioglu *et al.* 2008] H. Dibeklioglu, A. A. Salah and L. Akarun. *3D Facial Landmarking under Expression, Pose, and Occlusion Variations.* Proceedings of the International Conference on Biometrics: Theory, Applications and Systems, pages 1–6, 2008. 21, 23, 24, 66, 68

[Dorai & Jain 1997] C. Dorai and A. Jain. *Cosmos - a representation scheme for 3d free-form objects.* IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 19, no. 10, pages 1115–1130, 1997. 89

[Dryden & Mardia 1998] I. Dryden and K.V. Mardia. *Statistical Shape Analysis.* Wiley, London, 1998. 31

[Duda *et al.* 2000] R. O. Duda, P. E. Hart and D. G. Stork. Pattern classification (2nd edition). Wiley-Interscience, 2000. 47, 110, 113

[Ekman & Friesen 1971] P. Ekman and W. V. Friesen. *Constants Across Cultures in the Face and Emotion.* journal of Personality and Social Psychology, vol. 17, no. 2, pages 124–129, 1971. 78

[Ekman & Friesen 1978] P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists, 1978. 78

[Ekman *et al.* 2002] P. Ekman, W. V. Friesen and J. C. Haper. *Facial Action Coding System, The Manual on CD ROM.* 2002. xi, xii, 78, 164

[Faling *et al.* 2009] Yi Faling, Xiong Wei, Huang Zhanpeng and Zhao Jie. *A Multi-view Nonlinear Active Shape Model Based on 3D Transformation Shape Search.* Information Assurance and Security, 2009. IAS '09. Fifth International Conference on, vol. 2, pages 15 –18, aug. 2009. 16

[Faltemier *et al.* 2008] T.C. Faltemier, K. W. Bowyer and P.J. Flynn. *Rotated profile signatures for robust 3d feature detection.* Proceedings of the International Conference on Face and Gesture Recognition, 2008. 41, 66

[Farkas 1994] L.G. Farkas. *Anthropometry of the Head and Face.* Raven Press, vol. 2nd edition, 1994. 9

[Fasel & Luettin 2003] B. Fasel and J. Luettin. *Automatic Facial Expression Analysis: A survey.* Pattern Recognition, vol. 36, no. 1, pages 259–275, 2003. xi, 73, 74, 82

[Feris *et al.* 2002] R. S. Feris, J. Gemmell, K. Toyama and V. Krger. *Hierarchical Wavelet Networks for Facial Feature Localization.* 5th IEEE International Conference on Automatic Face and Gesture Recognition, pages 125–130, Washington DC, USA, 2002. 12, 13

[Gokberk *et al.* 2008] B. Gokberk, H. Dutagaci, A. Ulas, L. Akarun and B. Sankur. *Representation Plurality and Fusion for 3-D Face Recognition.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 38, no. 1, pages 155–173, 2008. 10

[Greenwald *et al.* 1989] M. Greenwald, E. Cook and P. Lang. *Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli.* journal of Psychophysiol, vol. 3, pages 51 – 64, 1989. 75

[Griffith] S. Griffith. *http://www.case.edu/pubs/cnews/2002/1-17/emotion.htm*. 75

[Gunes & Piccardi 2009] H. Gunes and M. Piccardi. *Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 39, no. 1, pages 64 –84, feb. 2009. 80, 83

[Hammal *et al.* 2007] Z. Hammal, L. Couvreur, A. Caplier and M. Rombaut. *Facial Expression Classification: An Approach based on the Fusion of Facial Deformation using the Transferable Belief Model*. International journal of Approximate Reasoning, 2007. 81, 83

[Hanif *et al.* 2008] Shehzad Muhammad Hanif, Lionel Prevost, Rachid Belaroussi and Maurice Milgram. *Real-time facial feature localization by combining space displacement neural networks*. Pattern Recognition Letter, vol. 29, no. 8, pages 1094–1104, 2008. 14

[Harasse *et al.* 2005] S. Harasse, L. Bonnaud and M. Desvignes. *People counting in transport vehicles*. International Conference on Pattern Recognition and Computer Vision, pages 221–224, 2005. 138

[He *et al.* 2009] Languang He, Xuan Wang, Chenglong Yu and Kun Wu. *Facial expression recognition using embedded Hidden Markov Model*. In Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, pages 1568 –1572, 11-14 2009. 80, 82, 88

[Hou *et al.* 2001] XinWen Hou, S.Z. Li, HongJiang Zhang and QianSheng Cheng. *Direct appearance models*. Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, pages I–828 – I–833 vol.1, 2001. 17

[Hu *et al.* 2004] Y. Hu, D. Jiang, S. Yan, L. Zhang and H. Zhang. *Automatic 3D Reconstruction for Face Recognition*. International Conference on Face and Gesture, 2004. 86

[Hu *et al.* 2008a] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou and T.S. Huang. *Multi-view facial expression recognition*. IEEE International Conference on Automatic Face and Gesture Recognition, pages 1–6, 2008. 84, 107

[Hu *et al.* 2008b] Yuxiao Hu, Zhihong Zeng, Lijun Yin, Xiaozhou Wei, Jilin Tu and T.S. Huang. *A study of non-frontal-view facial expressions recognition*. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1 –4, 8-11 2008. 82

[Huang *et al.* 2007] Yuchi Huang, Qingshan Liu and D. Metaxas. *A Component Based Deformable Model for Generalized Face Alignment*. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1 –8, oct. 2007. ix, 10, 16

[Huertas & Pears 2008] M. R. Huertas and N. Pears. *3D Facial Landmark Localisation by Matching Simple Descriptors.* Proceedings of the International Conference on Biometrics: Theory, Applications and Systems, pages 1–6, 2008. 19

[Hutton *et al.* 2003] T. Hutton, B. Buxton and P. Hammond. *Automated registration of 3D faces using dense surface models.* Proceedings of British Machine Vision Conference, pages 439–448, 2003. 19

[Irfanoglu *et al.* 2004] M.O. Irfanoglu, B. Gokberk and L. Akarun. *3d shape-based face recognition using automatically registered facial surfaces.* Proceedings of the 17th International Conference on Pattern Recognition, vol. 4, pages 183–186, 2004. 19

[Jahanbin *et al.* 2008a] S. Jahanbin, A. C. Bovik and H. Choi. *Automated facial feature detection from portrait and range images.* Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation, 2008. 22, 66

[Jahanbin *et al.* 2008b] S. Jahanbin, H. Choi, R. Jahanbin and A. C. Bovik. *Automated Facial Feature Detecton and Face Recognition Using Gabor Features on Range and Protrait Images.* Proceedings of 15th IEEE International Conference on Image Processing, pages 2768–2771, 2008. 22, 24

[Jia & Gong 2005] K. Jia and S. Gong. *Multi-Modal Tensor Face for Simultaneous Super-Resolution and Recognition.* International Conference on Computer Vision, pages 1683–1690, 2005. 158

[Johnson 1997] A. Johnson. *Spin-images: A representation for 3-D surface matching.* Ph.D. dissertation, vol. Robotics Institute, no. Carnegie Mellon University, page Pittsburgh, 1997. 19

[Kakadiaris *et al.* 2007] I. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, Y. Lu, N. Karampatziakis and T. Theoharis. *3D face recognition in the presence of facial expressions: an annotated deformable model approach.* IEEE Transaction of Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pages 640–649, 2007. 10, 19, 23, 86

[Kanade *et al.* 2000] T. Kanade, J. F. Cohn and Y. Tian. *Comprehensive database for facial expression analysis.* the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pages 46–53, 2000. 12

[Kim & Bien 2008] Dae Jin Kim and Zeungnam Bien. *Design of "Personalized" Classifier Using Soft Computing Techniques for "Personalized" Facial Expression Recognition.* Fuzzy Systems, IEEE Transactions on, vol. 16, no. 4, pages 874 –885, aug. 2008. 83

[Kim & Dahyot 2008] Donghoon Kim and R. Dahyot. *Face Components Detection Using SURF Descriptors and SVMs.* Machine Vision and Image Processing Conference, 2008. IMVIP '08. International, pages 51 –56, sept. 2008. 12, 15

[Kim *et al.* 2002] J. W. Kim, K. S. Choi, B. D. Choi and S. J. Ko. *Real-time Vision-based People Counting System for the Security Door*. International Technical Conference On Circuits Systems Computers and Communications, 2002. 138

[Kinoshita *et al.* 2006] K. Kinoshita, Y. Ma, S. Lao and M. Kawaade. *A fast and robust 3D head pose and gaze estimation system*. In ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces, pages 137–138, New York, NY, USA, 2006. 95

[Koelstra & Pantic 2008] S. Koelstra and M. Pantic. *Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics*. In Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –8, 17-19 2008. 83

[Kotsia & Pitas 2007] I. Kotsia and I. Pitas. *Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines*. Image Processing, IEEE Transactions on, vol. 16, no. 1, pages 172 –187, jan. 2007. 83

[Koudelka *et al.* 2005] M. L. Koudelka, M. W. Koch and T. D. Russ. *A prescreener for 3d face recognition using radial symetry and the hausdorff fraction*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 3, pages 168–168, 2005. 21, 23, 68

[Koutlas & Fotiadis 2008] A. Koutlas and D.I. Fotiadis. *An automatic region based methodology for facial expression recognition*. Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, pages 662 –666, oct. 2008. 80, 83

[Kozakaya *et al.* 2008] T. Kozakaya, T. Shibata, M. Yuasa and O. Yamaguchi. *Facial feature localization using weighted vector concentration approach*. Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –6, sept. 2008. 17

[Li *et al.* 2002] P. Li, B.D. Corner and S. Paquette. *Automatic landmark extraction from three-dimensional head scan data*. Proceedings of SPIE, vol. 4661, page 169, 2002. 20

[Li *et al.* 2008] He Li, J.M. Buenaposada and L. Baumela. *Real-time facial expression recognition with illumination-corrected image sequences*. In Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –6, 17-19 2008. 83

[Li *et al.* 2009] Zisheng Li, Jun ichi Imai and M. Kaneko. *Facial-component-based bag of words and PHOG descriptor for facial expression recognition*. In Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, pages 1353 –1358, 11-14 2009. 82

[Li 2009] S. Z. Li. Markov random field modelling in image analysis(3rd edition). Springer, 2009. 46

[Lisetti & Nasoz 2002] C.L. Lisetti and F. Nasoz. *MAUI: A multimodal affective user interface*. Proceedings of International Conference on Multimedia, pages 161 – 170, 2002. 73

[Littlewort *et al.* 2006] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind and J. Movellan. *Dynamics of facial expression extracted automatically from video*. Image and Vision Computing, vol. 24, pages 615–625, 2006. 80

[Lu & Jain 2005] X. Lu and A.K. Jain. *Multimodal Facial Feature Extraction for Automatic 3D Face Recognition*. Technical Report MSU-CSE-05-22, vol. Michigan State University, 2005. 22, 24

[Lu & Jain 2006] X. Lu and A. Jain. *Automatic feature extraction for multiview 3D face recognition*. Proceedings of 7th International Conference Automated Face and Gesture Recognition, pages 585–590, 2006. 22, 24, 41, 66, 68

[Lu *et al.* 2004] X. Lu, D. Colbry and A. K. Jain. *Three-dimensional model based face recognition*. Proceedings of the 17th International Conference on Pattern Recognition, 2004. 20

[Lu *et al.* 2006] X. Lu, A. K. Jain and D. Colbry. *Matching 2.5D Face Scans to 3D Models*. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 28, no. 1, pages 31–43, 2006. 10, 22, 23

[Lyons *et al.* 1998] M. J. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba. *Coding Facial Expressions with Gabor Wavelets*. Third IEEE International Conference on Automatic Face and Gesture Recognition, pages 200–205, 1998. 12

[Mahoor *et al.* 2009] M.H. Mahoor, S. Cadavid, D.S. Messinger and J.F. Cohn. *A framework for automated measurement of the intensity of non-posed Facial Action Units*. Computer Vision and Pattern Recognition Workshop, vol. 0, pages 74–80, 2009. 80, 83

[Mardia & Dryden 1998] K.V. Mardia and I.L. Dryden. *Statistical Shape Analysis*. Wiley, Chichester, 1998. 9

[Martin & Gross 2008] Ch. Martin and H.-M. Gross. *A Real-time Facial Expression Recognition System based on Active Appearance Models using Gray Images and Edge Images*. International Conference on Automatic Face & Gesture Recognition, 2008. 86

[Martin *et al.* 2008] C. Martin, U. Werner and H.-M. Gross. *A real-time facial expression recognition system based on Active Appearance Models using gray images and edge images*. In Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –6, 17-19 2008. 83

[Martins & Batista 2009] P. Martins and J. Batista. *Identity and expression recognition on low dimensional manifolds*. In Image Processing (ICIP), 2009 16th IEEE International Conference on, pages 3341 –3344, 7-10 2009. 83

[Mehta & Stonham 1996] P. A. Mehta and T. J. Stonham. *A system for counting people in video images using neural networks to identify the background scene.* Pattern Recognition, vol. 29, no. 8, pages 1421–1428, 1996. 137

[Mpiperis *et al.* 2008] I. Mpiperis, S. Malassiotis and M.G. Strintzis. *Bilinear models for 3-D face and facial expression recognition.* IEEE Transaction on Information Forensics and Security, vol. 3, no. 3, pages 498–511, 2008. 74, 85, 86, 87, 108, 130, 131

[Nair & Cavallaro 2009] P. Nair and A. Cavallaro. *3-D Face Detection, Landmark Localization, and Registration Using a Point Distribution Model.* IEEE Transaction on Multimedia, vol. 11, no. 4, pages 611–623, 2009. 19, 23, 24, 68

[Nelder & Mead 1965] J.A. Nelder and R. Mead. *A simplex method for function minimization.* Computer journal, vol. 7, pages 308–313, 1965. 36, 51

[Niese *et al.* 2008] R. Niese, A. A. Hamadi, F. Aziz and B. Michaelis. *Robust Facial Expression Recognition Based on 3-d Supported Feature Extraction and SVM Classification.* Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pages 1–7, 2008. 10, 83

[Obaid *et al.* 2009] M. Obaid, R. Mukundan, R. Goecke, M. Billinghurst and H. Seichter. *A Quadratic Deformation Model for Facial Expression Recognition.* Digital Image Computing: Techniques and Applications, 2009. DICTA '09., pages 264 –270, dec. 2009. 80, 83

[Orozco *et al.* 2008] J. Orozco, O. Rudovic, F.X. Roca and J. Gonzalez. *Confidence assessment on eyelid and eyebrow expression recognition.* In Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –8, 17-19 2008. 83

[Ortony & Tumer 1990] A. Ortony and T. J. Tumer. *What's basic about basic emotions?* Psychological Review, vol. 97, pages 315 – 331, 1990. vii, 76

[O'Toole *et al.* 2005] A.J. O'Toole, J. Harms, S.L. Snow, D.R. Hurst, M.R. Pappas, J.H. Ayyad and H. Abdi. *A Video Database of Moving Faces and People.* PAMI, vol. 27, no. 5, pages 812–816, May 2005. 12

[Pantic & Rothkrantz 2000] M. Pantic and L.J.M. Rothkrantz. *Automatic Analysis of Facial Expresssions : The State of the Art.* IEEE Transaction Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pages 1424–1445, 2000. 82

[Park & Kim 2008] Sungsoo Park and Daijin Kim. *Spontaneous facial expression classification with facial motion vectors.* Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pages 1 –6, sept. 2008. 80, 83

[Park & Shin 2008] Sang-Jun Park and Dong-Won Shin. *3D face recognition based on feature detection using active shape models.* Control, Automation and Systems, 2008. ICCAS 2008. International Conference on, pages 1881 –1886, oct. 2008. 16

[Park *et al.* 2008] Sungsoo Park, Jongju Shin and Daijin Kim. *Facial expression analysis with facial expression deformation.* Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1 –4, dec. 2008. 80, 83

[Phillips *et al.* 1998] P.J. Phillips, H. Wechsler, J. Huang and P.J. Rauss. *The FERET Database and Evaluation Procedure for Face-Recognition Algorithms.* Image and Vision Computing, vol. 16, no. 5, pages 295–306, April 1998. 12

[Phillips *et al.* 2005] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek. *Overview of the Face Recognition Grand Challenge.* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pages 947–954, 2005. 18, 37

[Rabaud & Belongie 2006] V. Rabaud and S. Belongie. *Counting Crowded Moving Objects.* International Conference on Computer Vision and Pattern Recognition, pages 705 – 711, 2006. 138

[Ramanathan *et al.* 2006] S. Ramanathan, A. Kassim, Y.V. Venkatesh and W. S. Wah. *Human Facial Expression Recognition using a 3D Morphable Model.* IEEE International Conference on Image Processing, pages 661–664, 2006. 85, 86, 108

[Rosato *et al.* 2008] M. Rosato, X. Chen and L. Yin. *Automatic Registration of Vertex Correspondences for 3D Facial Expression Analysis.* International Conference on Biometrics: Theory, Applications and Systems, pages 1–7, 2008. 85, 86, 87, 108

[Rubner *et al.* 1998] Y. Rubner, C. Tomasi and L. J. Guibas. *A Metric for Distributions with Applications to Image Databases.* International Conference on Computer Vision, pages 59–66, 1998. 145

[Salah & Akarun 2006] A. A. Salah and L. Akarun. *3D Facial Feature Localization for Registration.* International Workshop on Multimedia Content Representation, Classification and Security, 2006. 22

[Salah *et al.* 2007] A. Ali Salah, H. Cinar, L. Akarun and B. Sankur. *Robust Facial Landmarking for Registration.* Annals of Telecommunications, vol. 62, no. 1-2, pages 1608–1633, 2007. 22, 23, 24

[Sandwell 1987] D. T. Sandwell. *Biharmonic Spline Interpolation of GEOS-3 and SEASAT Altimeter Data.* Geophysical Research Letters, vol. 14, no. 2, pages 139–142, 1987. 28

[Saragih *et al.* 2009] J. M. Saragih, S. Lucey and J. Cohn. *Face Alignment through Subspace Constrained Mean-Shifts.* International Conference of Computer Vision(ICCV), 2009. 17

[Savran & Sankur 2009] A. Savran and B. Sankur. *Automatic detection of facial actions from 3d data.* ICCV09: Workshop on Human Computer Interaction, 2009. 85, 126

[Savran *et al.* 2008] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur and L. Akarun. *Bosphorus Database for 3D Face Analysis.* The First COST 2101 Workshop on Biometrics and Identity Management, 2008. 19, 53, 84, 124

[Savran *et al.* 2010] A. Savran, B. Sankur and M. T. Bilge. *Facial action unit detection: 3d versus 2d modality.* CVPR10: Workshop on Human Communicative Behavior Analysis, 2010. 80, 81

[Schlögl *et al.* 2003] T. Schlögl, B. Wachmann, H. Bischof and W. Kropatsch. *People Counting In Complex Scenarios.* Technical report, pages 1–8, 2003. 137, 138

[Schmidt & Cohn 2001] K. L. Schmidt and J. F. Cohn. *Dynamics of facial expression: Normative characteristics and individual differences.* IEEE International Conference on Multimedia and Expo, pages 547–550, 2001. 77

[Seemann *et al.* 2004] E. Seemann, K. Nickel and R. Stiefelhagen. *Head pose estimation using stereo vision for human-robot interaction.* FG, 2004. 102

[Segundo *et al.* 2007] M. P. Segundo, C. Queirolo, O. R. P. Bellon and L. Silva. *Automatic 3D Facial Segmentation and Landmark Detection.* Proceedings of 14th International Conference on Image Analysis and Processing, pages 431–436, 2007. 20

[Shan & Gritti 2008] C. Shan and T. Gritti. *Learning discriminative LBP-histogram bins for facial expression recognition.* British Machine Vision Conference, 2008. 120

[Shang & Chan 2009] Lifeng Shang and Kwok-Ping Chan. *Nonparametric discriminant HMM and application to facial expression recognition.* In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2090 –2096, 20-25 2009. 83

[Shih & Chuang 2004] F. Y. Shih and C. Chuang. *Automatic Extraction of Head and Face Boundaries and Facial Features.* Information Sciences, no. 158, pages 117–130, 2004. 12

[Sim *et al.* 2003] T. Sim, S. Baker and M. Bsat. *The CMU Pose, Illumination, and Expression Database.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 1, pages 1615 – 1618, December 2003. 12

[Sohail & Bhattacharya 2007] A.S.M. Sohail and P. Bhattacharya. *Classifying facial expressions using point-based analytic face model and Support Vector Machines.* Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, pages 1008 –1013, oct. 2007. 80, 83

[Song *et al.* 2009] M. Song, D. Tao, Z. Liu, X. Li and M. Zhou. *Image Ratio Features for Facial Expression Recognition Application.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. PP, no. 99, pages 1 –1, 2009. 83

[Soyel & Demirel 2008] H. Soyel and H. Demirel. *3D facial expression recognition with geometrically localized facial features.* Symposium on Computer Science and Information Technology, pages 1–4, 2008. 84, 107, 131

[Sun & Xie 2008] Chengzhi Sun and Mei Xie. *An enhanced Active Shape Model for facial features extraction.* Communication Technology, 2008. ICCT 2008. 11th IEEE International Conference on, pages 661 –664, nov. 2008. 16

[Sun & Yin 2008] Y. Sun and L. Yin. *Facial Expression Recognition Based on 3D Dynamic Range Model Sequences.* Proceedings of The European Conference on Computer Vision, Marseille, 2008. 10

[Sun *et al.* 2008] Yi Sun, M. 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, 17-19 2008. 126

[Sung & Kim 2008] J. Sung and D. Kim. *Pose-Robust Facial Expression Recognition Using View-Based 2D 3D AAM.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 38, no. 4, pages 852 –866, july 2008. 83

[Suwa *et al.* 1978] M. Suwa, N. Sugie and K. Fujimora. *A Preliminary Note on Pattern Recognition of Human Emotional Expression.* The 4th International Joint Conference on Pattern Recognition, pages 408 – 410, 1978. 73

[Szeptycki *et al.* 2009] P. Szeptycki, M. Ardabilian and L. Chen. *A Coarse-to-Fine Curvature Analysis-based Rotation Invariant 3D Face Landmarking.* IEEE 3rd International Conference on Biometrics: Theory, Applications and Systems, Washington DC, 2009. ix, 20, 38, 39, 40, 41, 68, 89, 90

[Tai & Chung 2007] S.C. Tai and K.C. Chung. *Automatic facial expression recognition system using Neural Networks.* TENCON 2007 - 2007 IEEE Region 10 Conference, pages 1 –4, 30 2007-nov. 2 2007. 80, 83

[Talafova & Rozinaj 2007] R. Talafova and G. Rozinaj. *Face Feature Detection for 3D Model of Talking Head with Speech Synthesis.* Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on, pages 137 –139, june 2007. 12, 13

[Tang & Huang 2008a] H. Tang and T. S. Huang. *3D facial expression recognition based on automatically selected features.* workshop, International Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008. 84, 85, 104, 107, 130, 131

[Tang & Huang 2008b] H. Tang and T.S. Huang. *3D facial expression recognition based on properties of line segments connecting facial feature points.* IEEE International Conference on Automatic Face and Gesture Recognition, pages 1–6, 2008. 84, 85, 107, 131

[Teferi & Bigun 2007] D. Teferi and J. Bigun. *Damascening video databases for evaluation of face tracking and recognition - The DXM2VTS database.* Pattern Recognition Letter, vol. 28, no. 15, pages 2143–2156, 2007. 12

[Tong *et al.* 2007] Y. Tong, W. Liao and Q. Ji. *Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships.* IEEE Transaction of Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pages 1683–1699, 2007. 81, 82, 109

[Tong *et al.* 2010] Y. Tong, J. Chen and Q. Ji. *A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding.* IEEE Transaction of Pattern Analysis and Machine Intelligence, vol. 32, no. 2, pages 258–273, 2010. 80, 81, 82, 88, 109, 126

[Tsalakanidou & Malassiotis 2009] F. Tsalakanidou and S. Malassiotis. *Robust facial action recognition from real-time 3D streams.* In Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, pages 4 –11, 20-25 2009. 83

[Tsishkou *et al.* 2004] D. Tsishkou, L. Chen and E.Bovbel. *Semi-automatic Face Segmentation for Face Detection in Video.* International Conference on Intelligent Access to Multimedia Documents on the Internet, pages 107–118, 2004. 138, 139

[Tu & Lien 2009] C.-T. Tu and J. J. J. Lien. *Automatic Location of Facial Feature Points and Synthesis of Facial Sketches Using Direct Combined Model.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. PP, no. 99, pages 1 –12, 2009. 17

[Uddin *et al.* 2009] M.Z. Uddin, J.J. Lee and T.-S. Kim. *An enhanced independent component-based human facial expression recognition from video.* Consumer Electronics, IEEE Transactions on, vol. 55, no. 4, pages 2216 –2224, november 2009. 80, 82

[Venkatesh *et al.* 2009] Y. V. Venkatesh, A. A. Kassim and O. V. R. Murthy. *A novel approach to classification of facial expressions from 3D-mesh datasets using modified PCA.* Pattern Recognition Letter, vol. 30, no. 12, pages 1128–1137, 2009. 84, 85, 87, 108, 131

[Viola & Jones 2002] P. Viola and M. Jones. *Fast and robust classification using asymmetric AdaBoost and a detector cascade.* Advances in Neural Information Processing System 14, pages 1311–1318, 2002. 37, 138, 139

[Vretos *et al.* 2009] N. Vretos, N. Nikolaidis and I. Pitas. *A model-based facial expression recognition algorithm using Principal Components Analysis.* In Image Processing (ICIP), 2009 16th IEEE International Conference on, pages 3301 –3304, 7-10 2009. 82

[Wang *et al.* 2002] Y. Wang, C. S. Chua and Y. K. Ho. *Facial feature detection and face recognition from 2D and 3D images.* Journal of Pattern Recognition Letters, vol. 23, no. 10, pages 1191–1202, 2002. 22, 24

[Wang *et al.* 2006] J. Wang, L. Yin, X. Wei and Y. Sun. *3D Facial Expression Recognition Based on Primitive Surface Feature Distribution.* International Conference on Computer Vision and Pattern Recognition, pages 1399–1406, 2006. vii, xi, 84, 85, 89, 93, 103, 104, 106, 107, 131

[Wang *et al.* 2009] Shuliang Wang, Xiao Feng, Hehua Chi and Xiuling Wang. *Localization and extraction on the eyes, nose and mouth of human face.* Granular Computing, 2009, GRC '09. IEEE International Conference on, pages 561 –564, aug. 2009. 13, 18

[Whitehill *et al.* 2008] J. Whitehill, M. Bartlett and J. Movellan. *Automatic facial expression recognition for intelligent tutoring systems.* In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, pages 1 –6, 23-28 2008. 83

[Wieczorkowska *et al.* 2005] Alicja. Wieczorkowska, P. Synak, R. Lewis and W. Z. Ras. *Extracting Emotions from Music Data.* Proceedings of 15th International Symposium, pages 456–465, 2005. xi, 77

[Wright *et al.* 2009] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Yi Ma. *Robust Face Recognition via Sparse Representation.* IEEE Transaction of Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pages 210–227, 2009. 126, 127

[Xu & Ma 2008] Hua Xu and Zheng Ma. *An Improved Active Shape Model for Facial Feature Location.* Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium on, vol. 3, pages 114 –118, dec. 2008. 16, 18

[Xu *et al.* 2006] C. Xu, T. Tan, Y. Wang and L. Quan. *Combining Local Features for Robust Nose Location in 3D Facial Data.* Pattern Recognition Letters, vol. 27, no. 13, pages 1487–1494, 2006. 19, 23

[Xua *et al.* 2006] C. Xua, T. Tana, Y. Wang and L. Quanc. *Combining local features for robust nose location in 3D facial data.* Pattern Recognition Letters, vol. 27, no. 13, pages 1487–1494, 2006. 41, 66

[Yang *et al.* 2002] M.-H. Yang, D.J. Kriegman and N. Ahuja. *Detecting faces in images: a survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pages 34 –58, jan. 2002. 138

[Yang *et al.* 2007] Peng Yang, Qingshan Liu and D.N. Metaxas. *Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition.* In Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1 –6, 17-22 2007. 80, 83

[Yin *et al.* 2006] L. Yin, X. Wei, Y. Sun, J. Wang and M. Rosato. *A 3D Facial Expression Database For Facial Behavior Research.* IEEE International Conference on Automatic Face and Gesture Recognition, pages 211–216, 2006. 19, 53, 82, 124

[Yin *et al.* 2008] L. Yin, X. Chen, Y. Sun, T. Worm and Michael Reale. *A High-Resolution 3D Dynamic Facial Expression Database.* The 8th International Conference on Automatic Face and Gesture Recognition, pages 1–6, 2008. 84

[Yoshida *et al.* 2002] H. Yoshida, Y. Masutani, P. MacEneaney, D. Rubin and A. Dachman. *Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study.* Radiology, no. 222, pages 327–336, 2002. xi, 91

[Yu & Yan 2009] Weiwei Yu and Nannan Yan. *Facial Feature Extraction on Fiducial Points and Used in Face Recognition.* Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on, vol. 3, pages 274 –277, nov. 2009. 17

[Yun & Guan 2009] Tie Yun and Ling Guan. *Automatic fiducial points detection for facial expressions using scale invariant feature.* Multimedia Signal Processing, 2009. MMSP '09. IEEE International Workshop on, pages 1 –6, oct. 2009. 12, 14

[Zeng *et al.* 2007] Zhihong Zeng, Jilin Tu, Ming Liu, Thomas S. Huang, Brian Pianfetti, Dan Roth and Stephen Levinson. *Audio-Visual Affect Recognition.* Multimedia, IEEE Transactions on, vol. 9, no. 2, pages 424 –428, feb. 2007. 83

[Zeng *et al.* 2008] Wei Zeng, Yun Zeng, Yang Wang, Xiaotian Yin, Xianfeng Gu and Dimitris Samaras. *3D Non-rigid Surface Matching and Registration Based on Holomorphic Differentials.* European Conference on Computer Vision, pages 1–14, 2008. 10

[Zeng *et al.* 2009] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang. *A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions.* IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pages 39–58, 2009. 82

[Zhang & Zhang 2010] C. Zhang and Z. Zhang. *A Survey of Recent Advances in Face Detection*. Technical Report: MSR-TR-2010-66, 2010. 138

[Zhang 1994] Z. 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. 26

[Zhao & Pietikainen 2007] Guoying Zhao and M. Pietikainen. *Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 6, pages 915 –928, june 2007. 80, 83, 88

[Zhao *et al.* 2008] Quanyou Zhao, Baochang Pan, Shenglin Zheng and Jian Liang. *A new facial key features location algorithm in color images*. Signal Processing, 2008. ICSP 2008. 9th International Conference on, pages 932 –936, oct. 2008. 12

[Zhi *et al.* 2008] Ruicong Zhi, Qiuqi Ruan and Zhenjiang Miao. *Fuzzy discriminant projections for facial expression recognition*. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1 –4, 8-11 2008. 82

[Zhi *et al.* 2009] Ruicong Zhi, M. Flierl, Qiuqi Ruan and B. Kleijn. *Facial expression recognition based on graph-preserving sparse non-negative matrix factorization*. In Image Processing (ICIP), 2009 16th IEEE International Conference on, pages 3293 –3296, 7-10 2009. 82

[Zhu & Zhao 2009] Shaojun Zhu and Jieyu Zhao. *Facial Feature Points Extraction*. Image and Graphics, 2009. ICIG '09. Fifth International Conference on, pages 195 –199, sept. 2009. 17

[Zhu *et al.* 2009] Yunfeng Zhu, F. De la Torre, J.F. Cohn and Yu-Jin Zhang. *Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection*. In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pages 1 –8, 10-12 2009. 83