1-1-2011

Human emotional state recognition using 3D facial expression features

Yun Tie
Ryerson University

HUMAN EMOTIONAL STATE RECOGNITION USING 3D FACIAL EXPRESSION FEATURES

by

YUN TIE

Master of Science, Kwangju Institute of Science and Technology, Korea, 2001
Bachelor of Engineering, Nanjing University of Science and Technology, China, 1996

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirement for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2011

Author's Declaration

I hereby declare that I am the sole author of this thesis.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

_____

(Yun Tie)

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____

(Yun Tie)

Borrower's Page

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Name	Signature	Address	Date

ABSTRACT

HUMAN EMOTIONAL STATE RECOGNITION USING 3D FACIAL EXPRESSION FEATURES

© Yun Tie

Doctor of Philosophy, Electrical and Computer Engineering,
Ryerson University, Toronto, Canada, 2011

In recent years there has been a growing interest in improving all aspects of the interaction between human and computers. Emotion recognition is a new research direction in human-computer interaction (HCI) which is based on affective computing that is expected to significantly improve the quality of HCI system and communications. Most existing works address this problem using 2D features, but they are sensitive to head pose, clutter, and variations in lighting conditions. In light of such problems, two 3D visual feature based approaches are presented in this dissertation. First, we present a recognition method based on the Gabor library for real 3D visual features extraction and an improved kernel canonical correlation analysis (IKCCA) algorithm for emotion classification. Second, to reduce the computation cost and provide a more general approach, we propose using a fiducial points' controlled 3D face model to recognize human emotion from video sequences. An Elastic body spline **(**EBS**)** technique is applied for deformation feature extraction and a discriminative Isomap (D-Isomap) based classification is used for the final decision. The most significant contributions of this work are detecting and tracking fiducial points automatically from video sequences to construct a generic 3D face model, and the introduction of EBS deformation features for

emotion recognition. The experimental results show the robustness and effectiveness of the proposed methods.

Acknowledgments

First and foremost I am deeply indebted to my advisor Dr. Ling Guan for his continuous guidance and encouragement throughout my research. I am very grateful to Dr. Guan for his insightful suggestions, valuable discussions, and great attention to details. This work would have been impossible without his feedback, patience and kindness.

I would like to thank the Electrical and Computer engineering Department of Ryerson University for providing a very well equipped and technically supported Ryerson Multimedia Research Laboratory (RML). My thanks are due to the School of Graduate Studies of Ryerson University for providing Graduate Student Scholarship. And I am very thankful to my thesis committee members for their invaluable advice on the dissertation.

I express my appreciation to the members of Ryerson Multimedia Research Laboratory. I would like to thank my present and past officemates, Dr. Ivan Lee, Dr. Matthew Kyan, Dr. Yifen He, Dr. Yongjin Wang, and Chunhao Wang, Rui Zhang, Ning Zhang, Muhammad Talal Ibrahim, Dong Nan, Adrian Bulzacki, for their help and support. It is enjoyable and a great pleasure to work in such a friendly and collaborative environment. And I am very grateful to the ELS staff at Ryerson University for the structure and language improvement on the dissertation.

Finally, I would also like to thank my family, including my parents, my parents-in-law, Yang Rong, Jiessie and Kiven for their consistently supporting and encouragement throughout these years. I especially would like to thank my wife Yang Rong for her complete patience and love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

EMOTION plays a critical role in human-to-human interaction, allowing people to express themselves beyond the verbal domain and understand each other from various modalities. Some emotions motivate human actions, and others enrich the meaning of human communication. The human computing paradigm suggests that user interfaces of the future need to be anticipatory and human-centered, and based on naturally occurring human communication [1]. Human-centered interfaces must have the ability to detect subtleties of, and changes in, the user's behaviour, and to initiate interactions based on this information, rather than simply responding to the user's commands [2].

The ability to recognize the human emotional or affective state is desirable to empower the intelligent computer to interpret, understand, and respond to human emotions, moods, and, possibly, intentions, which is similar to the way that humans rely on their senses to assess each other's affective state [3]. Automatically recognizing the human emotional or affective state can enhance not only the computer's performances in detecting and sensing affective states of the human, but also the abilities to interpret and respond appropriately to the user's affective feedback. Presented in [4], as an affective computing, emotion recognition enables human–computer interaction (HCI) for more naturally and in a more friendly manner. Many potential applications such as intelligent automobile

systems, game and entertainment industries, interactive video, indexing and retrieval of images or video databases can be obtained from this ability.

Research in computer-vision based emotion recognition has expanded rapidly in recent years due to the advances in imaging technology and newfound interests in psychology. It is well known that the human face provides an important and spontaneous channel for the emotional states. It contains powerful, natural and immediate information for human-to-human communication and social life. Facial expression functions as a conversation enhancer, communicates feeling and a cognitive mental state, shows empathy and acknowledges the actions of other people. Contemporary research in psychology reveals that certain emotions were associated with distinct facial signals. Analyzing facial expression in real time, without human intervention, provides an efficient and robust approach to recognizing human's emotions.

In this dissertation, we explore two 3D feature based approaches for automatic emotion recognition, *i.e.* the 3D Gabor feature and the 3D elastic body spline **(**EBS**)** features from video sequences. These methods open new research directions for human computer communication with applications to security systems, the intelligent home, a learning environment, and educational software to name a few.

1.1 Research Challenge

The vision based emotion recognition studies mainly focus on facial expression analysis because of the importance of the face in emotion expression and perception. Many studies on the machine analysis of facial expressions have seen much progress in the past

decade. In spite of considerable previous work documented in this area, many challenges still remain. Since the faces are non-rigid and have a high degree of variability in location, color and pose, several features of the face that are not common to other pattern recognition issues make facial expressions based emotion recognition more complex. Occlusion and lighting distortions, as well as illumination conditions can also change the overall appearance of the face. Such changes will cause the large intra-class variations of the emotion distribution to be highly nonlinear and complex in any space.

The majority approaches for solving the human facial expressions recognition problem are based on 2D spatiotemporal data: either 2D static images or 2D video sequences. Few efforts have been investigated on 3D face data for the vision-based emotions recognition. The performance of 2D based algorithms remains unsatisfactory, and is often unreliable under adverse conditions. It is difficult to handle pose variations, lighting illumination and subtle facial behaviour. Therefore, the 2D based approaches are limited to constrained environment. To achieve more robust performance, a growing body of research has been focused on addressing the problem using 3D information.

Automatically analyzing facial expressions for human emotion recognition in video sequences is also a challenging problem due to the fact that current techniques for the detection and tracking of facial expressions are sensitive to head pose, clutter and

variations in lighting conditions. To recognize emotional states, the facial feature extraction attempts to find the most appropriate way to represent the facial expressions. The representations of expression patterns should be considered carefully since it is important for feature extraction and classification, which will strongly affect the performance of an emotion recognition system. A good representation should have such characteristics as small within-class variations, large between-class variations, and robust to transformations without changing the class labels. Furthermore, its extraction should not depend much on manual operation.

Context dependency of facial expressions is also a largely unexplored research area for current emotion recognition systems. It is a fact that most present approaches to automatic facial expression analysis are context insensitive. However, the interpretation of human emotional signals is context dependent. To interpret an emotional signal, it is important to know the context in which this signal has been displayed.

1.2 Objectives of the Research

Motivated by the aforementioned challenges in this field, the proposed work aims to improve human–computer interaction intelligent (HCII) techniques with two 3D feature

based approaches: 3D Gabor feature-based method and 3D EBS feature-based method from video sequences.

Our work consists of 3D Gabor library for real 3D visual features extraction from 3D geometric information plus color/density information of the facial expressions, automatic detecting and tracking from video sequences, 3D EBS mesh modeling, active deformation extraction, and intelligent pattern classification. An active deformation approach using 3D EBS features for facial expressions transformation forms a major research activity in this work. It is expected to benefit the advancement of computer vision techniques and the applications in communication and information technology. The presented methods can be implemented on, among others, a mobile security robot for the detection and recognition of dangerous and suspicious intention and activities at airports, subway stations and other places of national and military importance, or on a domestic helper for assisting elderly and/or disabled people at home or community houses.

Key research objectives of this work are summarized as follows:

1. *Real 3D visual features extraction.* Generally, the existing 3D-based methods consider only geometric information for feature extraction. In this work, we present a real 3D visual feature-based method for human emotion recognition. The 3D

geometric information plus colour/density information of the facial expressions are extracted by the 3D Gabor library to construct visual feature vectors. The filter's scale, orientation, and shape of the library are specified according to the appearance patterns of the 3D facial expressions. An improved kernel canonical correlation analysis (IKCCA) algorithm is proposed for emotion decision.

2. *Video/images processing with intelligent detecting and tracking methods.* Automatical and robust detecting and tracking for face and fiducial points are primary for 3D EBS based recognition. Existing methods leave uncertainties and difficulties in practice and for real time applications. We try to solve such problems and apply to on-line uses.

3. *3D data modeling.* We have developed a new framework to construct view-independent facial expression recognition based on a generic 3D face model, which is controlled by fiducial point to synthesize the animated facial expressions. Few attempts have been reported so far based on the 3D face model toward the vision-based emotion recognition. The proposed method can be robust to arbitrary head movement occlusions, scene complexity like the presence of other people and dynamic background. It can also form the first step in the realization of facial expressions analysis capable of handling unconstrained environments.

4. *Active deformation analysis.* The EBS technique is applied on the 3D face mesh model to generate a smooth warp that reflects control point and to extract the deformation feature of the realistic expression from the neutral face. Very few existing studies have been made of the context-dependent interpretation of the observed facial expressions. The proposed EBS method can function as an interpolation approach to generate corresponding intrinsic geometries of the facial expressions and investigate the interpretation of emotional space.

5. *Intelligent classifications*. The IKCCA and a discriminative Isomap (D-Isomap) are applied in classification. The IKCCA is used for Gabor feature-based recognition. The semantic ratings that best describe the different facial expressions are computed by the IKCCA to generate a seven-dimensional semantic expression vector. The correlation with different testing samples is learned for classifying the associated prototypic facial expression with the trained facial feature distribution. For the D-Isomap classification, the deformation features of facial expressions are embeded into the low dimensional manifold with seven class centres, which span in a face space with six emotions and neutral. The discriminative information of a facial feature is considered so that it can reflect appropriately the discriminative structures of the emotional space on the manifold.

1.3 Outline of Dissertation

As mentioned in the previous sections, two 3D emotion recognition methods are proposed in this dissertation. Chapter 3 presents the work for the Gabor-feature based method. Since the EBS feature-based method consists of several crucial components, it is arranged into three Chapters: Chapter 4 – automatic face detection, Chapter 5 – fiducial point detection and tracking, and Chapter 6 – 3D face modeling and recognition. The rest of this dissertation is organized as follows:

Chapter 2: *Background and Related Works*, reviews the background and related works reported in the literatures for vision-based human emotion recognition.

Chapter 3: *3D Gabor Based Recognition*, conducts a 3D Gabor library-based method for 3D facial expressions' recognition. The IKCCA is used for the classification.

Chapter 4: *Face Detection*, discusses the automatic face region detection using the local normalization and optimal adaptive correlation (OAC) analysis.

Chapter 5: *Fiducial Points' Detection and Tracking*, describes the scale invariant feature examination for fiducial points' detection and multiple Differential Evolution Markov Chain (DE-MC) particle filters for tracking.

Chapter 6: *3D EBS Based Recognition*, details the 3D face modeling and feature extraction with EBS techniques, a D-Isomap based method is also described for final emotion classification.

Chapter 7: *Conclusions and Future Works*, discusses the results from our experiment and summarizes the contributions of this work. Future works are also presented.

Chapter 2

Background and Related Work

2.1 Background Study

EMOTIONS have been the study of intense interest in philosophy since the fourth century B.C., and the beginning of modern and scientific inquiry into the nature of emotion is thought by many to have begun with Charles Darwin's study of emotional expression in animals and humans [5]. Research into human emotion in psychology and neurophysiology has grown rapidly in recent years. It is generally accepted that all emotions are processed by a circuit of interconnected brain structures known as the limbic system [6]. The basic emotions may be coded by partially distinct brain systems. Based on discoveries made through neural mapping of the limbic system, the neurobiological explanation of human emotion is that emotion is a pleasant or unpleasant mental state organized in the limbic system. These states are manifestations of non-verbally expressed feelings. For instance, the amygdala has been revealed to be playing a

significant role in recognizing facial and vocal expressions of fear [7]. A survey of research in psychology about defining, studying, and explaining emotion can be found in [8].

Two main methods are often used to describe emotions. One is to label the emotions in discrete categories, such as joy, fear, love, surprise, sadness. The main advantage of a category representation is that people use this categorical scheme to describe observed emotional displays in daily life. The labeling scheme based on category is very intuitive and thus matches people's experience. But the problem with this approach is that it may contain blended emotions. Discrete lists of emotions fail to describe the range of emotions that occur in natural communication settings. Also, the choice of words that can describe the wide variety of emotional displays may be too restrictive, or culturally dependent.

Another way is to have multiple dimensions or scales to describe emotions, where an emotional state is characterized in terms of a small number of latent dimensions rather than in terms of a small number of discrete emotion categories. Two common dimensions are valence and arousal that are expected to reflect the main aspects of emotion. The different emotional labels can be plotted at various positions on a two-dimensional plane spanned by these two axes to construct a 2D emotion model [9]. Scholsberg [10] also

suggested a three-dimensional model in which attention–rejection was in addition to the above two. In contrast to categorical representation, dimensional representation enables raters to label a range of emotions. However, the matching of the high-dimensional emotional states onto a rudimentary 2D space results in the loss of information. Some emotions become indistinguishable and some emotions lie outside the space. This representation is not intuitive, and raters need special training to use the dimensional labeling system.

Most of the existing human emotion recognition systems attempt to recognize prototypic emotions. The most important and widely accepted set of measurement is the so-called "six-basic" emotions: *happiness, sadness, anger, fear, surprise,* and *disgust*, which were pioneered by Ekman and Friesen [11]. According to [11], the "six-basic" emotions are not culturally determined, but universal to human culture and thus biological in origin. They also indicate that humans perceive certain basic emotions with respect to facial expressions. This influence of a basic emotion theory has resulted in the fact that many of the existing studies of automatic emotion recognition focus on recognizing these basic emotions. There are also several other emotions, and many combinations of emotions, which have been studied but remain unconfirmed as universally distinguishable. In our present work, this set of six emotions is analyzed.

2.2 Emotional Behaviours

Emotion is complex, hidden under the expressions. There are many emotional states: anger, anticipation, boredom, disgust, fear, regret, sadness and more. Also, they may be combined with each other in face to face chatting or communications. In psychology and common use, those are aspects of a human being's mental state, normally based on or tied to the person's internal (physical) and external (social) sensory feeling.

Emotions arise spontaneously rather than through conscious effort. Every emotion is actually a neural impulse that moves an organism to action, prompting automatic reactive behaviour that has been adapted through evolution as a survival mechanism to meet a survival need. So, emotions are expressed through physiological functions such as facial expressions, heartbeat, and affect behaviours such as aggression, crying, or covering the face with hands [12]. When fear, the body frequently responds to shame by warmth in the upper chest and face, by a heightened heartbeat, increased "flinch" response, and increased muscle tension. The sensations connected with anger are nearly indistinguishable from fear. Happiness is often felt as an expansive or swelling feeling in the chest and the sensation of lightness or buoyancy, as if standing underwater. Sadness

brings a feeling of tightness in the throat and eyes, and relaxation in the arms and legs. Desire can be accompanied by a dry throat and heavy breath.

2.3 Facial Expression Based Emotion Recognition

The main characteristics of human emotions are the multiplicity and multimodality of communication channels. The psychological studies also indicated that facial expression in the visual channel is a most natural and primary cue for communicating the quality and nature of emotions, and correlates well with the body and voice [13]. Each of the six basic emotions corresponds to a unique facial expression. To the objectives of an emotion recognition system, facial expression analysis is considered to be the major indicator of a human affective state [14].

Automatically recognizing human emotion from facial expressions is inherently a multidisciplinary enterprise involving different research fields [15], including psychology, computer vision, feature data fusion, and machine learning. There are two main streams in the current research on the machine analysis of facial expressions: the recognition of affect and the recognition of facial muscle action. The most commonly used vision-based coding system is the facial action coding system (FACS) proposed by Ekman and Friesen [16] for the manual labeling of facial behaviour. FACS is a comprehensive and

anatomically based system that is used to measure all visually discernible facial movements. To recognize emotions from facial clues, FACS enables facial expression analysis through standardized coding of changes in facial motion in terms of atomic facial actions called Action Units (AUs). The changes in the facial expression are described with FACS in terms of AUs. FACS decomposes the facial muscular actions into 44 basic actions and describes the facial expressions as combinations of the AUs. As AUs are independent of interpretation, they can be used for any high-level decision-making process, including the recognition of basic emotions, the recognition of various affective states and the recognition of other complex psychological states. AUs of the FACS are very suitable to use in studies on human naturalistic facial behaviour, as the thousands of anatomically possible facial expressions can be described as combinations of 27 basic AUs and a number of AU descriptors. This work inspired many researchers to analyze facial expressions.

Different methods have been explored so far to perform facial expression analysis, which can be roughly categorized into two groups: the geometric feature-based methods and appearance-based methods. The geometric facial feature-based methods present the shape, or location information of prominent components such as the mouth, eyes, nose, eyebrow, and chin, which can cover the variation in the appearance of the facial

expression. The appearance-based methods, on the other hand, using image filters such as Gabor wavelets, generate the facial feature for either the whole-face or specific regions in a face image. Pantic and Rothkrantz [17] proposed an automatic facial action recognition system using a dual-view static image. The face was detected using watershed segmentation with markers method, in which the markers were extracted based on the Hue Saturation Value (HSV) colour model. A multi-detector approach to facial feature localization was utilized to spatially sample the profile contour and the contours of the facial components such as the eyes and mouth. They reported an average recognition rate of 86% by classifying facial actions into a group of 32 individual facial muscle actions occurring along or in combination using rule-based reasoning. Lyons *et al.* [18] used a set of multi-scale, multi-orientation Gabor filters to transform the images first. A grid was then automatically registered with the face using an elastic graph matching method. The Gabor coefficients sampled on the grid were combined into one single vector as the features. Principal Components Analysis (PCA) was applied to reduce the dimensionality of the feature space. They tested their system with a database of 193 images posed by 9 Japanese females, and achieved 75% expression classification accuracy by using Linear Discriminant Analysis (LDA). Silva and Hui [19] determined the eye and lip position using low-pass filtering and an edge detection method. They achieved an average

emotion recognition rate of 60% using a neural network (NN). Cohen *et al.* [20] introduced and tested different classifiers for recognizing human facial expression from video sequences. A face tracking algorithm called Piecewise Bezier Volume Deformation tracker (PBVD) was used, and 12 motion units (MU) were extracted as the basic features for classification. They introduced a multi-level hidden Markov model (HMM) classifier for dynamic classification, in which the temporal information was also taken into account. Two types of Bayesian network classifiers, Naive Bayes (NB), and Tree-Augmented Naive Bayes (TAN), and neural network were investigated to perform classification on a single frame. A person-independent experiment using their own database showed that the TAN classifier gave the best correct recognition rate of 66.53%. Guo and Dyer [21] introduced a linear programming based method for face expression recognition with a small number of training images of each expression. A pairwise framework for feature selection, instead of using all classes simultaneously, was presented and three methods were compared in the experiment part. Pantic and Patras [22] presented a method to handle a large range of human facial behaviour by recognizing facial muscle actions that produce expressions. AUs and their temporal models were automatically recognized from long, profile-view face image sequences. The algorithm performed both automatic segmentation of an input video into facial

expressions pictured and recognition of temporal segments of 27 AUs occurring alone or in a combination in the input face-profile video. Anderson and McOwan [23] presented an automated multistage system for real-time recognition of facial expression. The system used facial motion to characterize monochrome frontal views of facial expressions and was able to operate effectively in cluttered and dynamic scenes, recognizing the six emotions universally associated with unique facial expressions. Gunes and Piccardi [24] proposed an automatic method for temporal segment detection and affect recognition from face and body display. Facial expressions and body gestures were detected from each individual frame and temporal segments were analyzed. Individual classifiers were separately trained from face and body features. Affective states were then addressed in two ways of sequence-based and frame-based detection from the bimodal analysis of a video. Wang and Guan [25] constructed a bimodal system for emotion recognition. They used a face detection scheme based on a HSV colour model to detect the face from the background and Gabor wavelet features to represent the facial expressions. They achieved the best overall recognition rate of 82.14% using the proposed multi-classifier scheme.

Many difficulties still remain in facial expression recognition techniques due to head pose, clutter, variations in lighting conditions, and the variation across the human

population and to the context-dependent variation even for the same individual. Traditionally, the majority approaches for human facial expression recognition try to analyze either 2D static images or 2D video sequences. Unfortunately, the performance of 2D based algorithms is unsatisfactory, and often proves unreliable under adverse conditions. It is difficult to handle pose variations, lighting illumination and subtle facial behavior. Therefore, the above method is limited to a constrained environment.

In order to achieving a more robust approach, some research has addressed the problem using 3D information for recognizing and understanding facial expressions. The analysis of 3D facial expressions will facilitate the examination of the fine structural changes inherent in the spontaneous expressions. The 3D based algorithm allows the transfer of feature-like models from the given single view into new arbitrary views, thus making the solution far more pose invariant than current 2D solution. Furthermore, 3D data is by definition lighting invariant, thus eliminating errors associated with changes in illumination becomes more tractable with knowledge of the physical surfaces being considered.

Song *et al.* [26] presented a generic facial expression analogy technique to transfer facial expressions between arbitrary 3D face models, as well as between 2D face images. Geometry encoding for triangle meshes, vertex-tent-coordinates were proposed to

formulate expression transfer in 2D and 3D cases as a solution to a simple system of linear equations. Hu *et al.* [27] proposed a work on the non-frontal-view facial expression analysis by generating a multi-view from 3D data. Geometric salient facial points were manually labeled and then the geometric displacement between emotional and neutral expressions was calculated for the person at the corresponding angle. Various classifiers were investigated and the experiments showed that the support vector machine (SVM) returned the best performance with the average error rate of 0.335. Chin *et al*. [28] presented an emotional intensity-based facial expression modeling process by generating 3D customized face and facial expressions. The generated customized face integrated expression data used different expression intensities. They identified six universal expressions by determining the anatomical, parametric values of linear and sphincter muscles. Facial expressions were also simulated by intensity mapping.

There are other limitations of the previous approach such as the lack of temporal and detailed spatial information in the visual cues both at local and global scales. They may be caused by the principle difficulties and the sheer complexities of describing human facial movement. Moreover, no effort in automatic detection and the segments based on the face components in image sequences with respect to emotion recognition has been reported so far. Lots of psychological researches support that the timing of expressions is

a critical parameter in recognizing emotions and the detailed spatial dynamic deformation of the expression is important in expression recognition.

2.4 Proposed Methods

2.4.1 3D Gabor Feature Based Recognition

We first present a new 3D emotion recognition method using geometric features and the colour/density information. The main contribution of this work consists of applying the 3D Gabor library for feature extraction and an IKCCA algorithm for final classification. To the best of our knowledge, no similar work has been reported with 3D Gabor library for emotion or face recognition. The block diagram of this method is shown in Figure 2.1. We extract primitive 3D facial expression feature vectors by using the 3D Gabor library. The filter's scale, orientation, and shape of the library are specified according to the appearance patterns of the 3D facial expressions. Then the IKCCA is proposed for the final decision. From training samples, the semantic ratings that describe the different facial expressions are computed by the IKCCA to generate a seven-dimensional semantic expression vector. It is applied for learning the correlation with different testing samples. According to this correlation, we estimate the associated expression vector and perform expression classification.

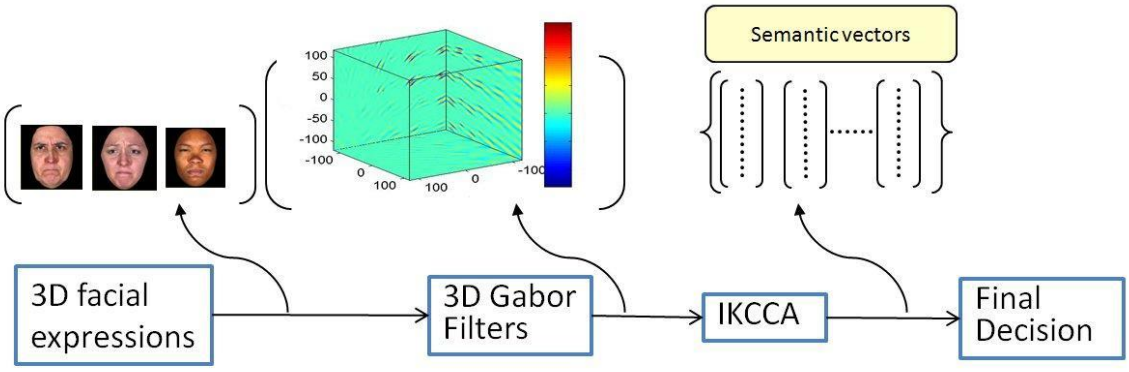


Figure 2.1 Block diagram of the 3D Gabor based method

To validate our proposed approach, we have conducted experiments for person-independent facial expression recognition on a public 3D facial expression database, i.e. the BU_3DFE database [29]. The experimental results yield an overall recognition rate of 85.39%.

2.4.2 3D EBS Feature Based Recognition

We present an automatic emotion recognition system from video sequences using a 3D physical face model with the EBS technique. The main contribution of this work is using the EBS-based method for automatic human emotion recognition from video sequences with the active deformation feature extraction depending on the 3D generic face model, which is driven by the key fiducial points, and thus to make it possible to generate the intrinsic geometries of the emotional space. The block diagram of this method is shown in Figure 2.2.

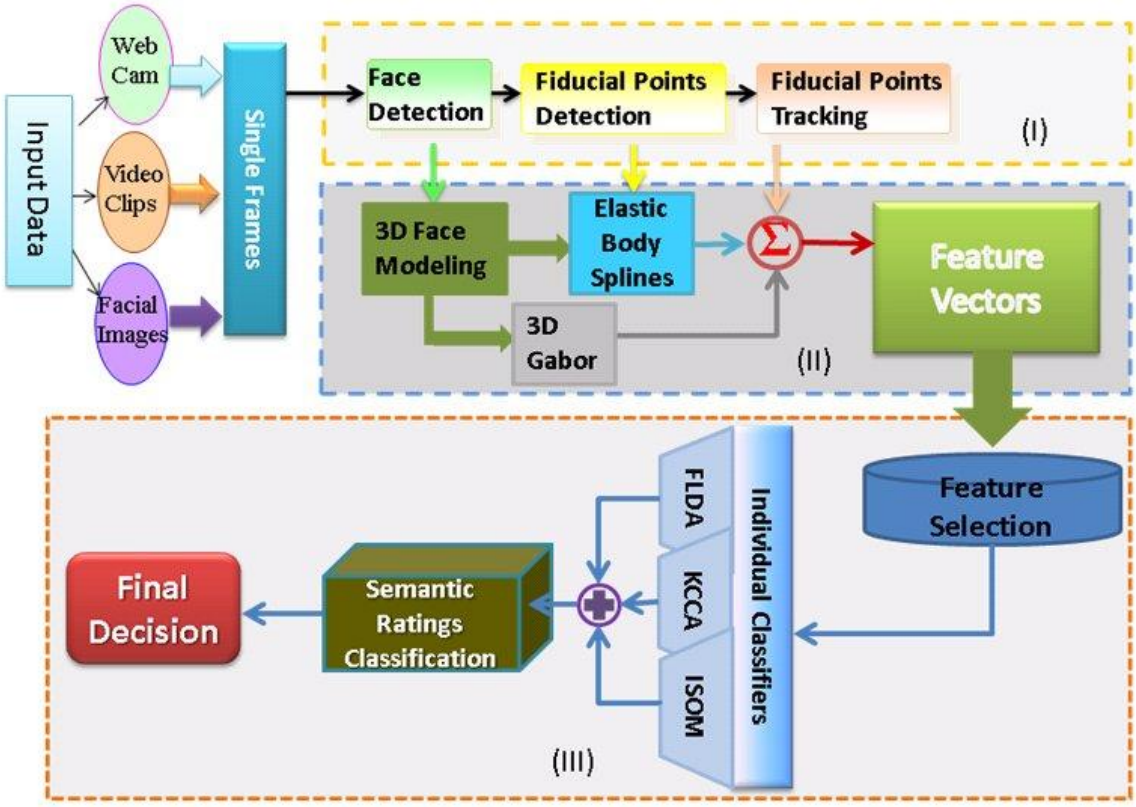


Figure 2.2 EBS Feature Based Recognition System (I) input data processing, (II) visual feature extraction, (III) emotional space classification

We detect the facial region in a video sequence that consists of feature selection and classification based on a local normalization technique and face detection algorithm using a Gabor wavelet transform and the Adaboost algorithm. This step incorporates a normalization technique based on normalized local histograms with OAC technique, which alleviates the illumination problem in conventional face detection methods. Scale space extrema are calculated on the facial region for selection of candidate points of interest. The direction of each candidate in its neighborhood is calculated by a gradient orientation histogram. The feature description for each fiducial point is obtained by

connecting the direction descriptions of it with its neighbours and is used for fiducial point detection. We apply multiple DE-MC particle filters to track the fiducial points depending on the locations of the current appearance of the spatially sampled features. A DE-MC particle filter leads to a more reasonable approximation to the proposal distribution and hence considerably improves accuracy for tracking by building a path connecting a sampling with measurement. The kernel correlation based on HSV colour histograms is used to estimate the observation likelihood and measures the correctness of particles.

We also construct a 3D generic face model based on the results of fiducial points detection and tracking. As a physics-based transformation, EBS is applied to the face model to generate a smooth warp that reflects control point correspondences and extracts the deformation feature of the realistic expression. D-Isomap based classification is used to embed the facial expression movements into a low dimensional manifold, which span in a 3D expressional face space.

Chapter 3

3D Gabor Based Recognition

TO recognize an emotional state from facial expressions, a set of feature vectors that can best describe the particular set of facial expressions needs to be extracted to discriminate between expressions. The feature vector includes the amount of information extracted from the particular facial expression and should not match with another one that belongs to some other expression. It is the most important aspect for a successful emotion recognition system. The general 3D approaches reported in the literature only use geometric information without any colour/density information of the face. However, these features can strongly affect the performance of an emotion recognition system.

In this chapter, we present a new emotion recognition method using 3D geometric features with facial density information. Gabor transform-based facial expression recognition systems show that their representation method has a high degree of correlation with the human semantic ratings [30]. It is a reasonable model of visual processing in the primary visual cortex and can be one of the most successful approaches

for processing images of the human face. The Gabor filter based feature extraction technique has proved to produce an extremely effective method for facial expression recognition, and is reported to yield good results on novel individuals applied in face recognition among several facial emotion recognition researches.

A Gabor filter is a multiple sinusoid modulated by a Gaussian function. It has wide applications in many research areas such as image processing and pattern recognition. Many previous research works [31-34, 42] have experimentally shown that the Gabor filter representation is optimal for classifying facial expressions since it captures various visual properties of facial regions using spatial localization, orientation selectivity and spatial-frequency characteristics. Compared with other popular transformations, such as traditional Fourier transform, DCT and various wavelets, the Gabor filter can be designed to be highly selective in frequency while displaying good spatial localization. It also has multi-resolving ability and tunable focus. Using multi-channel filtering, Gabor wavelets can be applied with different spatial-frequency properties to extract expression features from facial images. In the aspect of facial features extraction, Gabor wavelet transformation is insensitive to illumination variety, and the limited localization in space and frequency yields a certain amount of robustness against translation, distortion, rotation and scaling of the images. The observed expression images do not need to

correspond to the expression template strictly. Thus, the robustness of the whole system can be approved. All in all, Gabor filter based feature extraction can obtain more significant information and outperform other approaches. It is a promising feature extraction technique for face and expression recognition tasks.

3.1 3D Gabor Feature Extraction

Applying the 3D Gabor transform for feature extraction attracted more interests in recent years [35]. Feng *et al.* [36] used 3D Gabor for motion estimation with adjustable spatiotemporal resolution, Zhen *et al.* [37] applied the 3D Gabor library for MRI tagging sheet extraction and tracking, Kepenekci *et al.* [38] analyzed motion using the 3D Gabor kernels. In this work, we proposed to apply the 3D Gabor library to the 3D facial expressions for human emotion recognition.

The 3D Gabor library is used to extract 3D geometric information plus colour/density information by spatial localization, orientation selectivity and spatial-frequency characteristics to discriminate between expressions. The feature vector from the 3D Gabor library includes the amount of information extracted from the particular facial expression and will not match with another one that belongs to some other expressions.

3.1.1 3D Gabor Filter

A 3D Gabor transform is basically a product of a complex sinusoid wave modulated by a 3D Gaussian window. We have the 3D Gabor transform as the following:

$$G_{\gamma,\alpha,\beta}(x,y,z) = A \times H(x',y',z') \times S(x,y,z) \tag{3.1}$$

where $A = \frac{1}{(2\pi)^{3/2}\sigma_{x'}\sigma_{y'}\sigma_{z'}}$ is a normalization scale, $H(x',y',z')$ is a 3D Gaussian envelope that:

$$H(x',y',z') = \exp[-\frac{1}{2}((\frac{x'}{\sigma_{x'}})^2 + (\frac{y'}{\sigma_{y'}})^2 + (\frac{z'}{\sigma_{z'}})^2)] \tag{3.2}$$

and

$$S(x,y,z) = \exp[-j2\pi(Mx + Ny + Lz)] \tag{3.3}$$

where (x,y,z) is the non-rotated spatial coordinate, $\sigma_{x'}\ \sigma_{y'}\ \sigma_{z'}$ are defined as the width of the Gaussian envelop in different x, y and z axes, respectively, and can be tuned to the local structures. $(x',y',z')^T = R \times (x,y,z)^T$ is the rotated spatial coordinates of the 3D Gaussian envelope, R is a rotation matrix for transforming the Gaussian envelope to coincide with orientation of the sinusoid. These two coordinates are set to be the same value for normalization purposes. The Gaussian scale parameters $\sigma_{x'}\ \sigma_{y'}\ \sigma_{z'}$ may not be the same though they are normally set to be the same, and the sinusoid and the Gaussian envelop can have different orientations. Thus the shape of this Gaussian envelope can be an ellipsoid. (M,N,L) are the 3D frequencies of the complex sinusoid that

$$N = \gamma \sin\beta \sin\alpha$$

$$M = \gamma \sin\beta \cos\alpha$$

$$L = \gamma \cos\beta \tag{3.4}$$

where $\gamma = \sqrt{(M^2 + N^2 + L^2)}$ is the amplitude of the complex sinusoid wave with frequency (M, N, L), $0 \le \alpha \le \pi$ and $0 \le \beta \le \pi$ determine the orientation and spacing of the Gabor filter in the spatial domain. The rotation matrix of the Gaussian envelope R is given as

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\beta & -\sin\beta \\ 0 & \sin\beta & \cos\beta \end{bmatrix} \times \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.5}$$

which is the normalization process to make the Gaussian envelope have the same orientation as the complex sinusoid.

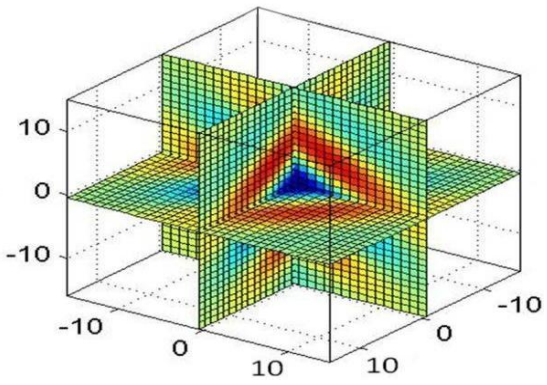


Figure 3.1 A slice view of a 3D Gabor filter

Figure 3.1 shows projections of a 3D Gabor filter with $\sigma_{x'} = \sigma_{y'} = \sigma_{z'} = \sigma$, $\gamma = 0.25$, $\alpha = \pi/2$, $\beta = \pi/2$, $\sigma = 1/\gamma$, and the size of the filter is $60 \times 60 \times 60$.

3.1.2 Gabor Library Design

Since the prior information about the 3D facial expression is unknown, we consider constructing the 3D Gabor library using a set of Gabor filters with different frequencies and orientations (γ, α, β) to obtain sufficient information as the following:

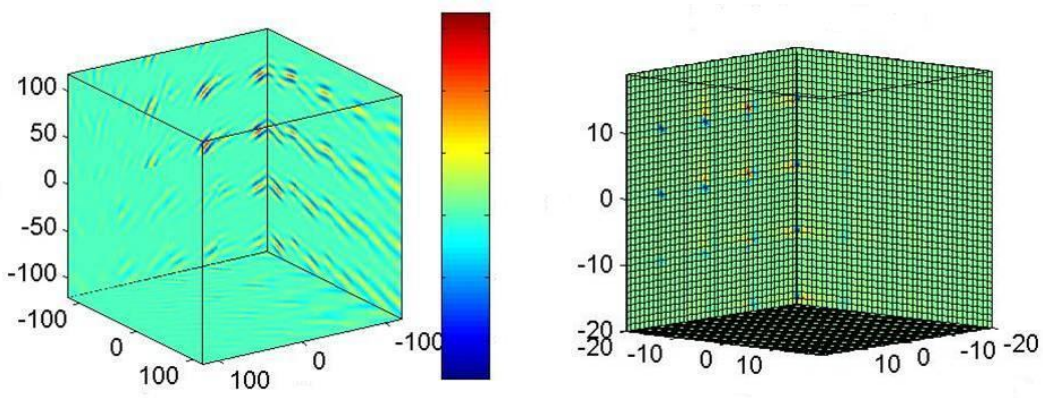
$$G_{\gamma_k,\alpha_v,\beta_w}(x,y,z) = \left(\frac{1}{(2\pi)^{3/2}\sigma_{x'}\sigma_{y'}\sigma_{z'}}\right) e^{-\frac{1}{2}\left[\left(\frac{x'}{\sigma_{x'}}\right)^2+\left(\frac{y'}{\sigma_{y'}}\right)^2+\left(\frac{z'}{\sigma_{z'}}\right)^2\right]} e^{-j2\pi(Mx+Ny+Lz)}$$

$$\gamma_k = \gamma_{\max}/(2)^{K/2},\ \alpha_v = v\pi/V, \text{ and } \beta_w = w\pi/W \tag{3.6}$$

where $\gamma_{\max}$ is the upper centre frequencies of the signal to be analyzed. We denote the wavelets as $G_{k,v,w}, \{k = 0,...,K-1, v = 0,...,V-1, w = 0,...,W-1\}$. The feature of 3D facial volume I can be extracted with the frequency and orientation information F from the voxel $\bar{T}$:

$$F(\bar{T}) = (I(\bar{T}) \otimes G_{k,v,w})(x,y,z) \tag{3.7}$$

Since the magnitude of the convolution result can express the response of a Gabor wavelet to the facial volume I, the useful information of the intensity changes at voxel $\bar{T}$ can be obtained by applying the Gabor library to a 3D facial expression. In this dissertation, the library parameters are set to be $4\times4\times4$, and in Figure 3.2, we show the designed 3D Gabor filter library in both space domain and frequency domain.



(a) (b)

Figure 3.2 The 3D Gabor library of $4\times4\times4$ (a) 3D Gabor library in space domain, (b) 3D Gabor library in frequency domain

3.1.3 Feature Representation

In this section, we describe our method for visual feature extraction using the 3D Gabor library. We normalize all the input data to be the same size and denote it as $I(x,y,z)$. The Gabor wavelet transform $F(x,y,z)$ of this volume $I(x,y,z)$ can be calculated from (3.7) as:

$$F_{kvw}(x,y,z)=\int I(x_1,y_1,z_1)G^{*}_{kvw}(x-x_1,y-y_1,z-z_1,)dx_1dy_1dz_1 \qquad (3.8)$$

where * indicates the complex conjugate. In this case, we obtain a very big coefficient matrix for each face. We use a total of $4\times4\times4=64$ Gabor filters, and thus the size of the matrix is $60\times60\times60\times64$. With a feature space of such a big size, the computation cost is very high, and thus it is not very suitable for a system which demands fast processing.

We take the mean μ_{kvw} and standard deviation σ_{kvw} of the magnitude of the transform coefficients of each sub-band filter to represent the features:

$$\mu_{kvw} = \iiint |F_{kvw}(x, y, z)| dxdydz$$
$$\sigma_{kvw} = \sqrt{\iiint (|F_{kvw}(x, y, z)| - \mu_{kvw})^2 dxdydz} \tag{3.9}$$

Then we can construct a feature vector for each emotional face using μ_{kvw} and σ_{kvw} as feature components. In this dissertation, we use four scales $K = 4$ and four orientations $V = W = 4$ in α and β directions, resulting in a feature vector of 128 dimensions for latter classification:

$$f_{3DGabor} = [\mu_{000} \quad \sigma_{000} \quad \mu_{001} \quad \sigma_{001} \quad ..... \quad \mu_{333} \quad \sigma_{333}] \tag{3.10}$$

3.2 IKCCA Classification

Canonical correlation analysis (CCA) is a powerful method to correlate the linear relations between arbitrary variables. It tends to combine the variables into a single-dimension in the new space if they are highly correlated. However, for emotion recognition tasks, multidimensional feature representations of the person-independent facial expressions are nonlinearly correlated. CCA may not correctly correlate the relationships between these feature representations.

Although the idea of kernelizing CCA is not new, all the previous works [39, 40] tackle the singularity problem using the regularization method or the eigenvalue

decomposition method. Fukumizu *et al.* [41] provided an improved kernel CCA algorithm to solve this problem based on a normalized cross covariance approach. It was applied for expressing the nonlinear dependence between two variables. In our work, we extend this method to multiple emotional states recognition with 3D visual features.

3.2.1 IKCCA Algorithm

IKCCA is a nonlinear extension of the CCA algorithm with the normalized cross covariance operator. It has the ability to infer semantic relations between the nonlinear feature representation variables. This method tries to find projections of each feature representation separately such that the representations are maximally correlated. This implies that the visual feature variables represent the same emotional expression from different faces. So we can extract person-independent emotions using the 3D visual features from the last section.

Let $X \in R^{n_x}$ denote the sample feature vector and $Y \in R^{n_y}$ the semantic expression vector. $A = a^T X$, $B = b^T Y$ are their projections. To determine the corresponding emotional classification for a given vector X , we need to find a pair of directions a_X and b_Y so that the canonical correlations ρ can be maximized:

$$\rho(X, Y; a, b) = \frac{E[AB]}{\sqrt{E[A^2]E[B^2]}} = \frac{a^T E[XY^T]b_Y}{\sqrt{a_X^T E[XX^T]a_X}\sqrt{b_Y^T E[YY^T]b_y}}, \qquad (3.11)$$

The canonical correlation ρ measures the strength of association between the sample feature vector X and the semantic expression vector Y. In practice, we estimate the objective function ρ as a desired individual expression pattern classification for a finite sample.

For the nonlinear emotion recognition case, let $\Phi(X)$ and $\Psi(Y)$ denote the diagonals of X and Y in Hilbert space through nonlinear mapping respectively, so the correlation function ρ can be reformed as:

$$\rho(\Phi(\mathrm{X}), \Psi(Y); a_{\Phi(X)}, b_{\Psi(Y)})$$

$$= \frac{a_{\Phi(X)}^T E[\Phi(X)\Psi^T(Y)] b_{\Psi(Y)}}{\sqrt{a_{\Phi(X)}^T E[\Phi(X)\Phi^T(X)] a_{\Phi(X)}} \sqrt{b_{\Psi(Y)}^T E[\Psi(Y)^T \Psi(Y)] b_{\Psi(Y)}}} \tag{3.12}$$

where $E[\Phi(X)\Psi^T(Y)]$ is the empirical covariance that:

$$E[\Phi(X)\Psi^T(Y)] = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi(X_i) - \frac{1}{n}\sum_{j=1}^{n}(\Phi(X_j))\right)\left(\Psi(Y_i) - \frac{1}{n}\sum_{j=1}^{n}(\Psi(Y_j))\right) \tag{3.13}$$

and $$E[\Phi(X)\Phi^T(X)] = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi(X_i) - \frac{1}{n}\sum_{j=1}^{n}(\Phi(X_j))\right)^2$$

$$E[\Psi^T(Y)\Psi(Y)] = \frac{1}{n}\sum_{i=1}^{n}\left(\Psi(Y_i) - \frac{1}{n}\sum_{j=1}^{n}(\Psi(Y_j))\right) \tag{3.14}$$

Since a and b can be rescaled without changing the problem, we can constrain them to be equal to 1. The objective function can be written as:

$$\max \quad \rho = a^T E[\Phi(X)\Psi^T(Y)] b$$

$$\text{subject to} \quad a_{\Phi(X)}^T \Phi(X)\Phi^T(X)a_{\Phi(X)} = 1$$

$$\text{and} \qquad b_{\Psi(Y)}^T \Psi^T(Y)\Psi(Y)b_{\Psi(Y)} = 1 \tag{3.15}$$

Using the corresponding Lagrangian, we can denominate the original objective as:

$$L = a^T\Phi(X)\Psi^T(Y)b - \frac{1}{2}\mu(a^T\Phi^2(X)a - I) - \frac{1}{2}\nu(b^T\Psi^2(Y)b - I) \tag{3.16}$$

where $\mu = \nu$ and

$$\mu^2 = \frac{a^T\Phi^T\Phi(\Psi^T(\Psi\Psi^T)^{-1}\Psi)\Phi^T\Phi b}{a^T\Phi^T\Phi\Psi^T\Psi b} \tag{3.17}$$

Thus, to compute the canonical correlations ρ is equivalent to finding a pair of directions a and b to maximize μ using (3.17).

3.2.2 IKCCA Classifier

Applying IKCCA for the emotion classification, first we set $\Phi(X_{test})$ to be an input test feature vector and $\Psi(Y_{\text{semantic}})$ the corresponding semantic expression vector. Let A_{test} be the projection of $\Phi(X_{test})$ onto the directions $a^i{}_{\Phi(X)}\big|_{i=1}^n$ and B_{semantic} the projection of $\Psi(Y_{\text{semantic}})$ onto the directions $b^i{}_{\Psi(Y)}\big|_{i=1}^n$, and then we have:

$$\begin{aligned} A_{test} &= [A_{test}^1, \ldots\ldots, A_{test}^n] \\ &= [a_{\Phi(X)}^1 \ldots\ldots a_{\Phi(X)}^n]^T \Phi(X_{test}) = P_x^T \Phi(X_{test}) \end{aligned}$$

$$\begin{aligned} B_{\text{semantic}} &= [B_{\text{semantic}}^1, \ldots\ldots, B_{\text{semantic}}^n] \\ &= [b^1{}_{\Psi(Y)} \ldots\ldots b^n{}_{\Psi(Y)}]^T \Psi(Y_{\text{semantic}}) = P_y^T \Psi(Y_{\text{semantic}}) \end{aligned} \tag{3.18}$$

where $P_x = [a_{\Phi(X)}^1 \ldots\ldots a_{\Phi(X)}^n]$ and $P_y = [b^1{}_{\Psi(Y)} \ldots\ldots b^n{}_{\Psi(Y)}]$. Solving (3.18) yields:

$$\Psi(Y_{\text{semantic}}) = (P_y P_y^T)^{-1} P_y B_{\text{semantic}} \tag{3.19}$$

We can rewrite (3.19) when $P_y P_y^T$ is singular:

$$\Psi(Y_{test}) = (P_y P_y^T + \zeta_n I)^{-1} P_y B_{test} \tag{3.20}$$

Using (3.15), (3.17) and (3.20), we can estimate the corresponding semantic expression vector Y_{semantic} for a given input feature vector X_{test}. The index of the most matched emotion class of the input sample is signed by:

$$C^* = \arg\max\left\{P_x^T \Phi(X_{test}), P_y^T \Psi(Y_{\text{semantic}})\right\}$$

$$= \arg\max_{i=1:n}[\rho(X^i{}_{test}, Y^i{}_{\text{semantic}}; a^i{}_{\Phi(X)}, b^i{}_{\Psi(Y)})] \tag{3.21}$$

3.2.3 Semantic Ratings Classification

We can further increase the recognition rate by using the semantic ratings classification. Directly applying the IKCCA classifier on face data for emotion recognition, all feature vectors are treated equally. The emotion class information is not used for distinguishing from others. By considering the overall correlations for the different classes with the class label information, the semantic ratings classification can reflect successfully the discriminant structures of the feature vector on the emotional space.

We construct the semantic ratings Y'_{semantic} from each training data by a weight factor w_{ij}, which is used to compact the training semantic expression vector Y^i and Y^j if they

share the same label, and expand Y^i and Y^j if they belong to different classes. Then we have:

$$Y'_{\text{semantic}} = \max_{i,j=1:n}[w_{ij}Y^i(X^i, a^i{}_{\Phi(X)}, b^i{}_{\Psi(Y)}) - Y^j(X^i, a^i{}_{\Phi(X)}, b^j{}_{\Psi(Y)})] \tag{3.22}$$

We need to construct the semantic expression vector first for the semantic ratings implementation and compute the overall correlations between the pairs. Note that w_{ij} is not the weight between training data, X^i and X^j, but an informative factor represented for semantic expression vectors, which is determined by class label information $L(i, j)$ and the correlations between semantic expression vectors Y^i and Y^j. By considering the overall correlations between the pairs, we have that:

$$w_{ij} = f[\rho(Y^i, Y^j), L(i, j), X^i, X^j] \tag{3.23}$$

where $\rho(Y^i, Y^j)$ gives the overall correlations between the pairs, and $L(i, j)$ is a label function defined as following:

$$L(i, j) = \begin{cases} 1 & if \quad Y^i, Y^j \quad \text{share the same label} \\ 0 & if \quad Y^i, Y^j \quad \text{have different labels} \end{cases} \tag{3.24}$$

w_{ij} would be able to affect the overall correlations between the pairs. When $L(i, j) = 1$, that means X^i and X^j come from the same emotion class. So we would like to preserve the projections of Y^i and Y^j getting higher correlation if they are from the same class.

And the inverse is also true, that means decreasing the correlation if they are from the different classes.

3.3 Experiment and Results

3.3.1 3D Facial Expression Database

A public person-independent 3D facial expression database, BU_3DFE database is used in our experiment to validate our proposed approach. The BU_3DFE database is the largest publicly available data set for 3D facial expression recognition research and contains images exhibiting substantial expression variation, which can cause problems for many recognition algorithms. Figure 3.3 shows some samples from this database.

In BU_3DFE 3D facial expression database, there are 100 subjects who participated in face scans, including undergraduates, graduates and faculty from State University of New York at Binghamton. The resulting database consists of about 60% female and 40% male subjects with a variety of ethnic/racial ancestries. Each subject in the database performed seven expressions (including neutral), captured by a 3D face scanner. With the exception of the neutral expression, each of the six prototypic expressions (happiness, disgust, fear, angry, surprise and sadness) includes four levels of intensity. From the database, we choose 560 samples for classifier training and 280 samples for testing, and each delivers

one of the seven facial expressions. There was no overlap between the training and testing subjects.

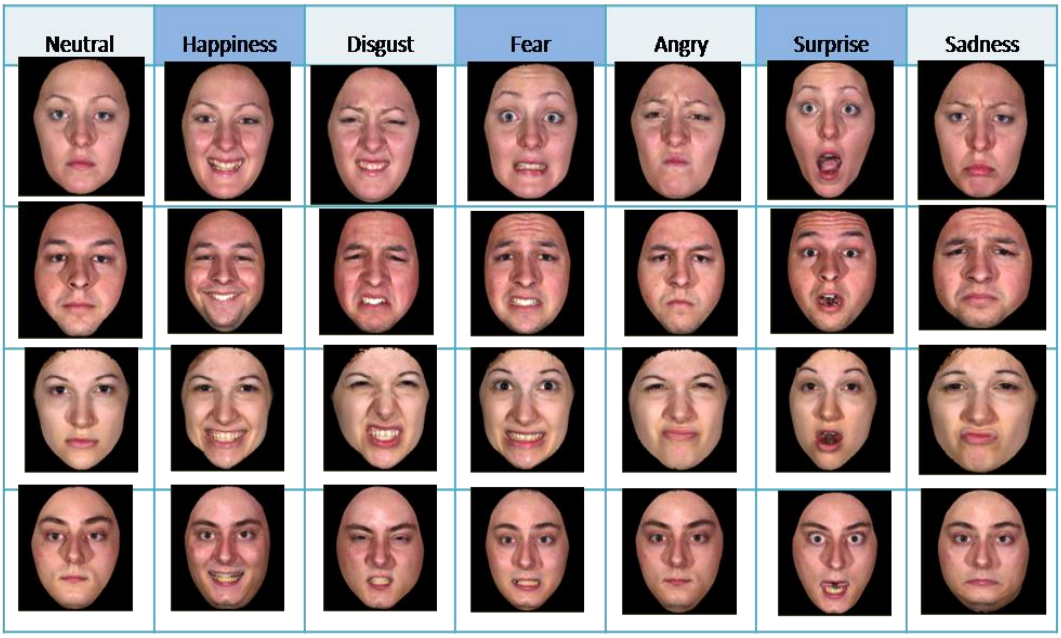


Figure 3.3 Sample Expressions of 4 subjects from BU_3DFE database

3.3.2 Feature Selection

The performance of the IKCCA based emotion recognition system depends on how to find out the best correlations between the feature vectors for the classification task. For a pattern recognition system, the length of the feature vector and the discriminating ability of the features, in terms of separating patterns belonging to different classes in the feature space, will critically affect the overall performance of the system. The importance of selecting relevant subset from the original feature set is closely related to the "curse of dimensionality" problem in function approximation, in which sample data points become

increasingly sparse as the dimensionality of the function domain increases. The finite set of samples may not be adequate for characterizing the original mapping and the computational requirement is higher for implementing a high dimensional mapping.

In this work, we use the PCA to reduce the dimensionality of the input feature vector to alleviate the aforementioned problems by reducing the number of transformed features, whilst retaining most of the intrinsic information content of the original data. The PCA technique arranges the feature vector in descending order of variance and is truncated at desired length such that the remaining feature vector is sufficient for accurate recognition of facial expressions.

In the PCA parameter determination, we first produce the data set with zero mean, and then compute the covariance matrix from the covariance between two dimensions. New data vectors are formed by projecting the original data onto the principal component vectors with the following criterion. The components of the new feature vector are derived from the first M eigenvectors that satisfy $\frac{\sum_{i=1}^{M}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \geq 0.9$, where λ_i is the *i*th largest eigenvalue, and N is the dimension of the original feature space [25]. The component vectors are determined and then can be used for the later emotion classification.

3.3.3 IKCCA Classifier

Table 3.1 Recognition Rates of the IKCCA classifier

	neutral	angry	disgust	fear	happiness	sadness	surprise
3D Gabor	70.12	77.92	63.66	67.86	69.67	63.90	76.96
3D Gabor +PCA	**70.74**	**84.69**	**64.62**	**65.18**	**82.71**	**66.52**	**81.04**

To clarify the relationship between feature vectors and the corresponding emotion category, in our experiment, all kinds of expression feature vectors of seven facial expressions collected from different people are extracted and compared using individual IKCCA classifiers. The IKCCA is carried out on the training samples of the 3D Gabor feature vectors, and then used to classify the data. This approach attempts to maximize the intra-class correlation following (3.15) and (3.17). We show the recognition rates by individual IKCCA classifiers with and without PCA feature selection in Table 3.1. From the table we can see that by applying PCA, the system performance is improved.

Table 3.2 Confusion matrix of IKCCA on BU_3DFE database

	Detected						
Desired	**Neutral**	**Happiness**	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Neutral	**70.74**	1.89	7.35	8.65	5.66	1.56	4.15
Happiness	2.95	**82.71**	5.56	3.32	1.45	1.04	2.97
Disgust	1.70	1.78	**64.62**	13.81	9.81	0.76	7.52
Fear	0.94	1.42	4.55	**65.18**	9.09	18.18	0.64
Angry	5.31	0.45	4.62	1.89	**84.69**	1.38	1.66
Surprise	1.85	1.60	1.26	6.12	3.56	**81.04**	4.57
Sadness	4.92	3.52	4.76	14.29	1.37	4.62	**66.52**

In Table 3.2, we show the confusion matrix of the IKCCA classifier on the BU_3DFE database. From Table 3.2, the same expressions collected from different subjects are very similar due to the fact that they are highly correlated within the same emotion class. The overall recognition rate using individual classifiers with PCA is about 73.64%.

From the results we find that expressions with a negative value correspond to a negative reaction in terms of arousal and stance, while positive values correspond to a positive reaction. Note, the expressions that are detected with high accuracy and low confusion are in the happiness, anger and surprise classes. The reason is that, in general, these emotions have strong negative or positive values with more distinguishable corresponding facial expressions. The confusion matrix also illustrates the most common misclassifications. In general, emotions of sadness, fear and disgust have low recognition rates, as they do not occur naturally alone but associated with other emotions. Moreover, the misclassifications can be attributed to the inherent difficulty of the classification into a few categories of expressions.

We compare the 3D Gabor based IKCCA classifier with and without PCA selection, and a 2D Gabor based method by [25]. Figure 3.4 summarizes the results of the best classification accuracy as an average percentage achieved by these methods. From the

figure, it is very clear that when classifying the seven prototypic facial expressions, using the 3D Gabor feature with PCA selection representation obtains a recognition rate of 73.64% which significantly exceeds the recognition rates of 49.29% by the 2D Gabor based method and 67.3% without PCA selection.

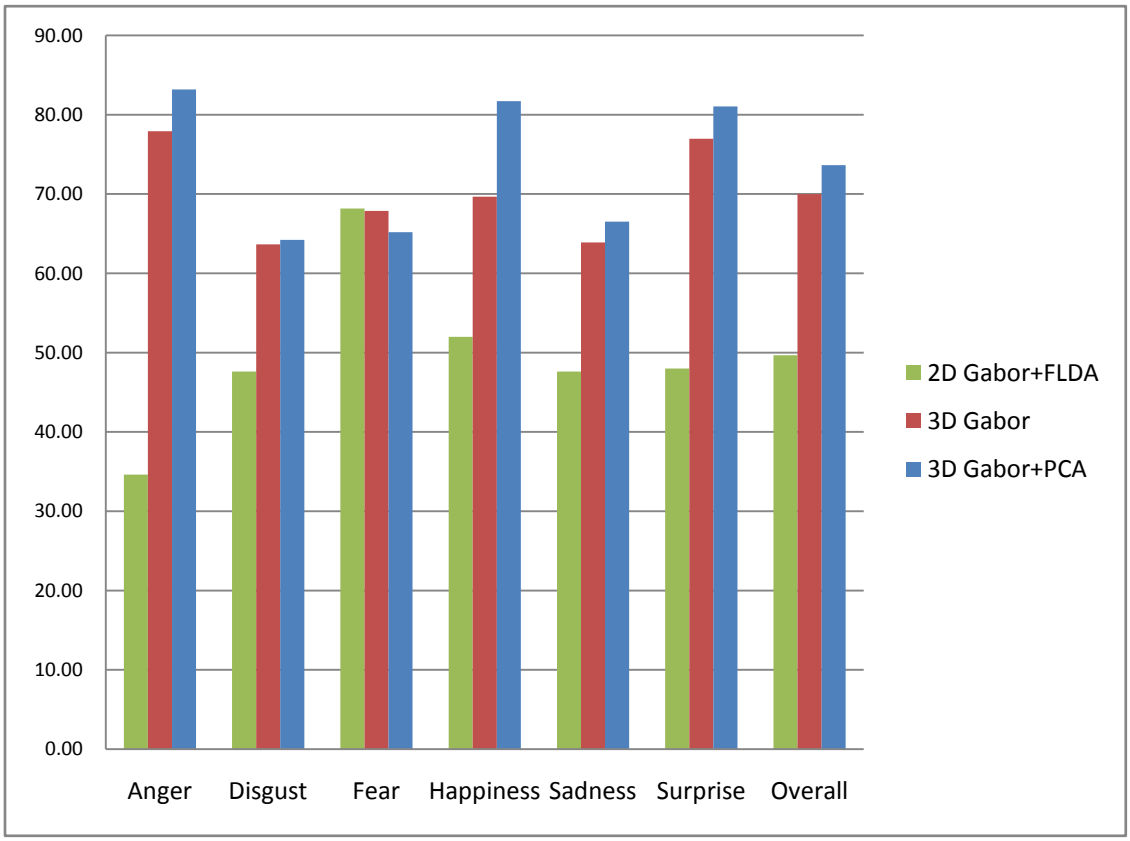


Figure 3.4 Classification Comparisons of 2D and 3D Gabor Filters

3.3.4 Semantic Ratings Classification

To improve the recognition rate, we introduce a semantic ratings classification method based on the individual IKCCA classifiers by considering the overall correlations for the different classes. We use the calculated semantic ratings by quantitatively evaluating the

seven emotional expression, while minimizing the overall correlations between the projected means of different samples we intend to separate.

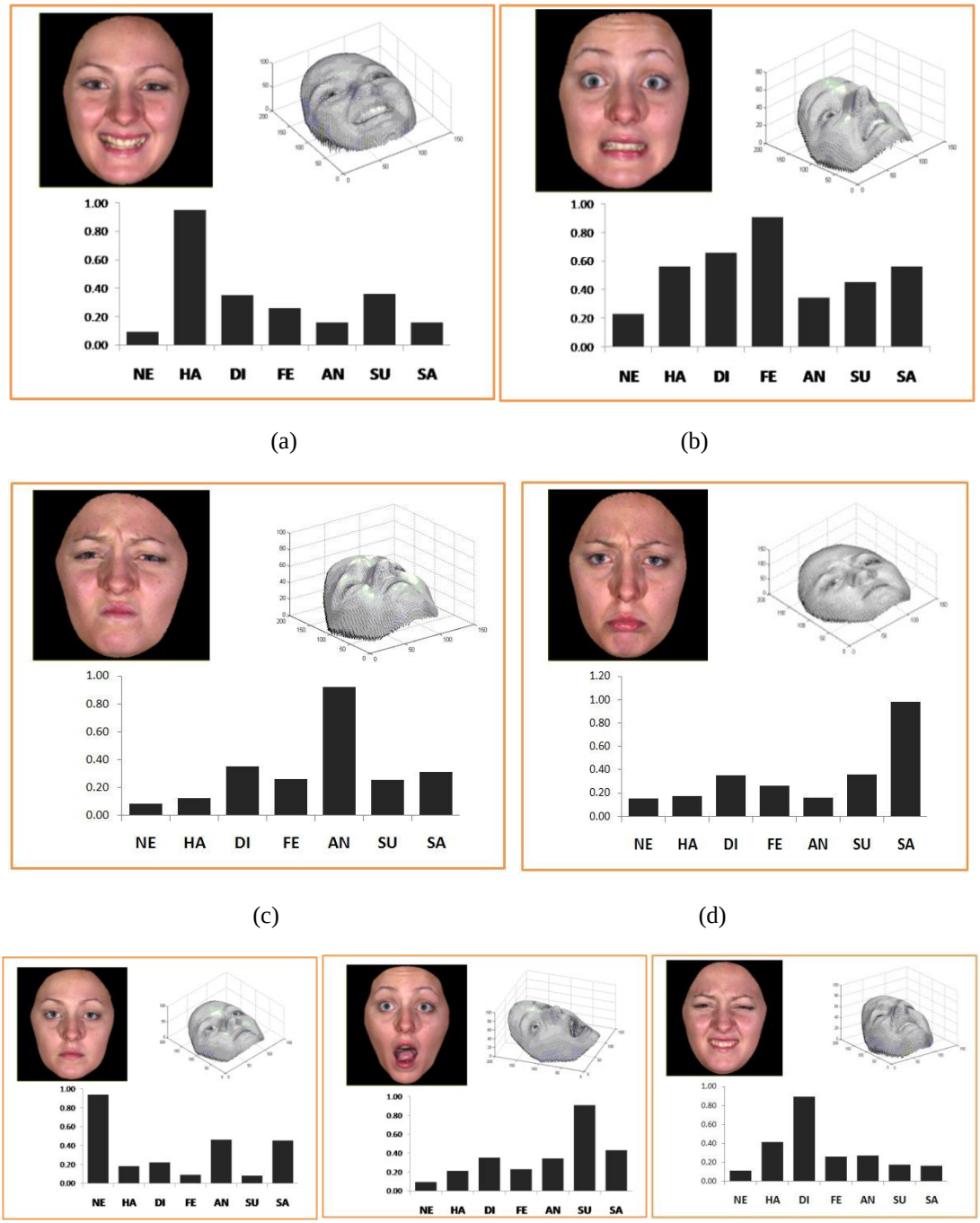


Figure 3.5 Samples of semantic ratings for different expressions of emotions (a) Happiness (b) Fear (c) Angry (d) Sadness (e) Neutral (f) Surprise (g) Disgust

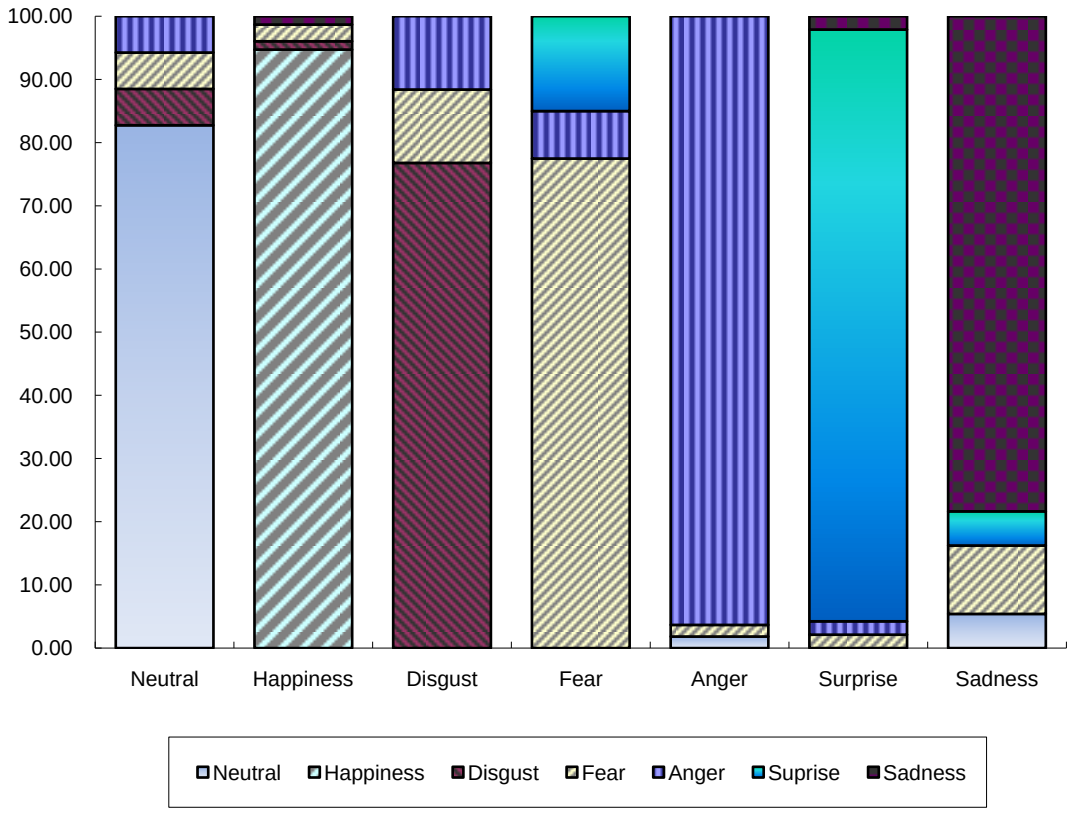
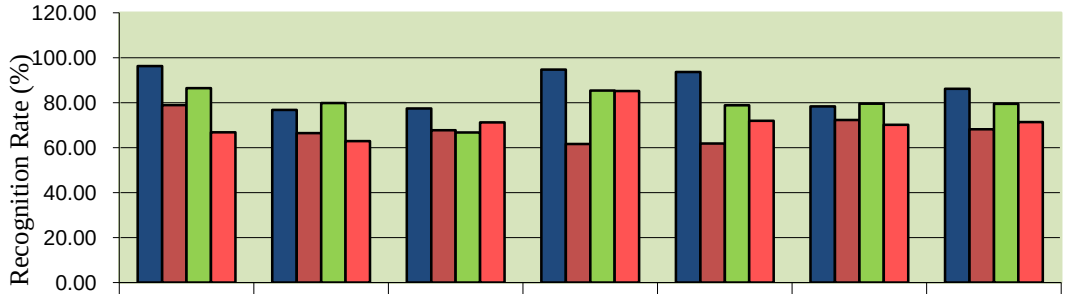


Figure 3.6 Final Emotion Recognition Rate with IKCCA Based Algorithm

The semantic ratings are computed from the training samples and combined into a seven- dimensional semantic expression vector for analysis. For a new 3D facial expression query, we first generate the feature vector using the 3D Gabor filters. The corresponding semantic ratings of each facial expression are estimated using (3.22), and then the emotion classification is performed according to (3.21) with the new trained Y'_{semantic}. From the result we can see that features representing different expressions exhibit diversity since the correlations between different emotions are relatively low.

Figure 3.5 illustrates some samples of the semantic ratings for the facial expression model.



	AN	DI	FE	HA	SA	SU	Ave
IKCCA	96.33	76.78	77.47	94.71	93.64	78.36	86.22
GMM	78.94	66.49	67.78	61.64	61.88	72.33	68.18
FLDA	86.46	79.85	66.73	85.44	78.86	79.57	79.49
NN	66.80	62.90	71.27	85.21	71.96	70.16	71.38

Figure 3.7 Comparison of Recognition results from different classifiers

The proposed method is also tested using the leave-one-out cross validation approach, which is the most frequently used approach for testing the generalization performance of a classifier. This approach is applied in order to make maximal use of the available data and produce averaged classification accuracy results. The facial expressions belonging to one subject are used as the testing data and the remainders as the training data. This is repeated for all the possible trials until all subjects are used as testing data. The recognition accuracy is calculated as the ratio of the number of correctly classified samples and the total number of samples in the data set. The experimental results are averaged to produce the final recognition rate and shown in Figure 3.6. From Figure 3.6,

we observe an overall recognition rate of 85.65% by this strategy, a 12% improvement over the individual classifier.

We conduct extensive experiments using different classification schemes, i.e., Gaussian Mixture Model (GMM), Fisher's Linear Discriminant Analysis (FLDA), and NN algorithm. The experimental results, for the performance comparison with the same data set, are drawn in Figure 3.7. The GMM classifier is implemented in a modular architecture. A separate GMM is trained for each individual class. The parameters including the weights, mean and standard deviation of each component are estimated by the Expectation Maximization (EM) algorithm. In out experiments, we try a range of k values, so that the distribution of the data can be modeled as the sum of k Gaussian functions. The applied FLDA classifier has six outputs corresponding to the six emotions. An input signal is labeled with the class that gives the maximum output value. In NN classification, a three-layer feed-forward neural network is investigated. The number of input layer neurons is equal to the dimension of the input feature set, while the output neurons correspond to the six emotion classes. The back-propagation algorithm is used to train the network. A new input is labeled with the class that produces maximum output value. From Figure 3.7 we can see IKCCA-based method achieves the best results for the final emotion recognition.

Comparisons of the recognition rates achieved by the 2D feature-based and 3D feature-based approaches are depicted in Table 3.3 and 3.4. We can see that 3D feature-based approaches have better performances due to the ability to handle pose variations and lighting illumination problems, and outperform the 2D-based methods that are sensitive to such conditions.

Table 3.3 Comparison with 2D feature-based Emotion Recognition approaches

Ref.	Lyons [18]	Silva [19]	Cohen [20]	Y. Wang [25]	Lu [43]	Wu [44]	Lajevardi [45]	This Work
Dimension	2D	2D	2D	2D	2D	2D	2D	3D
Feature	Gabor	geometric	PBVD	Gabor	Gabor	Gabor	Log-Gabor	Gabor
Classifier	LDA	NN	HMM	FLDA	NKFDA	LDA	NN	IKCCA
Recognition Rate	75%	60%	66.53%	78%	81%	78%	70%	85%

Table 3.4 Comparison with 3D feature-based Emotion Recognition approaches

Ref.	Hu [27]	Tang [46]	Soyel [47]	J. Wang [48]	This Work
Ini	Manu	Manu	Manu	Auto	Auto
Feature	Geometric points	Geometric	Gabor	geometric	Gabor
Classifier	SVM	AdaBoost	NN	LDA	IKCCA
Database	BU_3DFE	BU_3DFE	BU_3DFE	BU_3DFE	BU_3DFE
Recognition Rate	71%	95%	91.3%	83%	85%

Computationally, our proposed method has the advantages of automatic feature extraction using the real 3D visual features over the existing 3D based methods that only consider the 3D geometric information of the facial expressions. Note that the method proposed in [46] and [47] achieved an overall recognition rate of 95.1% and 91.3%

respectively. However, these methods are only tested on perfect manually aligned feature points and no experiments in fully automatic conditions were reported. Therefore, direct comparison with our proposed method, in terms of performance, will be biased and unrealistic.

3.4 Chapter Summary

In this chapter, we proposed a new facial emotion recognition method using real 3D visual features which are extracted automatically. We constructed a 3D Gabor library for facial feature extraction with an IKCCA algorithm for final decision making. We observed that using 3D visual features for emotion analysis has better performances with more visual feature information. Apparently, this work points to a promising direction for the analysis of 3D-based emotion recognition.

We applied the IKCCA algorithm to seven-dimensional semantic expression vector ratings to classify the prototypic facial expressions. Our work shows that IKCCA is a very effective method for correlating the nonlinear relationship between the facial features and the associated semantic features. It provides us with an effective way to predict the semantic expression information of a facial expression. We observed superior performance by the proposed method when compared with several other methods using

automatic feature extraction. To improve the performance of the semantic ratings-based approach, a better way could be obtaining more accurate semantic ratings of each facial image by constructing the corresponding semantic expression vector.

Although the quality of 3D Gabor library based recognition is typically high, the process is slow, costly. On the other hand, the general input devices are 2D-based, and it is difficult to collect 3D data for emotion recognition. Therefore, using simplified feature representation with a generic face model will be an appropriate solution for real time applications. We then present an automatic emotion recognition system from video sequences using a 3D physical face model with EBS technique.

Chapter 4

Face Detection

FACE detection is considered to be essential requirements for intelligent vision-based human computer interaction systems. Automatic face detection is considered to be the first primary step for our emotion recognition system from video sequences. In this Chapter, we present a robust and effective method to detect human faces that combines feature extraction and face detection based on local normalization, Gabor wavelets transform and Adaboost algorithm. The main contribution of this step is the incorporation of a normalization technique based on local histograms with OAC technique to alleviate a common problem in conventional face detection methods: inconsistent performance due to sensitivity to variation illuminations such as local shadowing, noise and occlusion. The approach uses a cascade of classifiers to adopt a coarse-to-fine strategy for achieving higher detection rates with lower false positives.

4.1 The State of the Art

Face detection techniques have been studied extensively in the past decades. Examples include feature based methods, using geometric information such as skin color [49-52], geometric shapes [53, 54], motion information [55-58], and machine-learning based approaches like neural networks [59-62], Gaussian mixtures [63], support vector machines [64-66] and statistical modeling [67, 68].

Facial feature based techniques use prior knowledge about the face's features. Low level feature analysis first deals with the segmentation of visual features using edges, intensity skin color, motion, or generalized measures, including those based on template matching where several correlation templates are used to detect local sub-features, considered as rigid in appearance or deformable. The visual features are organized into a more global concept of face through facial feature and constellation analysis using face geometry constraints. The main drawback of feature-based approaches is that these systems are often simple, but do not work well in practice. The global constraints and extracted features can be significantly influenced by noise, occlusions, changes in face expression and viewpoint, and motion information may be distracted by alternate motion in the video.

Machine-learning based techniques have been developed to handle difficult scenarios where multiple faces of different sizes and poses have to be detected in heavily cluttered

backgrounds. They require no prior, or relatively minimal, knowledge of what constitutes a face, and detect the type of faces they have been trained or defined on. Pre-classified face samples and non-face samples are usually used for training the system and are mostly processed offline. These approaches avoid the specific and possibly inaccurate face modeling by learning underlying rules contained in highly variable face patterns from large training sets of face examples. They have proven to be very tolerant to noise and distortions affecting the face patterns. However, most systems are complex and computationally expensive, and research is ongoing to improve the real-time performance of those systems.

The information between consecutive frames is highly correlated, so this property can be exploited for faces detection. The shortest weighted feature distance between the face pattern and the possible face candidates can be used for this intent. And also, the weights of the features can be updated adaptively in the successive frames. X. Li and X. Zhou [69] proposed a spatial-temporal mutual feedback scheme for faces detection and tracking. The prediction model and the observation model were generated from two consecutive frames and the templates were updated through Kalman filter and hypodissertation test. V. Pallavi et al [70] constructed a directed weighted graph to detect and track players in videos. Directed Graph Construction linked candidates in a frame

with the candidates in next consecutive frames, and then Trajectory Estimation was applied to find the specific trajectory for each player. However, their works need initialize manually.

On the other hand, most of the automatic face detection algorithms are implemented in such a way that at the initial step, a pyramid of downscaled copies of the given input image is produced. Then, a sliding window scans each of the downscaled images and finally a classifier is applied on all possible window locations to decide whether the region covered by the window contains a face object or not. Practically, the number of windows or equivalently the number of times the classification will be processed is typically in the tens of thousands depending on the image size and demagnification factor. Adaboost algorithm, suggested by Viola and Jones [71] originally in 2001, employed this method in a fast and robust way and has been widely investigated in video face detection systems such as face tracking or video surveillance. This algorithm is based on the observation that the presence probability of face objects in a scene is substantially smaller compared to that of non-face objects or, in other words, the fraction of non-face region in an input frame is relatively much larger than that of faces. The key point is that fast, but less discriminating classifiers can reliably reject most of the windows containing non-face objects while passing the windows containing the maybe-

face objects to a second level classifier, this is slower than the previous one but has higher discriminating power. This procedure iteratively continues rejecting windows containing non-face and passing the windows containing maybe-face objects to a higher-level classifier. Such an iterative process can provide high detection performance with much less computational expense. This approach combines a set of efficient classifiers in a cascaded structure to achieve fast front face detection, but has not been investigated on images in changing illumination conditions or under occlusions.

Since illumination is one of the most important factors that determine success or failure in face detection, many approaches have been proposed to handle the illumination problem. Such approaches include illumination insensitive representations, modeling of illumination variations and illumination normalization to a canonical form. From the theory of illumination-reflectance model, an image can be expressed in terms of its illumination and reflectance components. Most algorithms of face detection presume that the illumination variation is uniform or lighting must be controlled. X. Tan and B. Triggs [72] provided a system under uncontrolled lighting based on robust preprocessing and an extension of the local binary pattern (LBP) local texture descriptor. Georghiades et al. [73] demonstrated that face images with the same pose, under different illumination conditions, form a convex cone, the illumination cone. Ramamoorthi [74] and Basri and

Jacobs [75] independently used spherical harmonic representation to explain the low dimensionality of face images under different illumination conditions.

Li et al [76] presented a method for indoor, cooperative-user applications, including active near infrared (NIR) imaging hardware, algorithms, and system design, to overcome the problem of illumination variation: an illumination invariant face representation is obtained by extracting LBP features from NIR images. However, their solution is developed for cooperative user applications indoor and is not yet suitable for uncooperative user applications such as face recognition in video surveillance, nor is it suitable for outdoor use. W. Chen et al [77] proposed an illumination normalization approach using a discrete cosine transform (DCT) to compensate for illumination variations in the logarithm domain. Since illumination variations mainly lie in the low-frequency band, an appropriate number of DCT coefficients are truncated to minimize variations under different lighting conditions. Moreover, the advantage of their approach is that it does not require any modeling steps and can be easily implemented in a real-time face recognition system. Nevertheless, the shadowing and specularity problems are not sufficiently addressed because they lie in the same frequency band as some facial features. Furthermore, higher frequency facial features are more difficult to extract while poses and expressions change.

As a conclusion, the illumination component of an image varies a great deal, often more than the reflectance component. The non-uniform illumination will change the rules of human face gray level distribution, and the edge of face will be blurred and detection rates commonly drop quickly under this condition. In order to achieve advanced illumination invariant face detection in complex conditions, more robust and fast methods are required. Moreover, most previous face recognition and detection systems imposed strict restrictions on the input data and worked with the assumption that the location of the face within a frame is known. Although their works obtained good detection results, the requirement for calibrated multi-cameras and only being used for certain specific applications are two main limitations of these systems.

4.2 Methodology

The key step of this work is the incorporation of local normalization with OAC technique to conventional classifier for automatic face detection on video sequences. Each frame of the input video sequences is first extracted and regularized by local normalization. Face candidate regions are then roughly located by OAC. The Gabor wavelets filters are applied for local feature extraction after preprocessing. In the final step, the face region is

detected through a cascade classifier consisting of detectors with Adaboost algorithm. The system diagram of this work is shown in the Figure 4.1.

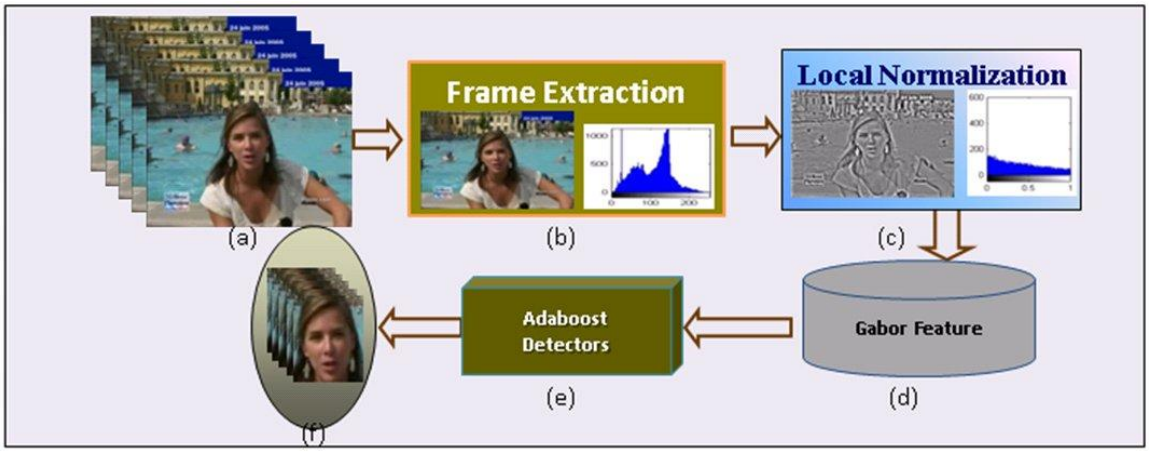


Figure 4.1 Face Region Detection (a) Video/Webcam Input, (b) Image from frame extraction, (c) Normalized image by local normalization, (d) Feature Extraction, (e) Classification, (f) Final face detection results

4.2.1 Local Normalization

Due to the fact that variant light condition definitely causes low detection rates and can be eliminated by illumination normalization, normalization techniques should be well considered in an automatic face detection system. The system's resistance can be evaluated to the most common classes of natural illumination variations. Most methods explored were typically characterized by relatively low spatial frequencies. We use local normalization in this important step in order to keep all the useful information in illumination invariant form. This facilitates accurate and robust feature extraction and face detection.

Illumination Compensation

Illumination compensation consists of several stages, including gamma intensity correction (GIC), difference of Gaussian (DoG), local histogram matching (LHM) and local normal distribution (LND).

The gamma correction of an image is a nonlinear gray-level transformation that replaces the input image s with its exponentiation s^{γ}. GIC, as mentioned in [78], corrects the overall brightness variation of the input image s to best match a predefined canonically illuminated image s_0. The predefined image s_0 is lighted under normal lighting condition. Given an input image $s(x, y)$, its GIC corrected image $s'(x, y)$ is computed by transforming the input image over its position (x, y) pixel by pixel with an optimal Gamma coefficient γ^*

$$s'(x, y) = G(s(x, y); \gamma^*) \tag{4.1}$$

where $G(s(x, y); \gamma) = c \cdot s^{1/\gamma}(x, y)$, c is a gray stretch parameter, and γ^* can be computed as

$$\gamma^* = \arg\min_{\gamma} \sum_{x,y} [G(s(x, y); \gamma) - s_0(x, y)]^2 \tag{4.2}$$

where $s_0(x, y)$ represents the predefined image. GIC can enhance the local dynamic range of the face in dark or shadowed regions, compress in bright regions and at highlights, and compensates for global brightness changes of an image. In our

implementation of GIC, γ^* is approximated using the golden section search with parabolic interpolation proposed in [78].

The intensity gradients such as shading effects are removed through a DoG filter [79], which is a popular method to obtain the resulting bandpass behavior for images and is defined as

$$\begin{aligned} DoG(x,y) &= L(x,y,D_2) - L(x,y,D_1) \\ &= \frac{1}{2\pi D_2^2}\exp[-(x^2+y^2)/2D_2^2] - \frac{1}{2\pi D_1^2}\exp[-(x^2+y^2)/2D_1^2] \end{aligned} \tag{4.3}$$

where L is the Guassian function and D_1 and D_2 are the deviations of inner and outer Gaussians respectively. The selected values of smaller or inner Gaussians D_1 are typically quite narrow so the detailed spatial information in high frequency is kept, while the outer ones D_2 might have more contents for low frequency.

We then apply LHM after GIC and DoG. Histogram matching, also known as histogram fitting or histogram specification, is the generalization of histogram equalization. The main idea of LHM is to produce an image with desired distributed brightness levels over local windows.

The gray levels of the input image s_k is firstly equalized by (4.4)

$$s_k = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n},\ k = 0,1,...,l-1 \tag{4.4}$$

where T denotes equalization function, n is the total number of pixels, n_j is the number of pixels with gray level r_j, and l is the number of discrete gray levels. The histogram distribution function $E(z)$ for the local window can be obtained by

$$E(z) = \sum_{0}^{z} p_z(z) \approx \sum_{i=0}^{z} \frac{n_i}{n} = s_k \tag{4.5}$$

where $p_z(z)$ represents the specified desirable probability density function (PDF) of local window. The inverse transformation function $z = E^{-1}(s)$ is applied to the levels obtained in (4.5). The new, revised version of the original image consists of gray levels characterized by the specified density $p_z(z)$ is then given by

$$z = E^{-1}(s) \quad \text{or} \quad z = E^{-1}[T(r)] \tag{4.6}$$

Finally, LND is applied on the resulting image from (4.6) by assuming the gray values drawn form a normal distribution. The output image $c(x, y)$ is normalized using (4.7)

$$c(x, y) = \frac{E^{-1}(s) - \mu_i}{\sigma_i} \tag{4.7}$$

where μ_i and σ_i are the mean and standard deviation of $E^{-1}[T(r)]$ over the whole image. This illumination compensation procedure can count the effects of illumination variations, local shadowing and highlights in the original image, which may preserve the essential elements for detection of visual appearances.

Candidates Selection

After illumination compensation, we propose to apply OAC technique [80] on the normalized images $c(x, y)$ to quickly locate face candidates. Compared with common automatic face detection algorithm, this method does not need to use the pyramid of downscaled copies of the input image and thus speeds up the processing.

A normalized image has similar power spectra and can be efficiently implemented in the spatial domain of a running window that approximately meets the requirements of the OAC process. The suggested algorithm is adaptive to the input normalized image, and designed to complete the segmentation in a single iteration in Hilbert space F , through the kennel transformation function. The transformation of the normalized image in F becomes a correlation image with normalized values ranging from zero to one. The OAC detector examines and segments this image according to the range of correlation values. The image is then split into two segments after correlation test that correspond to face candidates and background regions, which can be used conveniently by the later fine classifier with Gabor and Adaboost algorithm.

Assume we have a normalized image $c(x, y)$ with a multiple faces part $c_f(x, y)$ and a complex background part $c_b(x, y)$ that

$$c(x, y) = c_f(x, y) + c_b(x, y) \tag{4.8}$$

the face part c_f and background part c_b can be modulated as uncorrelated independent signals, so we have

$$c(x,y) = \sum_{k=1}^{l} \xi^k \otimes \bar{c}_f(x,y) + c_b(x,y) \tag{4.9}$$

where the face part c_f is now composed by the average face template $\bar{c}_f$ (or the egienface) though gain mapping matrix ξ, and l is the number of faces.

The OAC transform for the entire image is

$$H_C(X,Y) = C_f^T(X,Y) / \left|C_b(X,Y)\right|^2 \tag{4.10}$$

where C_f and C_b are the nonlinear mapping response in F of c_f and c_b respectively, T denotes the complex conjugate operation, and X,Y are the two-dimensional transform domain indices. The labeled graph (LG) generated by this transformation is the adaptive ratio of target (face) signal peak height to standard deviation of clutter (non-face) background of the image.

Since the background power spectrum $\left|C_b(X,Y)\right|^2$ is unknown, we instead estimate it from the spectrum of the entire image. Equation (3.10) then can be rewritten as

$$H(X,Y) = \frac{\bar{C}_f^T(X,Y)}{\left|C(X,Y)\right|^2} \tag{4.11}$$

where $\bar{C}_f^T(X,Y)$ is the nonlinear mapping response in F of $\bar{c}_f$. Now we can calculate the adaptive priori for a given image.

The OAC detector is defined as

$$D(X,Y) = g^T \otimes \left|C(X,Y)\right|^2 \tag{4.12}$$

where g is a m x m matrix and $m=5$ is assumed in this work.

We then use KCCA to get the nonlinear correlation between $D(X,Y)$ and $H(X,Y)$: to find a pair of directions ω_D and ω_H, such that the correlation $\rho(D,H)$ between the two projections ${\omega_D}^T$ and ${\omega_H}^T$ is maximized, where

$$\begin{aligned} &\rho(D,H;\omega_D\omega_H) \\ &= \frac{E\{\omega_D^T D H^T \omega_H\}}{\sqrt{E[\omega_D^T D D^T \omega_D]}\sqrt{E[\omega_H^T H H^T \omega_H]}} \end{aligned} \tag{4.13}$$

Assume $\Phi(D)$ is the diagonal of D in F so the correlation function in F can be formulated as

$$\begin{aligned} &\rho(\Phi(\mathrm{D}),H;\omega_{\Phi(D)},\omega_H) \\ &= \frac{\omega_{\Phi(D)}^T \Phi(D) Y^T \omega_H}{\sqrt{\omega_{\Phi(D)}^T \Phi(D)(\Phi(D))^T \omega_{\Phi(D)}}\sqrt{\omega_H^T H H^T \omega_H}} \end{aligned} \tag{4.14}$$

Therefore, to roughly locate face candidates is equivalent to finding a pair directions $\omega_{\Phi(D)}$ and ω_H that maximizes ${\omega_{\Phi(D)}}^T \Phi(D) Y^T \omega_H$ under the constraints

$${\omega_{\Phi(D)}}^T \Phi(D)(\Phi(D))^T \omega_{\Phi(D)} = {\omega_H}^T H H^T \omega_H = 1 \tag{4.15}$$

where $\Phi(D) = [\Phi(D_1), \Phi(D_2)...\Phi(D_N)]$ in the kennel space.

Let $\{(\omega^i{}_{\Phi(D)}, \omega^i{}_H)\}_{i=1}^t$ be the t pair directions of OAC, and $\rho_1,...\rho_t$ the t corresponding correlation values. Let a_i and b_i be the projections of the variables $\Phi(D)$ and H onto

the projection vectors $\omega^{i}{}_{\Phi(D)}$ and $\omega^{i}{}_{H}$ respectively. Thus, we get

$$a_i = (\omega^{i}{}_{\Phi(D)})^T \Phi(D)$$

$$b_i = (\omega^{i}{}_{H})^T H \tag{4.16}$$

where $i = 1,...,t$, a_i and b_i are approximately linearly correlated and both are centered projections, so that

$$a = [a^1,..a^t] = [\omega^1_{\Phi(D)},.....,\omega^t_{\Phi(D)}]^T \Phi(D) = P_D^T K$$

$$b = [b^1,..b^t] = [\omega^1_H,.....,\omega^t_H]^T H = P_H^T H \tag{4.17}$$

where

$$P_D = [(\omega^1_{\Phi(D)})^T,.....,(\omega^t_{\Phi(D)})^T]$$

$$K = (\Phi(D))^T \Phi(D)$$

$$P_H = [\omega^1_H,.....,\omega^t_H]$$

For a given normalized image, we can estimate the face candidates in the segmented image by the OAC value of $D(X,Y)$ and $H(X,Y)$ in term of a_i and b_i

$$C^* = \arg\max(a_i / b_i) \tag{4.18}$$

The segmentation image mask $M(x,y)$ for the original image $s(x,y)$ is then generated from the correlation image $c(x,y)$ as

$$\mathrm{M(x,\ y)} = \begin{cases} 0 & \mathrm{Th}_{c(x,y)} < C^* \\ 1 & \mathrm{Th}_{c(x,y)} \geq C^* \end{cases} \tag{4.19}$$

where $Th_{c(x,y)}$ is the thresholds parameter for pixels corresponding to the face candidates. Local normalization with nonlinear histogram equalization is used by taking into account histogram distribution over local window and combining it with global histogram distribution. Examples of the local normalization filtered results of the original images are shown in Figure 4.2.

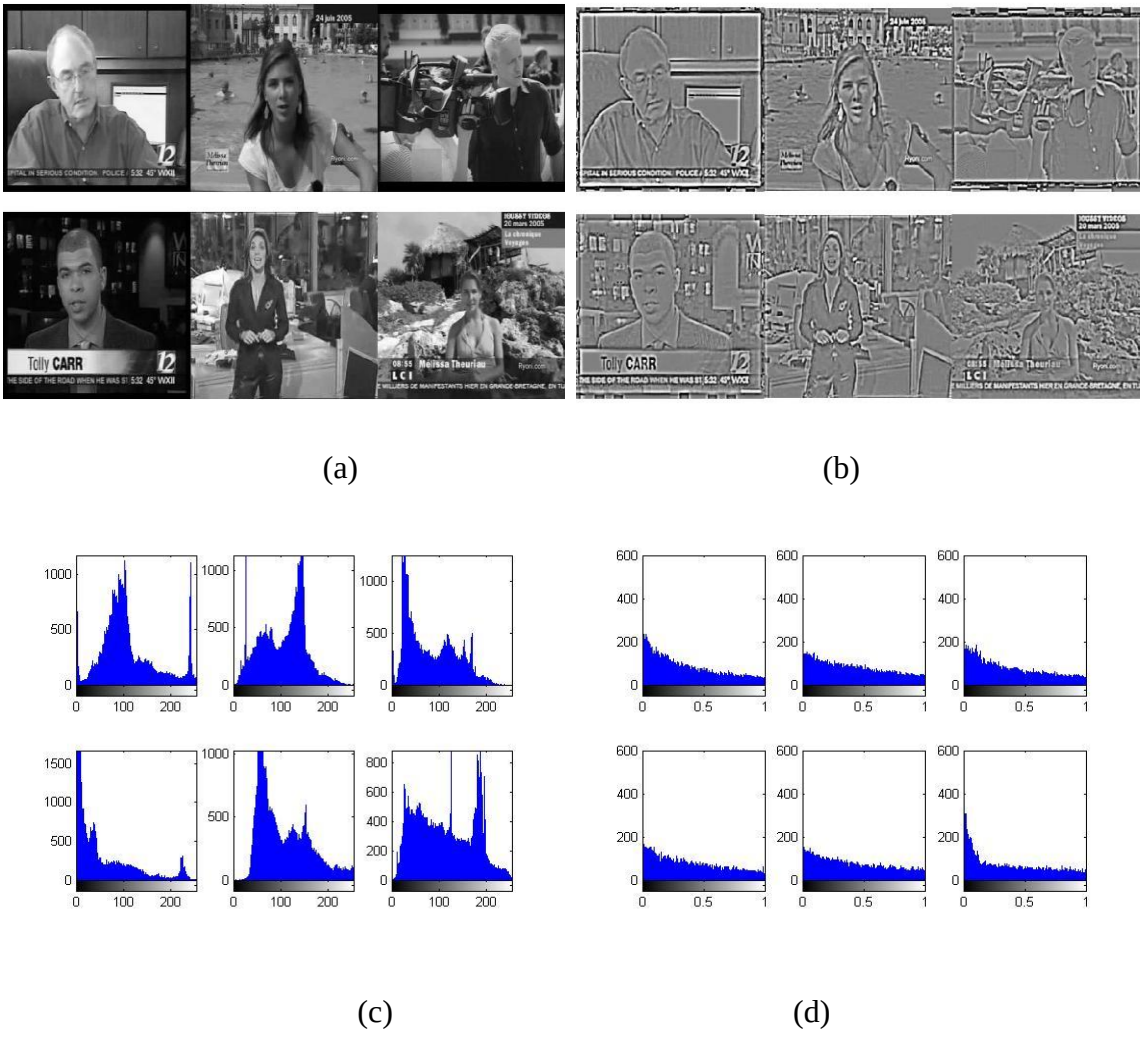


Figure 4.2 Samples of local normalizations for video sequences (a) Input images from video sequences, (b) Corresponding locally normalized images, (c) Histograms of original images, (d) Histograms of corresponding locally normalized images.

Figure 4.2 shows that the histograms of all input images are widely spread to cover the entire gray scale by local normalization, and the distribution of pixels is not too far from uniform. As a result, dark images, the histogram components of which are concentrated on the low side of the gray scales, bright images, the histogram components of which are biased toward the high side, and low contrast images, the histogram components of which are narrow and centered toward the middle of the gray scale, are much enhanced to have an appearance of high contrast. By applying local normalization, the system resistance to the natural illumination variations is improved.

4.2.2 Feature Extraction and Classification

The face detection stage of this step consists of two main components: Gabor wavelets feature extractions and Adaboost detection algorithms. Gabor wavelets demonstrate two desirable characteristics, spatial locality and orientation selectivity. Compared with other popular transformations, such as Fourier transform, DCT and various wavelets, the Gabor wavelets transform has shown its effectiveness in automatic face detection and recognition, and has been widely used by many researchers.

Gabor wavelet filters are applied for feature extraction. To design Gabor filter banks, we use 4 different scales and 8 orientations of Gaussian wavelets. The filter bank has 32 Gabor kernels with $k \in \{0,\ldots,7\}$ and $s \in \{0,\ldots,3\}$, here k represents orientation and s

represents scale. For training the original detector, a total of 15599 subjects (8754 positives and 6845 negatives) are used, and each set has 54x48x32=82,944 Gabor kernel features. All the features are trained through cascade Adaboost classifiers.

Boost algorithm has been proposed to reduce the redundancies of the high dimensional feature space and computational cost. The Adaboost algorithm by Viola and Jones4 for face detection is a typically successful example as it has a very low false positive rate and can detect faces in real time. It can be trained for different levels of computational complexity, speed and detection rate which are suitable for specific applications.

(a) (b)

Figure 4.3 Samples of face images and non-face images (a) face images, (b) non-face images.

For the training of the original face detector, we collect face images and non-face images from the publicly available face detection databases with large illumination variations: Extended Yale Database and Carnegie Mellon University (CMU) Database.

Figure 4.3 shows some samples of the training images. The original detector is trained to detect a face centered in a standard window with size of 54x48, and all training images are resized to 54x48 pixels.

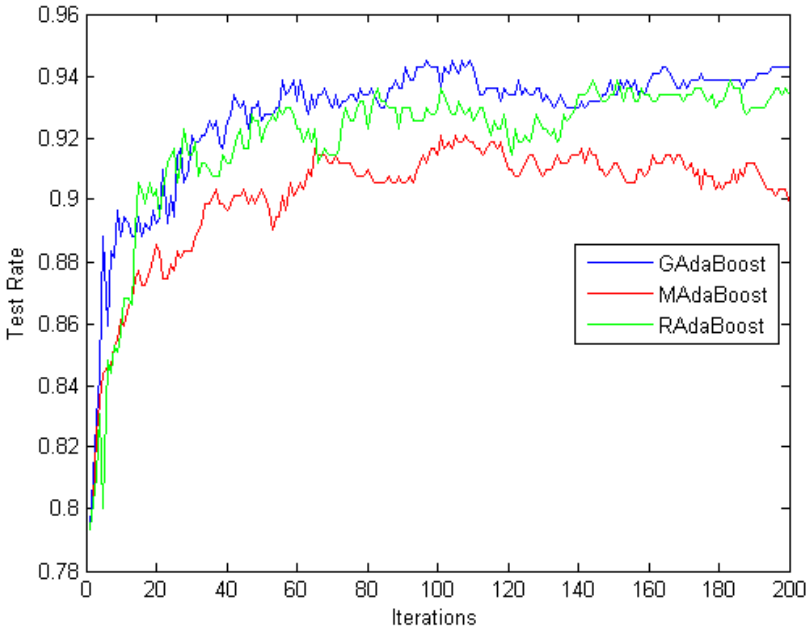


Figure 4.4 Test rates from three Adaboost algorithms

The performances of RealAdaboost [81], GentleAdaboost [82] and ModestAdaboost [83] for face detection are compared in our work based on video sequences using GML AdaBoost Matlab Toolbox [84]. RealAdaboost is the generalization of a basic Adaboost algorithm and treated as a fundamental boosting algorithm. GentleAdaboost is a more robust and stable version of RealAdaboost. It is identified that GentleAdaboost performs slightly better than RealAdaboost on regular data, and is considerably better on noisy data.

It is also much more resistant to outliers. ModestAdaboost is regularized tradeoff of Adaboost, it is mostly aimed for better generalization capability and resistance for certain specific sets of training data. RealAdaboost, GentleAdaboost and ModestAdaboost are compared for error checks with 200 boosting iterations (shown in Figure 4.4). GentleAdaboost returns better face detection rate in our work, and is selected as the detection algorithm for our system.

4.3 Experiment and Results

In this section, we evaluate the performance of the proposed method for human face detection and segmentation on two video datasets with different illumination conditions. One test video dataset is recorded under good brightness condition. It includes eight subjects and 520 video clips in total. All samples are running at 30 frames per second on images of 320x240 resolutions. Another dataset (647 clips in total) are from commercial films and videos available on the Web under complex illumination conditions. Videos in this dataset also contain single or multiple faces occurring at different sizes, in different poses, and at various positions with respect to each other. Note, the videos with good light conditions were collected for the purpose of human emotion recognition. Each human subject showed the six fundamental human emotional states: happiness, sadness,

anger, disgust, fear and surprise. The variations among the emotional states make the face detection task more challenging since the training images were essentially photographed in the neutral state.

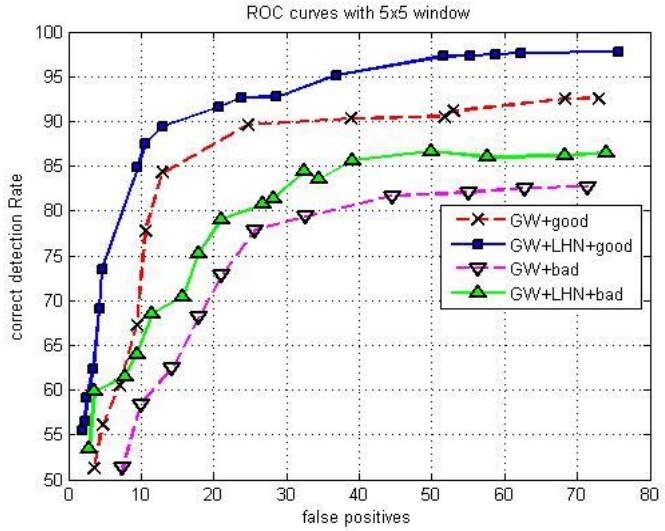
Table 4.1 Comparison of parameters used in experiments

Parameters	Set I	Set II
Local window size	5x5	8x8
GIC	γ=0.23	γ=0.25
DoG	$D_1 = 1, D_2 = 2$	$D_1 = 1, D_2 = 3$
LHM	K=255	K=255
LND	$\mu_i = 8, \sigma_i = 0.1$	$\mu_i = 10, \sigma_i = 0.1$

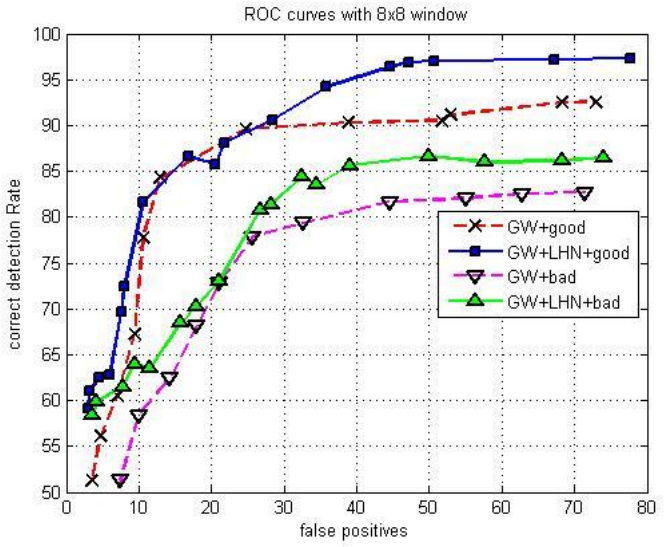
We performed the face detection with two different sets of parameters, listed in table 4.1. We first demonstrate the overall performance of the system in term of receiver operating curves (ROC) as shown in Figure 4.5.

From Figure 4.5 we can see that good detection results are obtained by setting the window size of the local normalization to *5*x*5*. We perform detections using Gabor wavelets (GW) features only and then we use the combined features of GW and local normalization (LN). In each figure, GW and LN+GW methods are also compared by correct detection rates vs. false positives in different illumination conditions. The experiment results demonstrated that the face detection accuracy is considerably

improved by about 10 to 15 percent by incorporating local normalization in the critical regions of detection rate vs. false positives. At the same time, the false detection rates are dropped by approximately 15 percent.



(a)



(b)

Figure 4.5 The ROC curves of face detection results with local normalization (a) shows detection result with parameter set I, (b) shows detection result with parameter set II.

Figure 4.6 Sample sequences from the test videos under good illumination condition, trained face detector with local normalization are applied on each frame. The frame numbers are marked above.

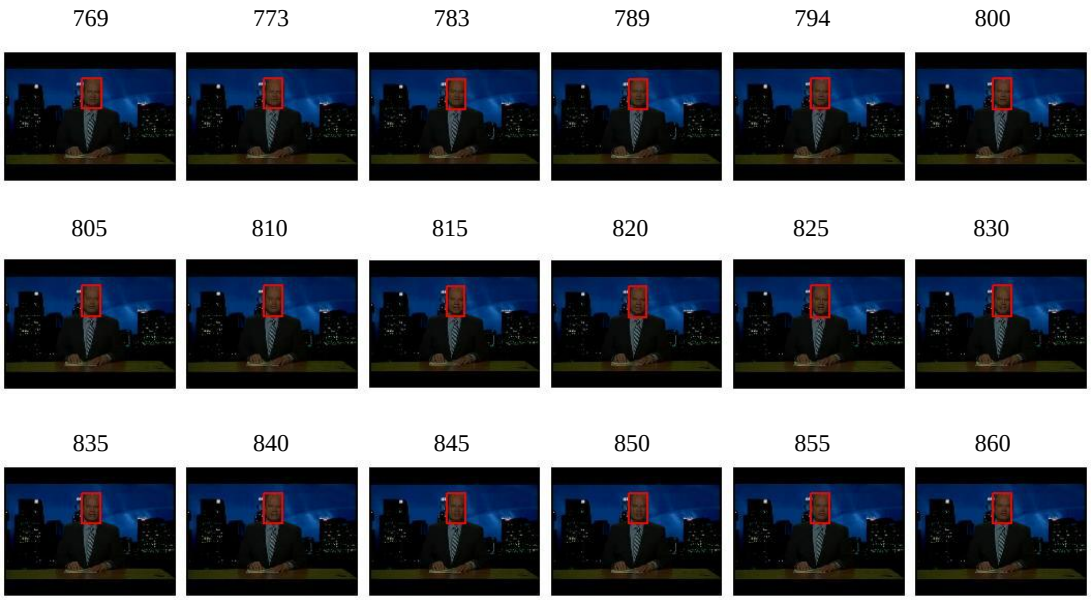


Figure 4.7 Sample sequences from the test videos under bad illumination condition

We then present some representative cases. Figures 4.6 and 4.7 show the face detection results by applying the proposed detectors to each frame of video sequences under good illumination conditions and bad illumination conditions. The size of the bounding box is

determined using the scale of the detected face on the image. From the results we can see that a face is actually detected even under varying illumination condition.

The results are carried out through pixel-based illumination normalization and are sensitive to optimal threshold $Th_{c(x,y)}$ of the local normalization, which was computed through (4.19). For a threshold higher than $Th_{c(x,y)}$, some of the faces in the sequences will not be detected, and on the other hand, much lower thresholds will lead to some spurious faces. Next we evaluate our method on the video sequences with rotating pose and varying sizes, the results are shown in Figure 4.8.

Figure 4.8 Sample sequences with changing size and head rotation

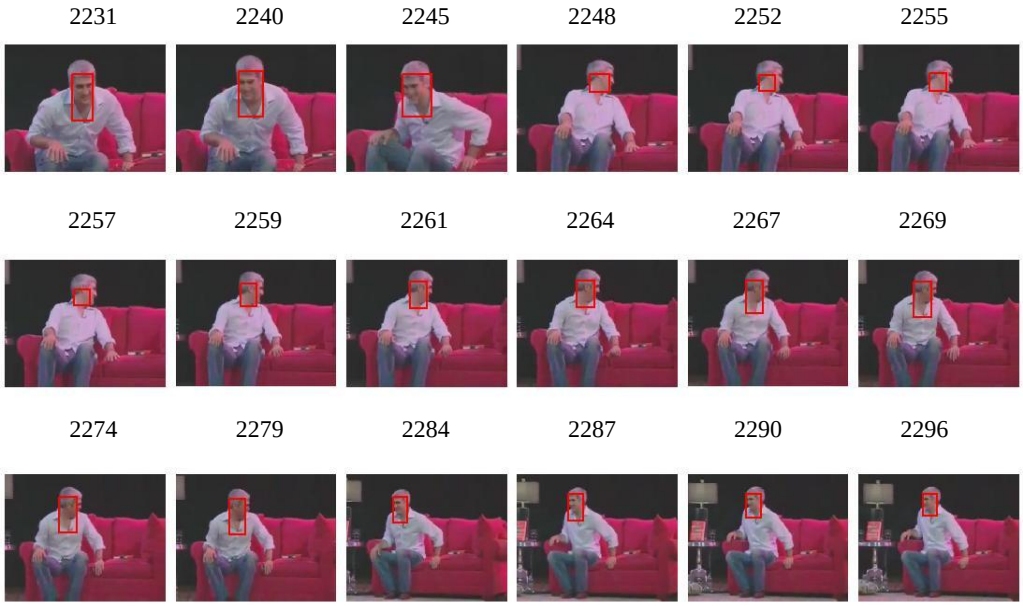


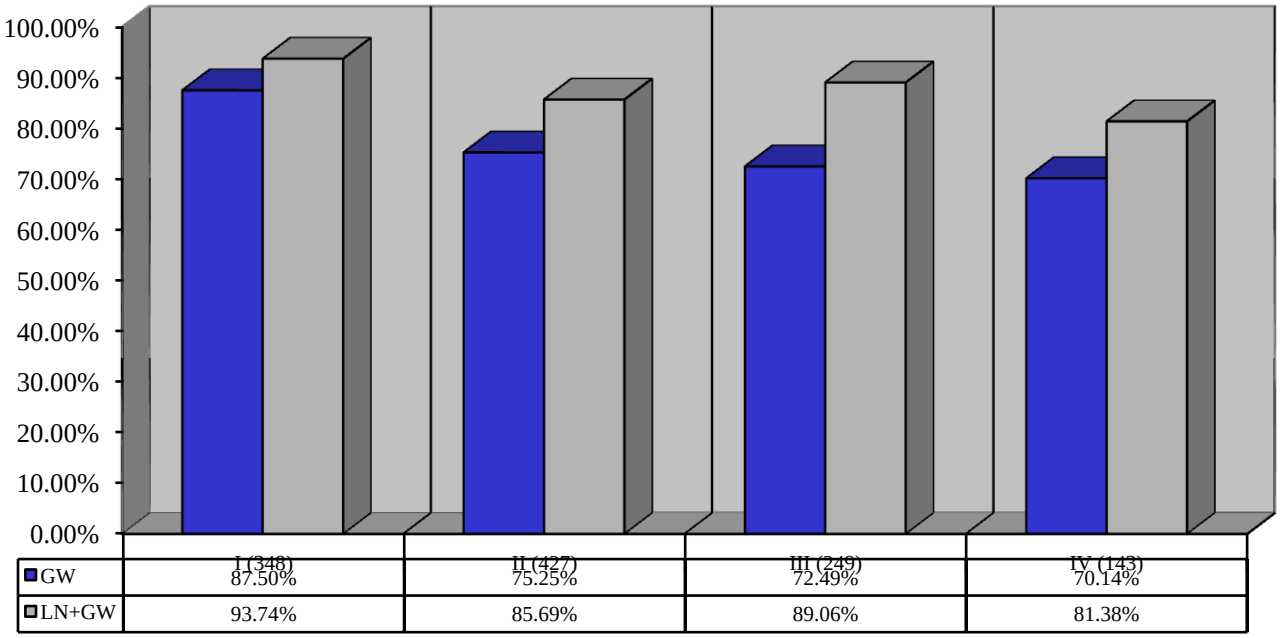
Figure 4.9 Sample sequences with head rotating after profile data are trained

Figure 4.10 Face Detections applied on sample sequences with multiple subjects

In Figure 4.8 we can see the face is lost after frame 2248 when a frontal face is rotating to a profile view. From frame 2284 and onward, the face is found again when the frontal face reappeared. This happened because, so far, we only trained our detectors using

frontal face images. To solve this problem, we added a wider range of face pose samples to the training database in order to improve the performance of the detectors. The improved result is shown in Figure 4. 9In Figure 4.10, we show the results of applying the proposed detectors in a sequence in which various faces appeared.

Table 4.2 Final detection rates for testing databases



	I (348)	II (427)	III (249)	IV (143)
GW	87.50%	75.25%	72.49%	70.14%
LN+GW	93.74%	85.69%	89.06%	81.38%

Two faces are in the center of the frames with slight movement and rotation. The third face first appeared on the left side of frame 18249, and then moves close to the faces in the center, stays a while, and leaves the scene in frame 18295 and onward. Our detectors are able to detect almost every face without dramatic movement and rotation, such as those in the center of all the frames in Figure 4.10. For the intensely varying objects, such

as the third face in frames 18249 to 18275, missing detection does occur as seen in frame 18265. However, it is redetected after a few frames. This example shows the robustness of our method to the various faces detection.

The final detection rates for different condition are given in table 4.2. Columns I, II, III, IV indicate video sequences with good illumination conditions, bad conditions, changing head poses/sizes and various faces, respectively; GW and LN+GW denote detection results using Gabor wavelets features only and those using combined features, respectively. As a result, the average face detection rate is considerably improved to about 10 percent by incorporating the proposed local normalization technique.

4.4 Chapter Summary

In this Chapter, we presented an effective and robust method for detecting faces in video sequences based on a coarse-to-fine strategy. Local normalization technique is incorporated into a conventional face detector to alleviate illumination variation problem. It is demonstrated that the method can improve the face detection rate and reduce the processing time. Compared with face detection without local normalization, our method has following advantages:

1. alleviate illumination variation problems in general face detection systems,

2. decrease computing time by candidates localization with optimal adaptive correlation techniques,
3. locate faces automatically on single frame and make it possible to eliminate the manual initiation step from head/face tracking algorithm, and
4. be able to deal with detecting various faces reliably

Chapter 5

Fiducial Point Detection and Tracking

FIDUCIAL points are a set of facial salient points, usually located on the corners, tips or mid points of the facial components. To be able to reasonably recognize the emotional expressions, the current appearance of the facial features must first be detected. To do so, one promising approach is to detect and track a set of fiducial points, the locations of which alter as the current appearance of the facial features changes with the facial expression. Due to high computing complexity, fiducial point techniques were not well studied. Automatically detecting and tracking fiducial points can extract the prominent characteristics of facial expressions with the distances between points and the relative sizes of the facial components to form the feature vector. On the other hand, finding feature points appropriately on the face can best represent the most important characteristics of the expressions and extract features more easily.

We choose 26 fiducial points on the face region, which are shown in Figure 5.1. The 26 fiducial points are selected according to the anthropometric measurement with the

maximum movement of the facial components during expressions. Table 5.1 gives the descriptions of these fiducial points. The fiducial points are detected in each fixed facial region by scale invariant feature based detectors, and tracked using multiple DE-MC particle filters with kernel correlation techniques.

Figure 5.1 Selected 26 Fiducial Points

Table 5.1 Description of the 26 fiducial points

Fiducial Points Description			
1	Top of the head	14	Top of the left eyebrow
2	Tip of the chin	15	Left eyebrow outer corner
3	Left of the head	16	Right eyebrow inner corner
4	Right of the head	17	Top of the right eyebrow
5	Left eye inner corner	18	Right eyebrow outer corner
6	Top of the left eye	19	Top of the nose
7	Left eye outer corner	20	Left nose corner
8	Bottom of the left eye	21	The medial point between left and right nostril centres
9	Right eye inner corner	22	Right nose corner
10	Top of the right eye	23	Left corner of the mouth
11	Right eye outer corner	24	Top of the upper lip
12	Bottom of the right eye	25	Right corner of the mouth
13	Left eyebrow inner corner	26	Bottom of the lower lip

5.1 Fiducial Point Detector

Automatically detecting fiducial points successfully in facial region plays an important role in numerous facial image interpretation tasks. We propose an automatic and robust fiducial point detection method using the scale invariant feature and the Adaboost algorithm for classification in this step. After the facial region is located from the face detection step, candidate points are selected over the facial region using local scale space extrema detection. The scale invariant feature of each candidate point is extracted for examination. We build the fiducial point detectors with Adaboost classifiers. All the candidate points in the facial region are examined through these detectors, and the 26 fiducial points can be detected.

Using fiducial points for facial expressions recognition is a challenge in computer vision systems. Most of the automatic feature point detection algorithms are implemented in such a way that every pixel in the input image is examined through feature detectors one by one to construct the feature vectors. Then the classifications are applied to the feature vectors to perform the feature point detection. Practically, the number of feature points, or equivalently, the number of times the classification will be processed, is typically in the tens of thousands, depending on the image size and demagnification

factor. We propose to use the scale space extrema method to efficiently detect the locations of candidate points in the facial region from the face detection step. Compared with common automatic feature point detection algorithm, our proposed method does not need to classify every pixel of the input image and thus speeds up the processing.

5.1.1 Candidate Selection

The scale space extrema can be detected using the Gaussian convolution kernel function convolved with the input image. The description function *L (x, y)* of the input image in different scale space is expressed as:

$$L(x, y, \sigma) = G(x, y, \sigma) * s(x, y) \tag{5.1}$$

L(x, y, σ) is the spatial scale image, where *s(x, y)* indicates input image of the facial region, and *G(x, y, σ)* is the Gaussian convolution kernel function that:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp[-(x^2 + y^2)/2\sigma^2] \tag{5.2}$$

σ is the scale factor. The image zooms with the change of *σ*, and then a series of scale images could be obtained. The scale space extrema are computed by the DoG function of the input image, which calculates the difference of two nearby scales separated by a constant multiplicative factor *k*

$$\begin{aligned} D(x, y, \sigma) &= [G(x, y, k\sigma) - G(x, y, \sigma)] * s(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{5.3}$$

where *D(x, y, σ)* is the DoG function of the input image. Each pixel in the DoG image is compared to its eight neighbors on the same scale, and each of its nine neighbors one scale up and down. The pixels with the local maximal or minimal values are chosen as candidate points, including the adjacent scale, the position and scale of the local extreme points. The points are generally the feature points of the image, located on contours, corners and edges.

5.1.2 Feature Vector Generation

After the position and scale σ of the candidate points are determined from the input image, a gradient orientation histogram is calculated for the direction of each interest point in its neighborhood. The gradient magnitude m (x, y) and orientation θ(x, y) are computed using pixel differences as:

$$m(x, y) = \sqrt{[L(x+1, y) - L(x-1, y)]^2 + [L(x, y+1) - L(x, y-1)]^2} \quad (5.4)$$

$$\theta(x, y) = \arctan[\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}] \quad (5.5)$$

where, L is the scale feature of the images. By choosing a neighborhood, $\boldsymbol{F}$, the center of each interest point and calculating the directions of points in $\boldsymbol{F}$, we can obtain the direction distribution and the statistical histogram. The gradient magnitude orientation is divided into 36 portions so as to be convenient in obtaining the direction distribution. The

direction of the candidate point is the maximal component of the 36 phases in the statistical histogram.

5.1.3 Fiducial Point Detector

To detect the fiducial points from the candidate points, a set of fiducial point detectors, with the feature description for the gradient orientation histogram of the input images, are constructed.

At the centre of each fiducial point, a 16×16 pixel neighborhood window - $\boldsymbol{F}$ is selected and divided into 16 subregions by 4×4. Using (5.4) and (5.5), the directions and amplitudes for all pixels in the subregions are obtained, and then accumulated into orientation histograms summarizing the contents over 4×4 subregions. These are eight direction distributions in the ranges of $(0,\pi/4,\pi/2,3\pi/4,\pi,5\pi/4,3\pi/2,7\pi/4,2\pi)$ with their lengths corresponding to the sum of the gradient magnitudes near that direction within the region. The amplitude and Gaussian function are also applied to the eight direction distributions to create the direction statistical histogram of subfields. The feature descriptor of each fiducial point is obtained by connecting the direction descriptions of all subfields. The total of the direction descriptions is 16, so the length of a fiducial point detector is 128=16×8, and should be normalized in order to ensure the illumination invariance.

5.2 Multiple Points Tracker

We use multiple DE-MC particle filters, which were first introduced by Du and Guan in [85], to track the fiducial points depending on the locations of the current appearance of the spatially sampled features which are automatically located in the initial step. Due to its generic nature, the DE-MC particle filter leads to a more reasonable approximation to the proposal distribution and hence considerably improves accuracy for tracking by building a path connecting a sampling with measurement.

A novel kernel correlation based observation likelihood is proposed for fiducial points with robust colour histograms. This likelihood attempts to deal with changes in the appearance of the face due to facial expressions. Furthermore, the fiducial points are tracked by utilizing prior knowledge on the facial feature configurations. We show how prior knowledge can be incorporated in our multiple DE-MC particle filter scheme.

5.2.1 The State of the Art

Automatically detecting and tracking fiducial points can extract the prominent characteristics of facial expression with the distances between points and the relative sizes of the facial components to form the feature vector. Due to high computing

complexity, automatic fiducial points' detecting and tracking for facial expression analysis have not been well studied in the past.

Cohn *et al.* in [86] presented an optical flow based approach, which automatically tracks the selected facial features with a hierarchical algorithm for estimating the optical flow. Maghami *et al.* [87] selected facial feature points from the first frame to the last using a maximum cross-correlation algorithm followed by a Kalman filter. The extracted feature vector was then given to different classifiers to classify the facial expressions within six basic emotions. The results showed that a Bayes optimal classifier can reach the average correct classification rate of 93.72% by this method. Lai *et al.* [88] proposed to use the integral optical density (IOD) to detect the fiducial points for near frontal face images and showed that the proposed algorithm was insensitive to the facial expression, small rotation, different types of glasses and hairstyle. Ersi and Kiana [89] presented a feature-based hybrid method to analyze local facial features located by a meta-version of the specification algorithm in the context of a LFA (Local Feature Analysis) technique. Fiducial points were determined based on genetic algorithm and the output points were decorrelated as much as possible. Michel Valstar and Maja Pantic [90] built an automatic facial expression recognition system from face video. Twenty fiducial points were detected by a localization method using individual feature GentleBoost templates. Then,

a particle filtering scheme was exploited to track the facial points. The AUs displayed in the input video and their temporal segments were recognized finally by Support Vector Machines (SVM). They achieved a 90.2% average recognition rate.

The particle filter, also known as a condensation or sequential Monte Carlo, is regarded as a powerful tool for computing posterior distributions and has been reported to have a better performance than the Kalman filter (KF) for multiple objects' tracking by some research groups [91-95, 100]. Many approaches have been investigated in multiple objects' tracking using particle filters [91]. Particle filters take a lot of forms across a variety of literature. Sminchisescu and Triggs [92] developed a proposal density based on uncertainty of local parameter estimation. The unscented particle filter (UPF) [93] drew samples from a proposal distribution which was determined by the calculation result of the unscented Kalman filter (UKF). In [94] the simulating annealing algorithm and the genetic algorithm (GA) together constituted the foundation of the annealed particle filter (APF). Particle filters are also applied for multiple targets or ambiguities when the posterior is multimodal. Hue, *et al* [95] developed a system for multiple targets tracking by expanding the state dimension to include component information. The Bayesian Multiple-BLob tracker [96] has an automatic object detection system that relies on modeling a fixed background. They used this model to identify foreground targets.

Vermaak, *et al* [97] introduced a mixture particle filter, where each component was modeled with an individual particle filter that formed part of the mixture particle filter.

Despite the successes in various applications, some difficulties still remain in fiducial point tracking tasks for particle filter techniques: the high dimensionality of the state space associated with the activities of facial expressions, high non-linear and non-Gaussian distributions of the observation models and targets. Moreover, when the object is small in appearance, cluttered background and occlusion lead to severe ambiguity. Meanwhile reliable detection is often unaffordable due to deficient features extracted from the object's small region in the image. It is observed that a traditional particle filter does not perform well when the dynamic system has a very small process noise, or if the observation noise has very small variance. In these cases, the particle set quickly collapses to one single point in the state space and the filter performance is severely affected.

To surmount these difficulties, in this step, multiple DE-MC particle filters are applied for fiducial point tracking and a kernel correlation analysis approach is proposed to improve the efficiency of sampling. Minimal amount computation is introduced by making use of the intermediate results obtained in particle allocation. From experimental results, our proposed method demonstrates impressive performance.

5.2.2 DE-MC Particle Filter

A particle filter provides a robust Bayesian framework for the visual tracking problem. It maintains a particle based representation of the *a* posteriori probability $p(X_k|Y_{1:k})$ of the state X_k given all the observations $Y_{1:k} = \{Y_1, Y_2, ..., Y_k\}$ up to and including the current time k instance, according to

$$p(X_k|Y_{1:k}) = \lambda_k p(Y_k|X_k) \int p(X_k|X_{k-1}) p(X_{k-1}|Y_{1:k-1}) dX_{k-1} \tag{5.6}$$

In (5.4), the state X_k is the location of a fiducial point while the observation set $Y_{1:k}$ is the set of image frames up to the current time instant. λ_k is a normalization constant that is independent of X_k. $p(X_k|X_{k-1})$ is the motion model that is conditioned directly on the immediate preceding state and independent of the earlier history if the motion dynamics are assumed to form a temporal Markov chain. The distribution is represented by discrete samples *N* through particle filtering. The *N* samples (particles) are drawn from a proposed distribution $p(X_k^{(i)}|X_k^{(i-1)}, Y_k)$, $i = 1,2,...,N$ and assigned with weights $w(X_k^{(i)})$.

Suppose that at a previous time instance $k - 1$, we have a particle based representation of the density, that is, we have a collection of *N* particles and their corresponding weights $\{X_{k-1}^{(i)}, w(X_{k-1}^{(i)})\}_{i=1}^N$. At time step *k*, select a new set of samples $\{\widehat{X}_{k-1}^{(i)}\}_{i=1}^N$ from $\{X_{k-1}^{(i)}\}_{i=1}^N$ with the probability proportional to $w(X_{k-1}^{(i)})$. The samples with a larger weight should be

selected with a higher probability. Then, applying a constant velocity dynamical model to the samples yields:

$$X_k^{(i)-} = \hat{X}_k^{(i)} + V_{k-1} \tag{5.7}$$

where V_{k-1} is the velocity vector computed in time step *k*-1. The particle set $\{X_k^{(i)-}\}_{i=1}^N$ then acts as the initial population for a *T*-iteration DE-MC processing. Then the weights of particles are subject to update by the DE-MC. At the end of this step, we take the output population as the particle set of current the time step $\{X_k^{(i)}, w(X_k^{(i)})\}_{i=1}^N$.

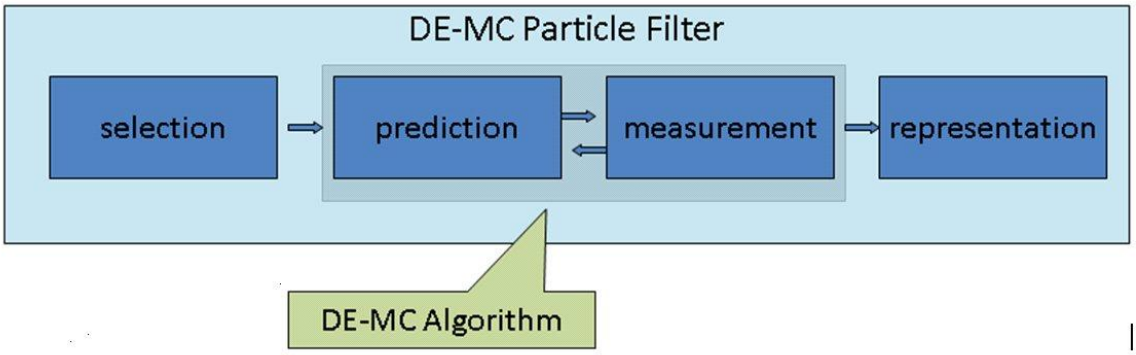


Figure 5.2 DE-MC Particle Filter

We then estimate the state at time step *k* as:

$$X_k = \arg\max_{X_k^{(i)}; i=1,\ldots,N} w(X_k^{(i)}) \tag{5.8}$$

and we update the velocity vector of current time step $V_k = X_k - X_{k-1}$. The step size of random jumping for current DE-MC iteration is reduced if the survival rate of the last DE-MC iteration is high or inflated otherwise. The algorithm of the DE-MC particle filter is shown in Figure 5.2.

5.2.3 Kernel Correlation Based Observation Likelihood

The kernel correlation based on HSV colour histograms is used to estimate the observation likelihood and measure the correctness of particles, since HSV decouples the intensity (value) from colour (hue and saturation) and corresponds more naturally to human perception. We set each feature point at the centre of a 16x16 pixel neighborhood window as the observation model. The kernel density estimate (KDE) $K(X_k)$ for the colour distribution of the object X_k at time step *k* is given as

$$K(X_k^{(i)};r) = \frac{1}{\zeta}\sum_{i=1}^{N}\frac{(c(X_k^{(i)}) - c(r))}{d^{i_x}} \tag{5.9}$$

where the $c(X_k^{(i)})$ indicates candidate region within a search region *R* centred at X_k of 40×40 pixels at time step *k*, which is sufficiently large to reach the maximum facial point movement without overlapping with any neighboring windows. $c(r)$ is a target region with *r* position translation in the search region *R*. ζ is a normalizing constant that ensures $K(X_k;r)$ to be a probability distribution, $\sum_{i=1}^{N} K(X_k;r) = 1$. The kernel width d^{i_x} is used to scale the KDE $K(X_k;r)$, and the optimal solution for kernel width d^{i_x} that minimizes the Mean Integrated Square Error (MISE) is given by

$$d_{opt} = \left(\frac{4}{(i_x+2)N}\right)^{1/(i_x+4)} \tag{5.10}$$

where i_x is the given particle set at time k and d_{opt} denotes the optimal solution for kernel width d^{i_x}. If we denote $K^*(X_k;r)$ as the reference region model and $K(X_k;r)$ as a

candidate region model, we can measure the data likelihood to track the facial points' movements by considering the maximum value of the correlation coefficient between the colour histograms in this region and in a target region. The correlation coefficient $\rho(X_k^{(i)};r)$ is calculated as

$$\rho(X_k^{(i)},r)=\left|\frac{\sum_{i=1}^{N}\sum_{r\in R}\left|K^*(X_k;r)-E(K^*(X_k;r))\right|\left|K(X_k;r)-E(K(X_k;r))\right|}{\sqrt{\sum_{i=1}^{N}\sum_{r\in R}\left|K^*(X_k;r)-E(K^*(X_k;r))\right|^2}\sqrt{\sum_{i=1}^{N}\sum_{r\in R}\left|K(X_k;r)-E(K(X_k;r))\right|^2}}\right| \quad (5.11)$$

where $E(K(X_k;r))$ and $E(K^*(X_k;r))$ are the means of the vectors $K(X_k;r)$ and $K^*(X_k;r)$, the average intensities of the colour model, respectively. Finally, we define the observation likelihood of the colour measurement distribution using the correlation coefficient $\rho(X_k;r)$ as:

$$p(Y_k|X_k^{(i)})=e^{-\frac{\rho^2(X_k^{(i)};r)}{\tau_i}} \quad (5.12)$$

where τ_i is a scaling parameter, which helps the result evaluated by (5.12) more reasonably be distributed in the range of (0,1).

5.2.4 Fiducial Point Tracking

In this section, we present the use of multiple DE-MC particle filters for fiducial point tracking over time. Once the observation model is defined we need to model the transition density and to specify the scheme for reweighting the particles. A single

particle filter weights particles based on a likelihood score and then propagates these weighted particles according to a motion model. Simply running particle filters for multiple fiducial point tracking needs a complex motion model for the identity between targets. Such an approach suffers from exponential complexity in the number of tracked targets. In contrast to traditional methods, our approach addresses the multi-target tracking problem of combining the colour based kernel correlation technique for the observation likelihood with a DE-MC particle filtering distribution. A set of weighted particles is used to approximate a density function corresponding to the probability of the location of the target given observations.

To avoid sampling from a complicated distribution, the M-component non-parametric mixture model is adopted for the posterior distribution over the state X_k of the all targets M according to:

$$p(X_k|Y_{1:k}) = \sum_{j=1}^{M} Pi_{j,k}\, p_j(X_k|Y_{1:k}) \tag{5.13}$$

where $M = 26$, $p_j(X_k|Y_{1:k})$ is the *a* posteriori probability of the fiducial points with the M-component non-parametric mixture model, and *Pi* is the mixture weights satisfying $\sum_{m=1}^{M} Pi_{m,k} = 1$. The likelihood $p(Y_k|X_k)$ is the measurement model and expresses the probability of observation Y_k. We utilize training data to learn the interdependencies between the positions of the fiducial points for the reweighting scheme. The motion

model $p(X_k|X_{k-1})$ predicts the state X_k given the previous state X_{k-1}. Using the filtering distribution computed from (5.13), the predictive distribution becomes:

$$p(X_k|Y_{1:k-1}) = \sum_{j=1}^{M} Pi_{j,k-1} p_j(X_k|Y_{1:k-1}) \tag{5.14}$$

where $p_j(X_k|Y_{1:k-1}) = \int p(X_k|X_{k-1}) p_j(X_{k-1}|Y_{1:k-1}) dX_{k-1}$. Hence, the updated posterior mixture takes the form

$$\begin{aligned} p(X_k|Y_{1:k}) &= \sum_{j=1}^{M} Pi_{j,k} p_j(X_k|Y_{1:k}) \\ &= \lambda_k \sum_{j=1}^{M} Pi_{j,k} p_j(Y_k|X_k) \int p_j(X_k|X_{k-1}) p_j(X_{k-1}|Y_{1:k-1}) dX_{k-1} \end{aligned} \tag{5.15}$$

where the new weights are given by:

$$Pi_{j,k} = \frac{Pi_{j,k-1} \int p_j(Y_k|X_k) p_j(X_k|Y_{0:k-1}) dX_k}{\sum_{l=1}^{M} Pi_{l,k-1} \int p_l(Y_k|X_k) p_l(X_k|Y_{0:k-1}) dX_k} \tag{5.16}$$

When tracking the multiple modalities, multiple trackers start with a mode-seeking procedure, the posterior modes are subsequently detected through the HSV colour histograms based kernel correlation analysis. Using a trained color-based observation model allows us to track different fiducial points. Here, we have M different likelihood distributions. At time k we sample candidate particles from an appropriate proposal distribution $\{\widehat{X}_{k-1}^{(i)}\}_{i=1}^{N}$ from $\{X_{k-1}^{(i)}\}_{i=1}^{N}$ and weight these particles according to the probability proportional:

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p(Y_k \left| \widehat{X}_k^{(i)} \right.) p(\widehat{X}_k^{(i)} \left| X_{k-1}^{(i)} \right.)}{p(\widehat{X}_k^{(i)} \left| X_{0:k-1}^{(i)}, Y_{1:k} \right.)} \tag{5.17}$$

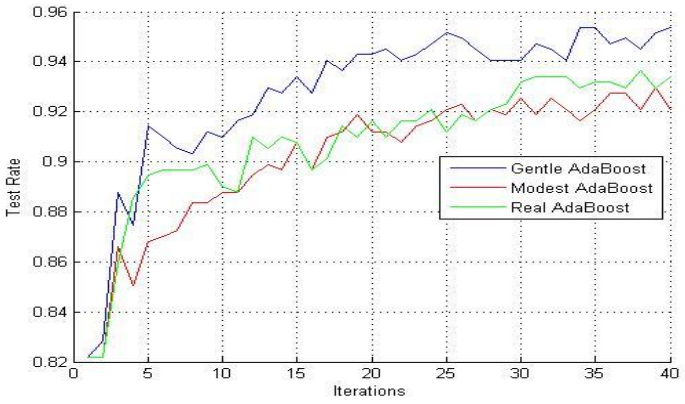
In our work, scaling is normalized by person-related scaling factors that are estimated from the positions of the facial features at the first frame, such as the dimensions of the mouth. This scheme simply processes with the prior knowledge by sampling from the transition priors and updating the particles using importance weights derived from (5.17).

5.3 Experiment and Results

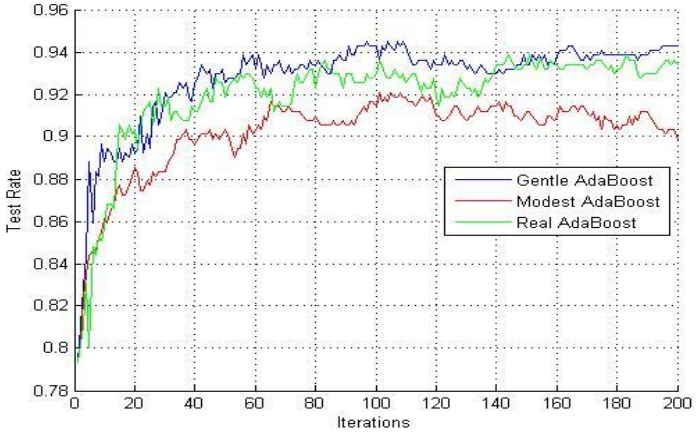
5.3.1 Detecting Result

In this section, we evaluate the performance of the proposed fiducial point detecting method using two video datasets, Cohn-Kanade database [98] and Mind Reading database [99]. The Cohn-Kanade database consists of approximately 2000 image sequences in nearly frontal view from over 200 subjects, who are 18 to 50 years old; 69 % female and 31 % male; and 81 % Caucasian, 13 % African, and 6 % other groups. Each video pictures a single facial expression and ends at the apex of that expression while the first frame of every video sequence shows a neutral face. Image sequences from neutral to target display are digitized into 640×480 pixel arrays with either 8-bit gray-scale or 24-bit colour values. The Mind Reading database is an interactive computer-based resource for face emotional expressions, developed by Cohen and his psychologist

team. It consists of 2472 faces, 2472 voices and 2472 stories. Each video pictures the frontal face with a single facial expression of one actor (30 actors in total) of varying age ranges and ethnic origins. All the videos are recorded at 30 frames per second, last between 5 to 8 seconds, and the resolution is 320×240. 360 image sequences of 100 subjects from the Cohn-Kanade database and 120 image sequences of 20 subjects from the Mind Reading database are selected randomly for our work, which constitute a total of 480 image sequences of 120 subjects.



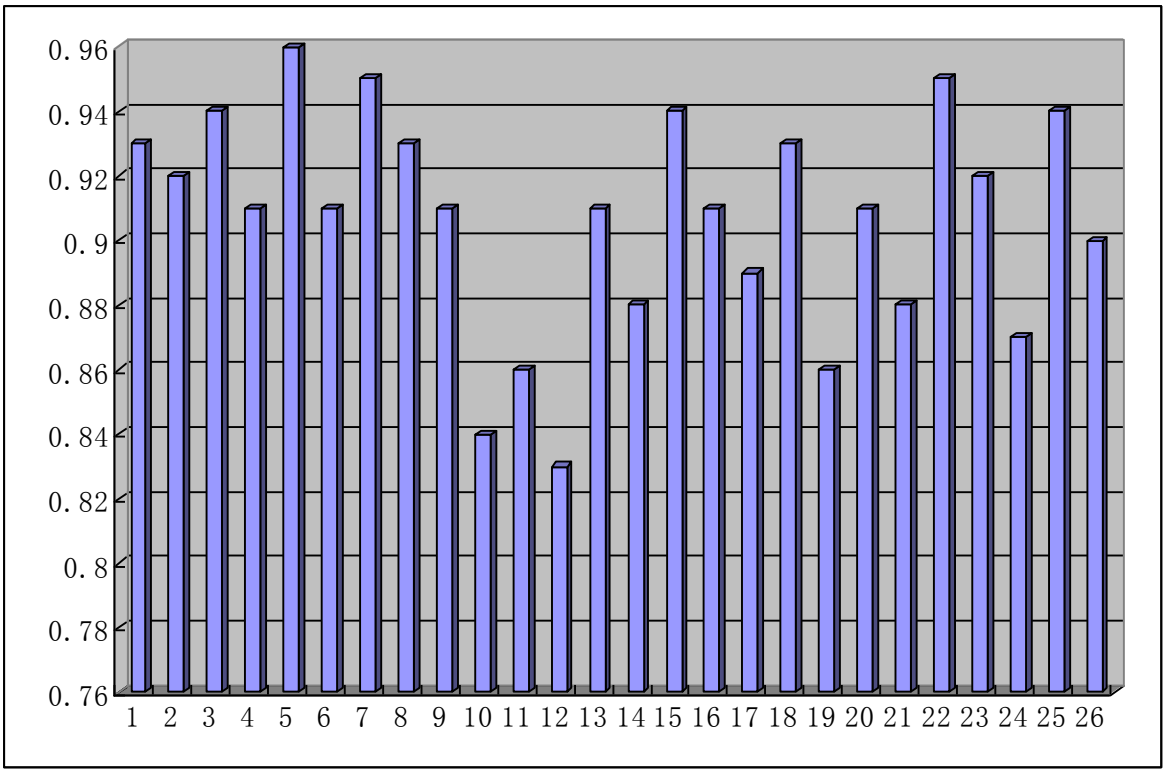
(a)



(b)

Figure 5.3 Test rates from Adaboost algorithms, a) 40 boosting iterations, b) 200 boosting iterations

Table 5.2 Detection Rate of the 26 Fiducial Points



We divide all the 480 image sequences into training and testing subsets containing 240 sequences each. For the training of the 26 fiducial point detectors, the representative sets of positive and negative samples are selected from the facial region. We use 10 frames from each training sequence and manually label each fiducial point on the face region. We get 10×240 positive samples for each detector. We also choose another five arbitrary points in the same frame and get 5×10×240 negative samples, and in total have 14400 samples for each detector. As discussed in 5.1, the feature length of each one of the detectors is 128, so we have a 128×6 size feature vectors from one frame and a 768×2400

feature matrix for each training detector. The representation of features is highly redundant and computing the complete set is computationally expensive. Thus, we apply the Adaboost algorithms for the dimensionality reduction and detector classification.

RealAdaboost, GentleAdaboost and ModestAdaboost are compared for error checks with 40 and 200 boosting iterations, shown in Figure 5.3. GentleAdaboost returns the best detection rates from the results, and is selected as the classification algorithm for our system.

Figure 5.4 Sample sequences from the test videos for facial point detection

We compare the automatically located fiducial points with the manually located points to evaluate the performance of the detection method. In general, the detecting and tracking methods usually are regarded as SUCCESS if the bias of automatic labeling result to the manual labeling result is less than 30% of the true inter-ocular distance. However, this is unacceptable in the case of facial expression analysis. To follow the subtle changes in the facial feature appearance, we define a SUCCESS case if the bias of a detected point to the true facial point is less than 10% of inter-ocular distance in the test image. The overall detection rates for each point are shown in Table 5.2. And the proposed method achieves 90.69% average detection rate for the fiducial point detection.

We illustrate some representative cases in Figure 5.4. The proposed method is applied on each frame of the input video sequences, and the 26 fiducial points are automatically detected. We can also see that the detection is actually successful even under varying illumination conditions.

5.3.2 Tracking Result

System Criteria

The performances of the proposed method are conducted with the system missing rates and false alarms by comparison between the output and the ground truth. We use recall

and precision as the performance measures to evaluate our proposed method and these are defined as:

$$\begin{aligned} \mathrm{Re}call &= \frac{N_{truth}}{N_{truth} + N_{miss}} \times 100\% \\ \mathrm{Pr}ecision &= \frac{N_{truth}}{N_{truth} + N_{false}} \times 100\% \end{aligned} \tag{5.18}$$

where N_{truth} stands for the number of ground-truth for detections and tracking, N_{miss} stands for the number of missed detections and tracking, and N_{false} stands for the number of false alarms. The sum $N_{truth} + N_{miss}$ is the total number of each fiducial point in the entire video sequence.

System Performances

Two facial expression video datasets are considered for checking the performances of the proposed tracking method: Mind Reading database [99] and RML Emotion database [25]. The RML Emotion database was originally recorded for language and context independent emotional recognition with the six fundamental emotional states: happiness, sadness, anger, disgust, fear and surprise. It includes eight subjects in nearly frontal view (2 Italian, 2 Chinese, 2 Pakistani, 1 Persian, and 1 Canadian) and 520 video sequences in total. Each video pictures a single emotional expression and ends at the apex of that expression while the first frame of every video sequence shows a neutral face. Video

sequences from neutral to target display are digitized into 320×340 pixel arrays with 24-bit color values.

Table 5.3 Tracking Results on Databases

D = Total Detected Points, R = Recall, P = Precision

Points	RML Emotion Database 180 Sequences						Mind Reading Database 240 Sequences					
	D	N_{truth}	N_{miss}	N_{false}	**R (%)**	**P (%)**	**D**	N_{truth}	N_{miss}	N_{false}	**R (%)**	**P (%)**
P1	15764	14238	949	1526	93.75	90.32	41009	38741	4584	2268	89.42	94.47
P 2	15778	14066	1121	1712	92.62	89.15	47178	41328	1997	5850	95.39	87.60
P 3	16099	14394	793	1705	94.78	89.41	45517	40041	3284	5476	92.42	87.97
P 4	15254	13922	1265	1332	91.67	91.27	40903	39066	4259	1837	90.17	95.51
P 5	15212	14652	535	560	96.48	96.32	41904	39499	3826	2405	91.17	94.26
P 6	15200	13855	1332	1345	91.23	91.15	42331	38902	4423	3429	89.79	91.90
P 7	15968	14564	623	1404	95.90	91.21	45548	40920	2405	4628	94.45	89.84
P 8	16064	14209	978	1855	93.56	88.45	40801	39157	4168	1644	90.38	95.97
P 9	15201	13939	1248	1262	91.78	91.70	42053	39547	3778	2506	91.28	94.04
P 10	14506	12825	2362	1681	84.45	88.41	41723	39833	3492	1890	91.94	95.47
P 11	14135	13073	2114	1062	86.08	92.49	39928	38642	4683	1286	89.19	96.78
P 12	15014	12705	2482	2309	83.66	84.62	44620	39083	4242	5537	90.21	87.59
P 13	14805	13957	1230	848	91.90	94.27	43321	41328	1997	1993	95.39	95.40
P 14	15409	13494	1693	1915	88.85	87.57	42613	39664	3661	2949	91.55	93.08
P 15	15224	14411	776	813	94.89	94.66	46641	41007	2318	5634	94.65	87.92
P 16	14904	13889	1298	1015	91.45	93.19	43139	40292	3033	2847	93.00	93.40
P 17	14915	13652	1535	1263	89.89	91.53	41550	38637	4688	2913	89.18	92.99
P 18	15470	14163	1024	1307	93.26	91.55	42513	40336	2989	2177	93.10	94.88
P 19	13924	13179	2008	745	86.78	94.65	41293	38460	4865	2833	89.07	93.14
P 20	15005	13955	1232	1050	91.89	93.00	45418	41358	1967	4060	95.46	91.06
P 21	15025	13399	1788	1626	88.23	89.18	44033	38828	4497	5205	89.62	88.18
P 22	16639	14493	694	2146	95.43	87.10	43146	40721	2604	2425	93.99	94.38
P 23	16136	14324	863	1812	94.32	88.77	42860	41137	2188	1723	94.95	95.98
P 24	14256	13324	1863	932	87.73	93.46	44002	41133	2192	2869	94.94	93.48
P 25	15400	14388	799	1012	94.74	93.43	43536	41063	2262	2473	94.78	94.32
P 26	16024	14391	796	1633	94.76	89.81	42333	38997	4328	3336	90.01	92.12

180 image sequences of six subjects from the RML Emotion database and 240 video sequences of 20 subjects from the Mind Reading database are selected for experiment, which constitute a total of 420 sequences of 26 subjects with six emotions. We list the

final experiment results in Table 5.3 based on the two databases. As explained above, we also consider that a tracked point displaced in any direction less than 10% of inter-ocular distance from the true point is regarded as a SUCCESS point in the final results.

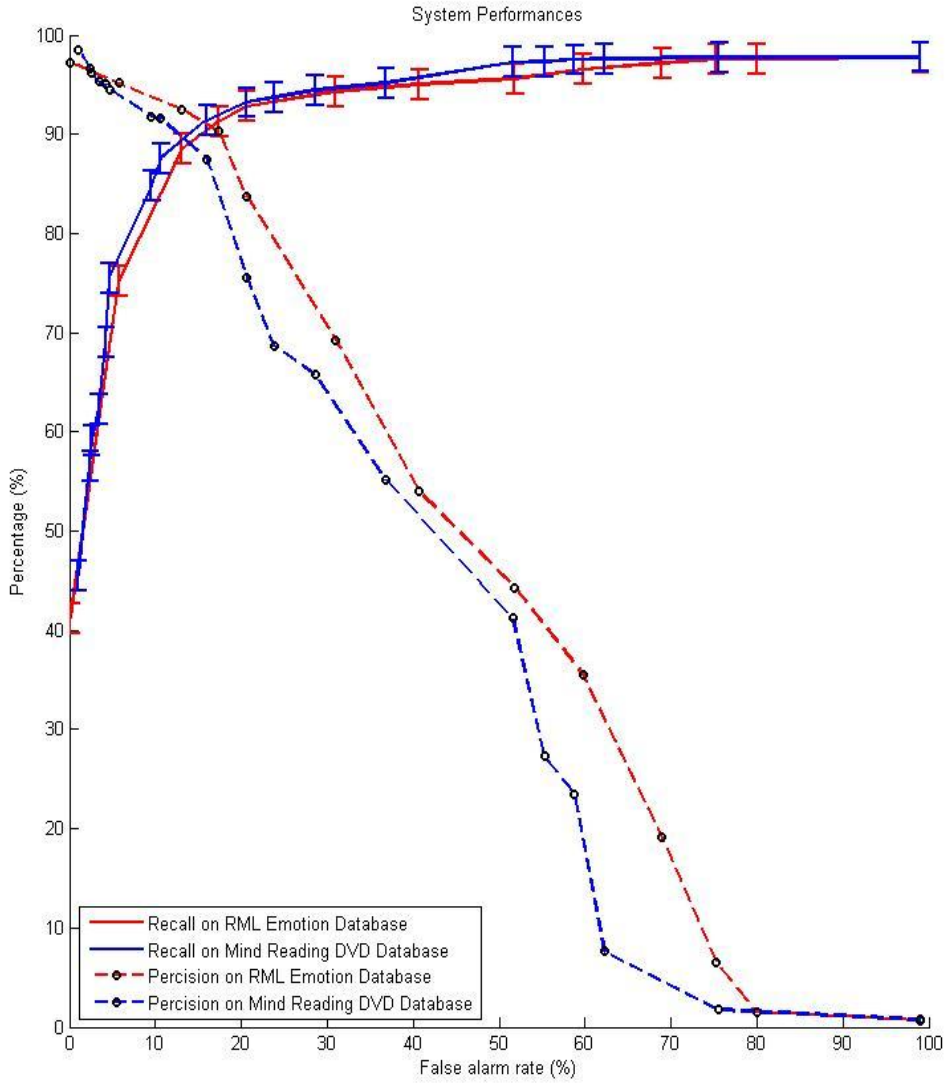


Figure 5.5 Recall and precision against false alarm rate for the test databases

The overall performance of the system in terms of false alarm rate is illustrated in Figure 5.5. From this figure, we can see that the precision is decreasing and recall is increasing with the increment of false alarms. Note, in the graph, a system performance of recall 92.45% and precision 90.93% is achieved simultaneously.

Tracking Result

In this section, we present some representative cases using the proposed method, exploring various practical aspects for fiducial point detection and tracking. Figures 5.6 and 5.7 summarize the experimental results for two different emotional expressions.

The fiducial points are first initialized by the point detectors in the first frame and then tracked by the kernel correlation based multiple DE-MC particle filters. For all figures, the white dots represent the positions of the fiducial points to be detected and tracked, which are all labeled with the associate numbers. In Figure 5.6, the subject exhibits a set of sadness expression from a neutral face at the beginning and ends at the apex of that expression. Figure 5.7 shows an anger expression while talking at the same time. As expected, all the points are tracked reliably for all of the whole sequences. Since the motions of the faces are not intensive and the facial appearances are not heavily changed, the features extracted from consecutive frames are highly correlated and the results achieve a very impressive tracking rate.

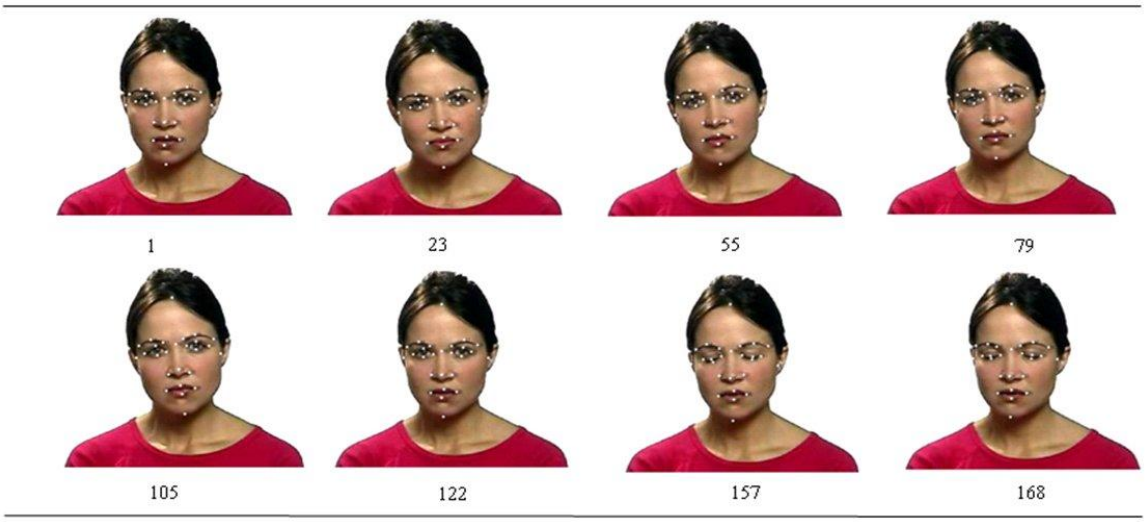


Figure 5.6 Sample sequences for facial expression: Sadness. The frame numbers are marked below.

Figure 5.7 Sample sequences for facial expression: Anger with talking simultaneously. The frame numbers are marked below

We then apply the proposed method to the zoomed case, as shown in Figure 5.8. When the camera zooms, the factors assigned with the colour-based kernel correlation keep changing and descending, as a result of (5.9) and (5.10) which can be seen from frame

63. But the fiducial points can still be tracked with the updating weights $Pi_{j,k}$ from (5.17), as we keep track of the points from the previous frame. It shows the fact that the use of the priors for the multiple filters provides constraints that are sufficient for the reliable tracking of the points at the presence of the facial appearances.

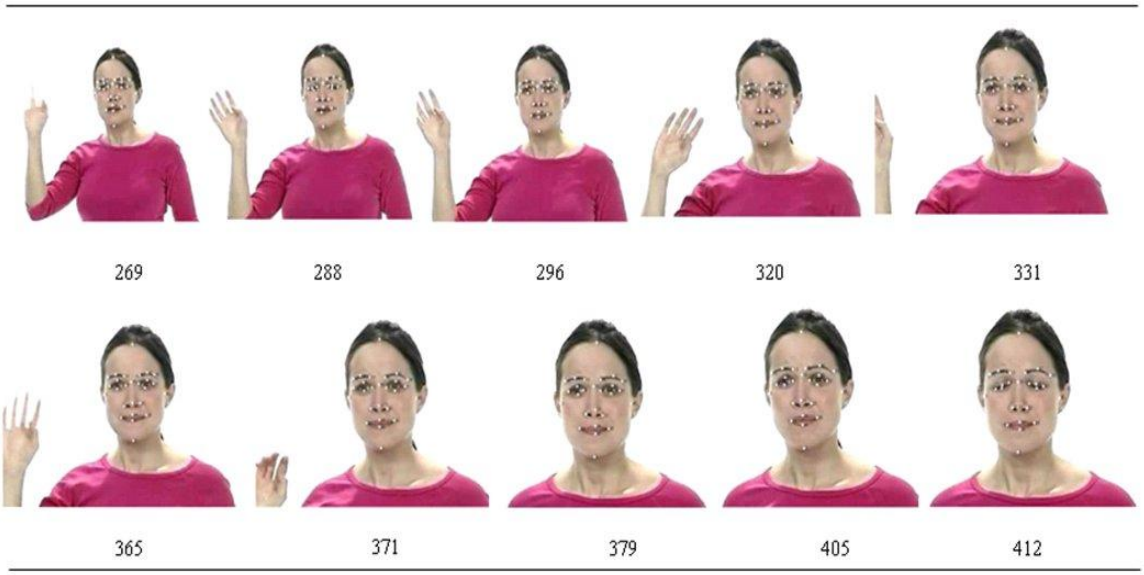


Figure 5.8 Sample sequences for the zoomed case. The frame numbers are marked below

While performing the experiments, we also consider the cases with the head's rotation or being occluded, as shown in Figure 5.9. In Figure 5.9 we can see the points 3, 7, 15, 20 and 23 are lost after frame 78 when a frontal face is rotating to a profile view. So far, the multiple detectors and trackers are based on different configurations of colour intense regions. If both detectors and particle trackers fail for several consecutive frames, the proposed approach will eventually fail.

Figure 5.9 Sample sequences for the head's rotation case. The frame numbers are marked below

Figure 5.10 The improved case for the head's rotation sample sequence

To solve this problem, we execute a conservative way to update the trackers temporally with the response distribution [101] for the next n frames when the missing

points firstly occurred. This step length n can be changed by the user and should not be crucial to the system. If the trackers respond correctly after a few frames, the trackers are able to recover due to the accumulation of probabilities. However, when the step length n continues to grow, due to incorrect responses of the detector, the colour correlation of the observation likelihood drops and the trackers will begin to lose track. After that, "point lost" will be declared. We then stop estimating its motion V_k and discard the motion likelihood term. The trackers will be reinitialized by the point detectors in the following frames. All the 26 points can be detected with a new set of parameters if the facial region appears again in the scene. The improved result is shown in Figure 5.10 that reinitialization executes and all fiducial points are found again after frame 183.

5.3.3 Comparison with the State Of The Art

A comparison of the detection rates and tracking performances achieved by feature point-based methods, with automatic initialization for facial expression recognition, is depicted at Table 5.4. It shows that the proposed method has achieved the best detection rate among state-of-the-art. Moreover, the proposed method has demonstrated its ability to handle pose variations problems and can be used for both image and video based facial expression recognition. Computationally, the proposed method has the advantages of

automatic initialization by using the scale invariant features extraction over the other methods that examine pixels one by one**.**

Table 5.4 Comparison of Feature Point Tracking Methods

Ref.	Sequences	Features	Tracker	DR
The Proposed	480	Scale Invariant Feature	DE-MC PF	90.69%
[90]	300	Gabor	Particle Filter	90.2%
[86]	504	Optical Flow	Optical Flow	87.3%
[92]	743	Color and Edge	NA	86%
[93]	400	Gabor wavelet	NA	87%

Note that the method proposed in [91] achieved a better overall recognition rate (93.72%). However, this method is only tested on perfect manually aligned image sequences and no experiments in fully automatic conditions were reported. In addition, only 13 sequences were experimented on in [91]. Therefore, the result is far from conclusive.

5.4 Chapter Summary

Automatic fiducial point detecting and tracking is a challenging task in facial expression analysis. In this Chapter, we proposed an automatic approach to detect and track fiducial points for varying facial expressions. We first constructed a set of fiducial point detectors with a scale invariant feature. Locating feature points automatically on a single frame makes it possible to eliminate the manual initiation step from tracking algorithms. We

also presented multiple DE-MC particle filters for the fiducial point tracking. This approach combines colour based kernel correlation techniques for the observation likelihood with DE-MC particle filtering distribution for multiple point tracking. Effective tracking performance is achieved by forming the proposal distribution for the particle filter from a mixture of the kernel correlation in the current frame and the dynamic model predicted from the previous step. Different from simply applying the single DE-MC particle filter for multiple point tracking, we adopt the M-component non-parametric mixture model for the multiple DE-MC particle filter posterior distribution over the states of all the target points.

Chapter 6

3D EBS Based Recognition

IN this Chapter, the 3D EBS based emotion recognition method is proposed using a physical face model with D-Isomap for classification. We synthesize emotional facial expressions with the generic mesh model based on the fiducial points obtained in Chapter 5 as the landmarked control points. Function as physics based interpolation transformation, EBS is applied on the 3D mesh model to generate a smooth warp that reflects control point correspondences and to extract the deformation feature of realistic expressions. The corresponding intrinsic geometries of the facial expression can be generated and interpreted as the emotional space. D-Isomap based classification is used to embed the deformable facial expressions into the low dimensional manifold with seven class centers, which span in a face space with 6 emotions and one neutral state.

The main contribution of this work is using the EBS based method for automatic human emotion recognition from video sequences with the active deformation feature extraction depending on the 3D generic face model, which is driven by the key fiducial

points, and thus to make it possible to generate the intrinsic geometries of the emotional space.

6.1 3D Face Modeling

Presently, state-of-the-art of 3D face modeling is by the physically-based modeling paradigm, which will be a key research of emotion recognition for the next-generation HCII [102]. These models can be categorized into two major classes: one is based on the finite element method (FEM) and the other is the deformable mesh (DM) based method. The FEM is a numerical approach to approximate the physics of the 3D face [103]. It implicitly defines interpolation functions between nodes for the physical properties of the face with the solution of the inherent finite differential equations. The FEM equations are complex and their solutions are computationally expensive. The FEM based approach has more utility in applications where a very high accuracy of the tissue movement is required such as surgery simulation, simulation of the tongue and simulation of skin closure. In contrast, the DM based method is better suited to facial expressions recognition due to the fact that the simple formulation is easily implementable and supports topological and geometric flexibility through the local geometric operations. In addition, the DM model is easy and quick to render, easy to modify with existing editors

and easy to represent. The computational cost of a DM model is proportional to the size and complexity of the face for emotion recognition.

Platt and Badler [104] presented a 3D facial model with muscles represented as collections of functional blocks in defined regions of the facial structure. Forces applied to elastic meshes, through muscle arcs, generate realistic facial expressions. Cootes *et al.* [105] proposed an active appearance model for matching statistical models of appearance to images, by employing interactive algorithms. Peyras *et al.* [106] presented a method of fitting active appearance models for unseen faces. The method allows variations in poses and expressions solved by active appearance models. Song *et al.* [107] introduced a generic facial expression analogy technique to transfer facial expressions between arbitrary 3D face models, as well as between 2D face images. Hu *et al.* [27] proposed a work on the non-frontal-view facial expression analysis by generating multi-views from 3D data. Chin *et al*. [108] introduced an emotional intensity-based facial expression modeling process by generating a 3D customized face and facial expressions.

There are several issues with the DM face models that are still not properly addressed. The number of vertices of a DM model should be selected to get the most from the trade-off: being neither too large to complicate the model nor too small to exclude useful microstructure information. It is worth considering carefully how to select the minimal

facial features and the corresponding topological structure of the DM face model with geometric knowledge which fully cooperates with each other. In addition, developing a complete feature set which can be automatically and easily manipulated is extremely difficult [109].

In light of the above problems, this Chapter presents an EBS based method for automatic face modeling. The goal of this step is to generate facial expressions using a physically-based mesh modeling approach. This can be done according to the input video sequence from the deformable feature perspective, and executing with the control points in an acceptable time, for emotion recognition applications. Merits of this proposed approach are:

a) we propose a physically-based model of human face with fiducial points for driving the deformation of the face according to the muscle movement parameterization. The face can be modeled as an elastic body that is deformed under a tension force field. Muscles are interpreted as forces deforming the polygonal mesh of the face. The factors affecting the deformation are tension of the muscles, elasticity of the skin and zone of influence. Higher-level parameterizations that are easier to use for emotional expressions can be defined in terms of low-level parameters.

b) We extend the DM face model by a set of well-designed polygons with an EBS

structure which can be efficiently modified to establish the facial expression model. A 3D face is decomposed into area or volume elements, each endowed with physical parameters embedded in the EBS model according to the surface curvature. The deformable element relationships are computed by integrating the piecewise components over the entire face.

c) The control points are predefined by the landmarked fiducial points. The number of control points is small and the control points can be identified robustly and automatically. Once the control points are adjusted, the emotional face model can be established using the transform function of the EBS and can be extended to obtain expression parameters for final recognition.

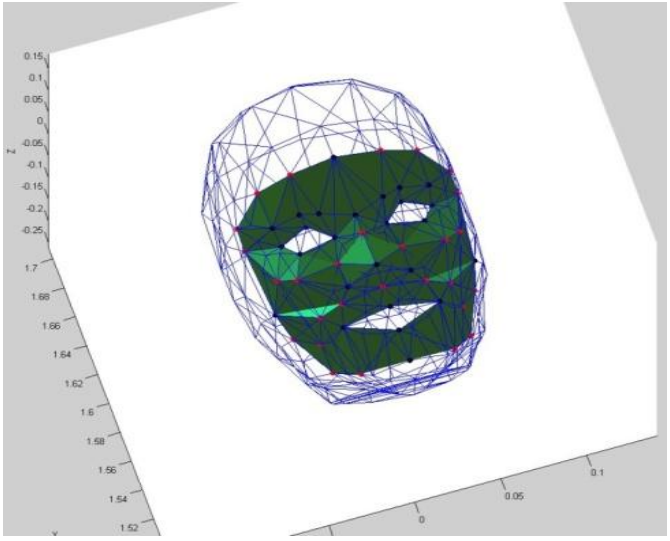


Figure 6.1 The proposed 3D mesh model with 26 fiducial points and 28 characteristic points

Our generic face model is actually a mesh wireframe model consisting of characteristic feature points and deformable polygons with the EBS structure. We can deform the wireframe model to best fit a human face without or with any expressions. The 3D affine transformation realizes the facial expressions by imitating the facial muscular actions. It formulates the deforming rules according to the FACS coding system by the 26 fiducial points as the control points. Figure 6.1 shows the proposed model based on this standardized coding system.

In practical applications, not all feature points in the model can be easily detected from the input sequences [110], so we use 54 characteristic feature points for facial expression parameterization. Characteristic feature points include: a) the 26 significant fiducial point-based control points, and b) 28 dependent points which are determined by the fiducial points. We also assume that the physical property of the EBS structure is the same within the facial region. The EBS deformation analysis is presented in the following section.

6.2 EBS Parameterization

EBS is applied for generating different facial expressions with a generic face model from a neutral face. By varying the position of control points, EBS mathematically describes

the equilibrium displacement of the facial expressions subjected to muscular forces using a Navier partial differential equation (PDE). The deformable face model equations can be expressed in 3D vector form with the interpolation spline relating the set of corresponding control points. Let $X_i = [X_{i1} \quad X_{i2} \quad X_{i3}]^T (i = 1,2,...,N)$ denote a set of N control points in the 3D face model of neutral face, $Y_i = [Y_{i1} \quad Y_{i2} \quad Y_{i3}]^T$ be the corresponding control points with expressions, and $L(\bar{x})$ be the displacement of all points within the face model from the original position (neutral face). The displacement between the feature points sets are $L_i = Y_i - X_i$. To find an appropriate physical property for an expressional model, muscular forces are assumed to be distributed on the homogeneous isotropic elastic body of the face model to obtain smooth deformation. Solving the PDEs, we can form the splines as linear combination of translated versions of the solution. The coefficients of the spline are determined from points X_i, the displacements of the control points sets L_i. The spline relaxes to an affine transformation as the distance from the point approaches infinity.

We have the Navier equilibrium PDE as:

$$\alpha \nabla^2 L(\bar{x}) + (\alpha + \beta) \nabla [\nabla \bullet L(\bar{x})] + F(\bar{x}) = 0 \tag{6.1}$$

where $F(\bar{x})$ is the muscular force field being applied on the face, α and β are the Lam é coefficients to describe the physical properties of the face, ∇^2 and ∇ denote the Laplacian

and gradient operation, respectively, and $\nabla \bullet L(\bar{x})$ is the divergence of $L(\bar{x})$. To obtain smooth deformation of the face model, the muscular force field $F(\bar{x})$ can be formulated as external forces that:

$$F(\bar{x}) = Wd(\bar{x}) \tag{6.2}$$

where $d(\bar{x}) = |\bar{x}| = \left|x_1^2 + x_2^2 + x_3^2\right|^{1/2}$, and $W = [w_1, w_2, w_3]^T$ is the strength of the force field.

So, the PDE's solutions of (6.1) can be computed as:

$$L(\bar{x}) = E(\bar{x})W \tag{6.3}$$

$$\text{and } E(\bar{x}) = [\lambda d(\bar{x})^2 I - 3\bar{x}\bar{x}^T]d(\bar{x}) \tag{6.4}$$

where $\lambda = (11\alpha + 5\beta)/(\alpha + \beta)$ is the Poisson's ratio, I is a 3×3 identity matrix, and $\bar{x}\bar{x}^T$ is an outer product. Using linear combination of the PDE's solution in (6.3), we can calculate the EBS that represents all the displacements of an expressional face with the translated version:

$$L_{EBS}(\bar{x}) = \sum_{i=1}^{N} E(\bar{x} - X_i)W_i + A\bar{x} + \bar{b} \tag{6.5}$$

$A\bar{x} + \bar{b}$ is the affine portion of the EBS, $A = [\bar{a}_1, \bar{a}_2, \bar{a}_3]$is a 3×3 matrix. The summation in (6.5) can be expressed in the matrix-vector form as

$$\begin{aligned} E_{EBS\_M} &= H^{-1} L_{EBS\_M} \\ &= \begin{bmatrix} W_1^T & W_2^T & \cdots & W_N^T & \bar{a}_1^T & \bar{a}_2^T & \bar{a}_3^T & \bar{b}^T \end{bmatrix}^T \end{aligned} \tag{6.6}$$

where E_{EBS_M} is a $(3N + 12) \times 1$vector with all the EBS coefficients.

L_{EBS_M} is defined as a vector with all the displacements and augmented by zeros that $L_{EBS\_M} = \begin{bmatrix} L_1^T & L_2^T & \cdots & L_N^T & O_1^T \end{bmatrix}^T$, O_1 is a column vector of 12 zeros, and H is the transfers function that:

$$H = \begin{bmatrix} K & C \\ C^T & O_2 \end{bmatrix}_{(3N+12)\times(3N+12)} \tag{6.7}$$

where $K = \begin{bmatrix} E_{(\Delta_{11})} & E_{(\Delta_{12})} & \cdots & E_{(\Delta_{1N})} \\ E_{(\Delta_{21})} & E_{(\Delta_{22})} & \cdots & E_{(\Delta_{2N})} \\ \vdots & \vdots & \ddots & \vdots \\ E_{(\Delta_{N1})} & E_{(\Delta_{N2})} & \cdots & E_{(\Delta_{NN})} \end{bmatrix}_{3N\times 3N}$

$$C = \begin{bmatrix} X_{11}I & X_{12}I & X_{13}I & I \\ X_{21}I & X_{22}I & X_{23}I & I \\ \vdots & \vdots & \vdots & \vdots \\ X_{N1}I & X_{N2}I & X_{N3}I & I \end{bmatrix}_{3N\times 12}$$

O_2 is a 12×12 matrix of zeros, and $\Delta_{ij} = X_i - X_j$.

We have 26 control points, so $N = 26$ in our system. The control point positions X_i, Y_i and the displacements of the control point sets L_i are obtained from the detection and tracking steps in Chapters 4 and 5. We solve (6.6) from the requirements that the spline displacements equal the control point displacements with a constant Poisson's ratio λ all over the face region. The flatness constraints which are expressed in terms of second or higher order (e.g. X_i^2, X_j^2 or $X_i X_j$) are set to zero that enforce the conservation of linear and angular momenta for an equilibrium solution. These constraints cause the force field

to be balanced so that the EBS face model is stationary. The spline coefficient E_{EBS_M}, the spline basis function H and the control point locations X_i can be used in (6.5) to compute the value of the spline for the 28 nonsignificant points.

The muscular force field $F(\bar{x})$ in (6.2) is calculated from the solution of EBS that:

$$F(\bar{x}) = \sum_{i=1}^{N} [w_{i1} \quad w_{i2} \quad w_{i3}]^T d(\bar{x} - X_i) \tag{6.8}$$

By the principle of superposition for an elastic body, the external forces must be minimized according to the roughness measurement constraints [111]. This ensures that the forces are optimally smooth and sufficient to deform the elastic material so that the EBS equals the given displacements at the control point locations. By varying the values of Poisson's ratio in (6.3), we can calculate each corresponding muscular force field respectively. To find the minimum muscular force field, $\left|F(\bar{x})\right|_{\min}$, we obtain the appropriate physical property λ' and the associate EBS coefficients E_{EBS_M}. We then construct the deformable visual feature f^d for classification with λ' and E_{EBS_M}. The deformation feature extraction step is summarized as follows.

1. Initialize the control point positions X_i in the 3D face model for neutral face according to the detection results for the 26 fiducial points
2. Set the Poisson's ratio λ for facial region as 0.01

3. Update the corresponding control point positions Y_i in the expressional face model subject to the tracking results
4. Calculate the displacements of the control point sets L_i in the facial region
5. Solving the EBS in (6.5) and obtain associate spline coefficient E_{EBS_M}
6. Compute the position of nonsignificant points in the facial region based on the EBS's solution in the previous step
7. Calculate the muscular force field $F(\bar{x})$ in (6.2) from the solution of the EBS
8. Sweep the Poisson's ratio from 0.02, 0.03, …, to 0.5 and repeat steps 5, 6 and 7 to obtain the new muscular force fields
9. Find the minimum muscular force field $\left|F(\bar{x})\right|_{\min}$, fix the Poisson's ratio λ' and the EBS coefficients E_{EBS_M}
10. Construct the deformable visual feature f^d for classification with λ' and E_{EBS_M}

6.3 D-Isomap Based Classifier

Once the deformable facial features have been established with the EBS, we use D-Isomap based method for emotion classification. Isomap was proposed by Tenenbaum [112], and is one of the most popular manifold learning techniques for promising nonlinear dimensionality reduction. It attempted to learn complex embedding manifolds

using local geometric metrics within a single global coordinate system. The Isomap algorithm uses geodesic distances between points instead of simply taking Euclidean distances, thus encoding the manifold structure of the input space into distances. The geodesic distances are computed by constructing a sparse graph in which each node is connected only to its closest neighbors. The geodesic distance between each pair of nodes is taken to be the length of the shortest path in the graph that connects them. These approximated geodesic distances are then used as input to classical multidimensional scaling (MDS).

Yang proposed a face recognition method based on Extended Isomap [113]. In his work, an extended Isomap method for face recognition that utilized Fisher Linear Discriminant (FLD) was introduced. The main difference between this method and the original Isomap method is that after a geodesic distance is obtained, the extended Isomap algorithm uses FLD to achieve the low dimensional embedding while the original Isomap algorithm uses MDS to achieve the embedding. X. Geng [114] proposed an improved version of Isomap to guide the procedure of nonlinear dimensionality reduction. The neighborhood graph of the input data is constructed according to a certain kind of dissimilarity between data points, which is specially designed to integrate the class information.

The Isomap algorithm generally has three steps: *Construct neighborhood graph*, *Compute shortest paths*, and *Construct D-dimensional embedding*. Classical MDS is applied to the matrix of graph distances to obtain a low-dimensional embedding of the data. Thus the prime difference between MDS and Isomap is the use of geodesic distances in Isomap. However, since the original prototype Isomap does not discriminate data acquired from different classes, when concerned with multi-class data, several isolated sub-graphs will result in undesirable embedding. On the other hand, the Extended Isomap [113] can only be used when handling the problem in which the number of the classes is less than three. When the number of classes becomes larger, the classes may construct their own spatially intrinsic structure. The Extended Isomap and improved version cannot recover the classes' intrinsic structures of the high-dimensional data.

In order to cope with such problems, in this section, we adopt a discriminative Isomap [115] based method for emotion classification. The discriminative information of facial feature is considered so that it can reflect successfully the discriminative structures of the emotional space on the manifold. The discriminative Isomap has the capability of discovering nonlinear degrees of freedom and finding the globally optimal solution guaranteed to converge for each manifold [116, 117].

The EBS feature f^d for each emotional face model can be seen as one point in a high dimensional space. As the result from the last step, there are 54 characteristic feature points in the 3D face model, and every feature has 175 dimensions. Given the variations of facial configurations during emotional expressions, these points can be embedded in a lower dimensional space. We define high dimensional facial EBS feature set $F_{\text{original}} = \{f_i\} \subset R^{n \times M}$ as the input data, $i = 1,...,n$ is the input sample number, $M = 175$ is the dimensionality of the original data. Let $Y_{Isomap} = \{y_i\} \subset R^{n \times m}$ denote the embedding space of F_{original} into a low dimensional manifold with dimension of m, which preserves the manifold's estimated intrinsic geometry [118].

We compute Euclidean distance d_{ij} between any pairwise points in input space F_{original} with discriminative weight factor $\Psi (0 < \Psi < 1)$ that:

$$d_{\text{ij}}(f_i, f_j) = \begin{cases} \sqrt{\|f_i - f_j\|^2} & if \quad L(f_i) \neq L(f_j) \\ \Psi \sqrt{\|f_i - f_j\|^2} & if \quad L(f_i) = L(f_j) \end{cases} \tag{6.9}$$

where $L(f_i)$ denotes the class label which the input data f_i belongs to. For pairwise points with the same class label, the Euclidean distance is shortened by weight factor Ψ. The parameters of the compacting and expanding are needed to be empirically defined for discriminative matrix [119]. We construct a neighborhood graph G according to the distance between the points. A point is a neighbor of any other point if it lies within a

fixed radius or is one of the closest points to it. The neighboring graph G is constructed by connecting point pairs with edges equal to the distance between the points. The distances between all point pairs are computed based on the chosen distance metric. We then calculate geodesic distance matrix D_{G} between all pairwise points by computing the shortest paths in the neighborhood graph G :

$$D_{\mathrm{G}}(f_i, f_j) = \min_{p \subset P_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{\mathrm{ij}}(f_i, f_j), \tau_{ij})$$

$$\text{and} \quad \tau_{ij} = \begin{cases} 1 & if \quad f_i, f_j \in p \\ 0 & if \quad f_i, f_j \notin p \end{cases} \tag{6.10}$$

where P_{ij} denote the set of all paths connecting f_i and f_j, $p \subset P_{ij}$ is a path that is acquired by adding up a sequence of edges between two neighboring points. To make sure the matrix is symmetric the geodesic distance matrix D_{V} between all points is set as:

$$D_{\mathrm{G}} = \min(D_{\mathrm{G}}, D_G^T) \tag{6.11}$$

To convert the distance matrix to inner products of matrix, we need to construct a translation map D_{trans} that:

$$D_{\mathrm{trans}} = \frac{-HD_V^2 H}{2} \tag{6.12}$$

where H is the cantering matrix, given by $H = I - \frac{1}{N} ee^T$, and $e = [1,..,1]^T \in R^M$. Compute the largest eigenvalue and the top m eigenvectors of D_{trans}, we obtain the

eigenvector matrix $E \in R^{n \times m}$ and the eigenvalue matrix $M \in R^{m \times m}$. The embedding matrix U_{Isomap} in low dimensional space can be calculated that:

$$U_{Isomap} = M^{1/2} E^T \tag{6.13}$$

The discriminative Isomap can discover the discriminative structure on the manifold, and provides a simple way to obtain the low dimensional embedding as well [119]. A labeled class center $\{u_i\}_l$ for an emotional space is calculated depending on the result from discriminative Isomap:

$$\{u_i\}_l = \frac{1}{k} \sum_{i=1}^{k} ([V_i]_l U_{Isomap} - D_{\mathrm{G}}(f_i, f_j)) \tag{6.14}$$

where $l = 7$ is the emotional space for labeling, k is the sample number in one emotional space. $[V_i]$ represents the ith-element of the eigenvector matrix.

We use Nearest Class Center (NCC) algorithm to determine the emotion class of a test data. Compute the class centers for the test data that

$$u' = \frac{1}{n_i} \sum_{u_i \in C_l} u_i \tag{6.15}$$

where i is number of data in class u_i and u' is the center coordinates of class C_l. Using the nearest class center u_{Cl} by Euclidean distance, we can obtain the class label Cl for the test data.

$$Cl = \arg\min_{u' \in C_l} (d_{ij}(u', \{u_i\}_l)) \tag{6.16}$$

6.4 Experiment and Results

To evaluate the performance of our proposed emotion recognition method, we implement experiments on a Pentium IV 2.8 GHz PC with 3.25 GB memory, under the Windows XP operating system, and all are coded in MATLAB. Two facial expression video datasets are used for experiment: RML Emotion database and Mind Reading DVD database.

The positions of 26 fiducial points are obtained from detecting and tracking step and for calculating the positions of 28 dependent points. These positions are 2D data in the video sequences and cannot be applied with the 3D EBS analysis directly. All the fiducial points need to be aligned to our 3D model first in this work. We use a flexible generic face modeling (FGFM) algorithm [120] for fitting each face image to the 3D mesh model. To remove the individual differences in the facial expressions, each face shape from the video sequences is normalized to the same scale. We use the length (distance between fiducial point 1' and 2') and width (distance between fiducial point 3' and 4') of neutral face for scale normalization. The 26 control points on the 3D face model are initially estimated by the fiducial points using the back projection technique with the set of predefined unified depth values. The original dependent points are also predefined in the model. A classified FGPFM ratio features is selected with a minimal Euclidean

distance between the estimated and the codebook-like ratios database. The depth values of control points and curvature parameters are obtained for reconstructing the EBS face model from the selected ratio features classifier.

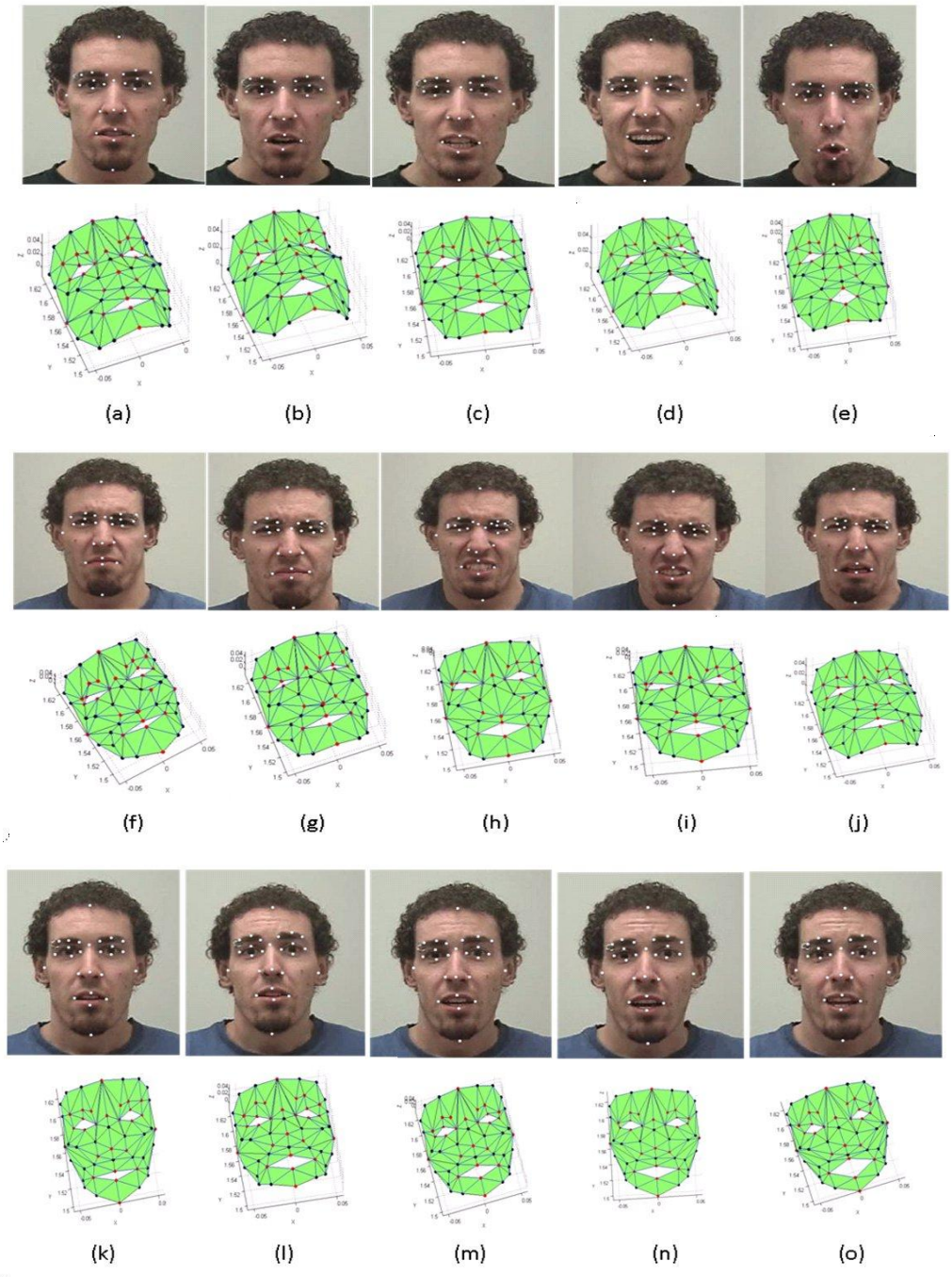


Figure 6.2 Emotional EBS model construction, (a-e) anger faces (f-j) disgust faces (k-o) fear faces

Figure 6.2 shows some representative sample results for emotional model construction with our proposed method. Our objective here is to find the positions of dependent points after emotional facial deformation under the availability of the fiducial point's position. The basic six emotions are analyzed in this experiment, i.e. *happiness, sadness, anger, fear, surprise,* and *disgust*. The best-fit mesh model of a given face is estimated from the first input frame with neutral emotion. Based on the known tracking information, the positions of all characteristic feature points are calculated and the EBS model is reconstructed with any particular expressions. From the experimental results we can see that our method provides good construction results following the variations of the control points.

We provide more experiment results in Figure 6.3 to verify the consistency of the proposed method. Figure 6.3 presents the results of emotional face model for different peoples. The Poisson's ratio is assumed to be constant for the whole facial region and determined under the condition of minimum muscular force field generation. Figure 6.3 (a-d) shows the results when Poisson's ratio is set to 0.27, 0.41, 0.31 and 0.25 respectively. Subjectively, the proposed method provides a good face model under different peoples and facial expressions.

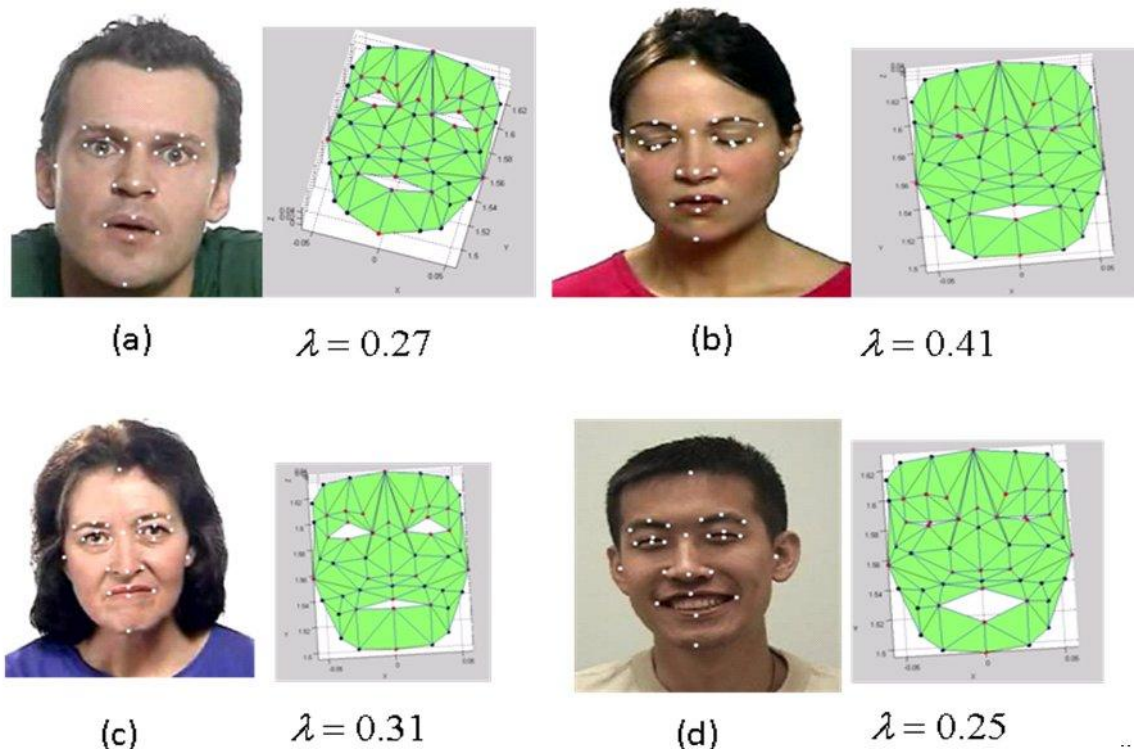


Figure 6.3 EBS face model constructions with different Poisson's ratio (a) a male anger face (b) a female sadness face (c) a female anger face (d) a male happiness face

Using EBS transform can interpolate the positions of characteristic feature points such that the 3D face model of an expressive expression can be generated from the input video frame. Based on the arrangement of facial muscle fiber, our EBS model calculates elastic characteristics for each emotional face by modeling the facial muscle fiber as elastic body. The affine or rigid elastic body coordinate transformation is fitted to the displacements of the facial expression with the continuity condition. The spline obtained by this method is mathematically identical to computed coefficients from the original displacements of the control points directly. Moreover, the resulting spline is added to the

initial mesh of the elastic body transformation to give the overall coordinate transformation. Simulation results show that the face model generated by our method demonstrates good performance under the availability of control points' positions.

D-Isomap Based Classification

In this section, 280 video sequences of 8 subjects from RML Emotion database and 420 video sequences of 12 subjects from Mind Reading DVD database are randomly selected for D-Isomap based classifier evaluation, which constitutes a total of 700 sequences of 20 subjects with six emotions and neutral faces.

The facial EBS feature for every frame is extracted from the last step and to construct a 175 dimensional vector. It is too large to manipulate directly. We use D-Isomap algorithm for dimensionality reduction. Since each feature vector can be seen as one point in a 175 dimensional space, the D-Isomap is utilized to finds the embedding manifold in a low-dimensional space to represent the original data. These representations should cover most of the variances of the observation based on the continuous variations of facial configurations during emotional expressions. In our system, the low-dimensional space structures are extracted to facilitate the manifold's estimated intrinsic geometry due to the D-Isomap's capability of nonlinear analysis and the convergence of globally optimal solution.

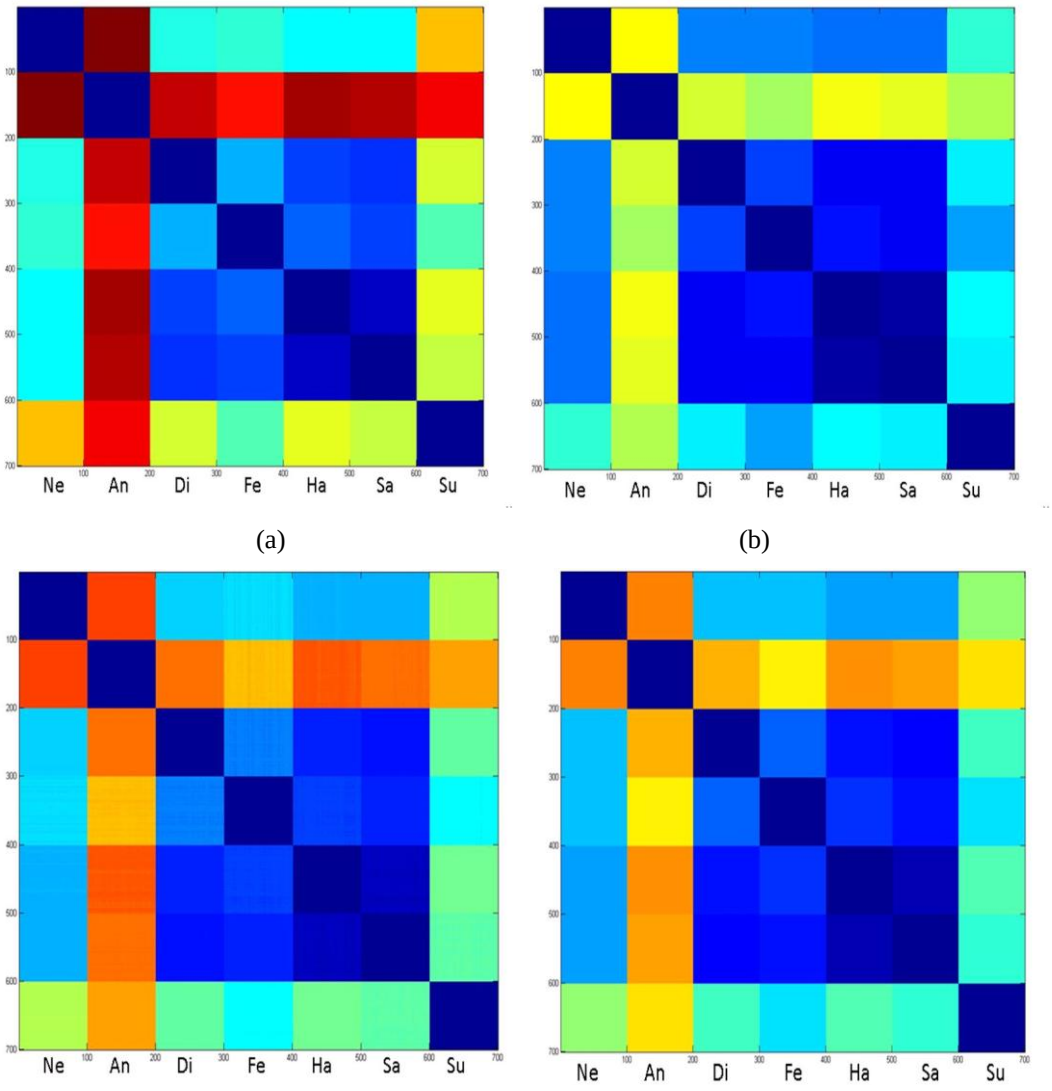


Figure 6.4 Distance matrix graph with different weight factor, higher values are shown in red, lower values in blue, that (a) $\Psi = 0.1$ (b) $\Psi = 0.25$ (c) $\Psi = 0.5$ (d) $\Psi = 0.75$

The geodesic distance graph from (6.12) is used for D-Isomap based embedding. Figure 6.4 shows examples of distance matrix with discriminative weight factor Ψ for 7 emotional expressions of randomly selected subjects. The distance graph reflects the intrinsic similarity of the original expression data and consequently is considered for determining true embedding in our system. From Figure 6.4 we can see, by applying the

weight factor, the points from the same cluster can be projected closer in low dimensional space, thus the distance is compacted. On the other hand, the distance between different clusters could be expanded by the weight factor.

Increasing the dimension of the embedding space, we can calculate the residual variance for the original data. The true dimension of data can be found out by considering the trend of decrease in the residual value. The embedding results using Isomap and adopted D-Isomap with different clusters k are presented in Figure 6.5. Figure 6.5 (a)–(f) shows the results when cluster k is set to be 7, 12, and 20, respectively. From the results we see that our proposed method achieves an average of 10% improvement comparing with the original Isomap. The best performance can be obtained when cluster k is 12 and the dimension of embedded space is reduced to 20, which covers more than 95% variances of the observation from the input data. Therefore, these 20 dimensional components are used here to represent facial expressions in the input videos.

We also provide expressional configurations to show emotional apparent variation in Figure 6.6 with the sample numbers of 700 and 7000 respectively. By applying NCC algorithm to the embedding results from the D-Isomap using (6.15), we can determine the emotion class for a test video. We label the emotion class centers on the embedded feature space, which are shown in Figure 6.6.

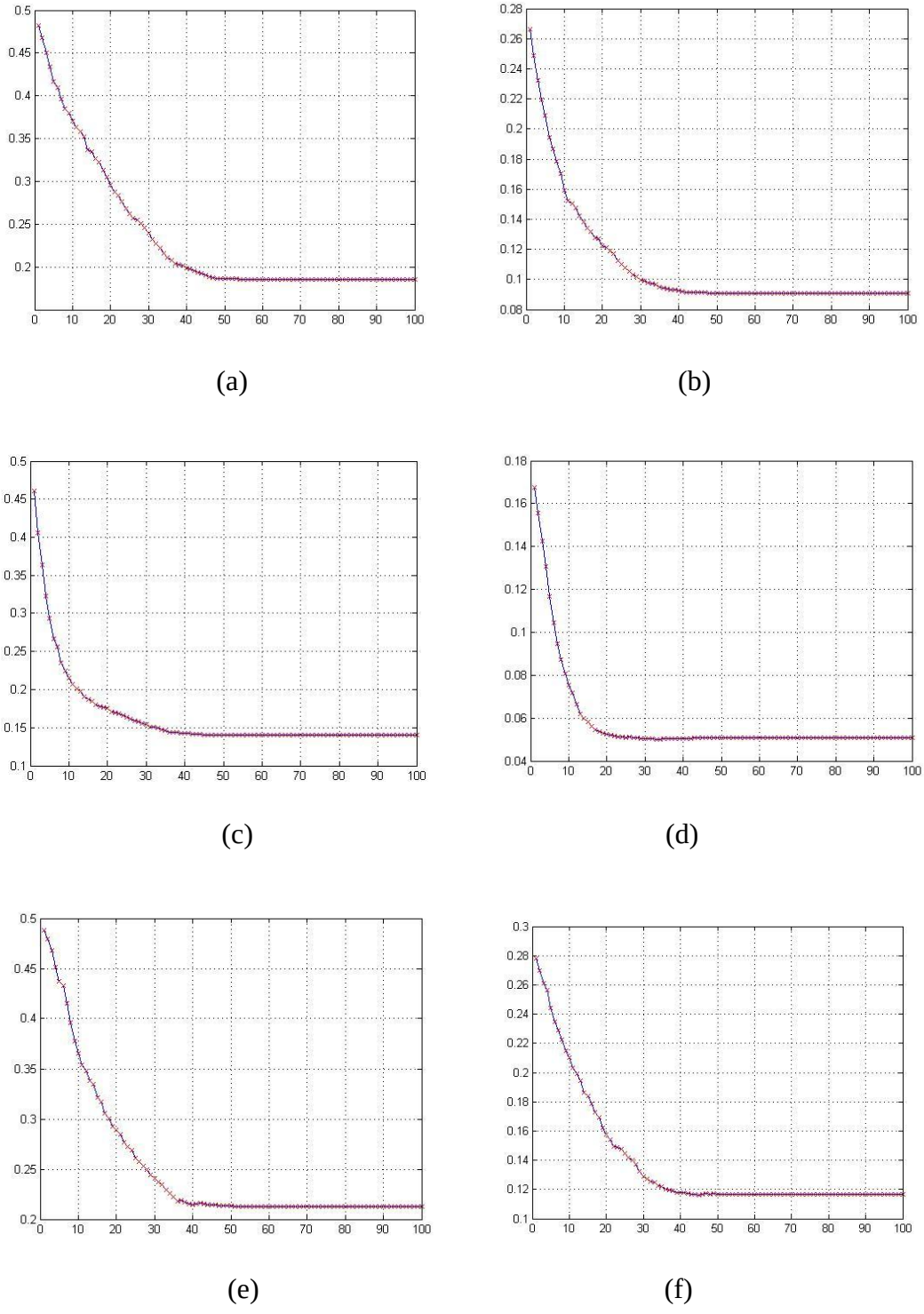
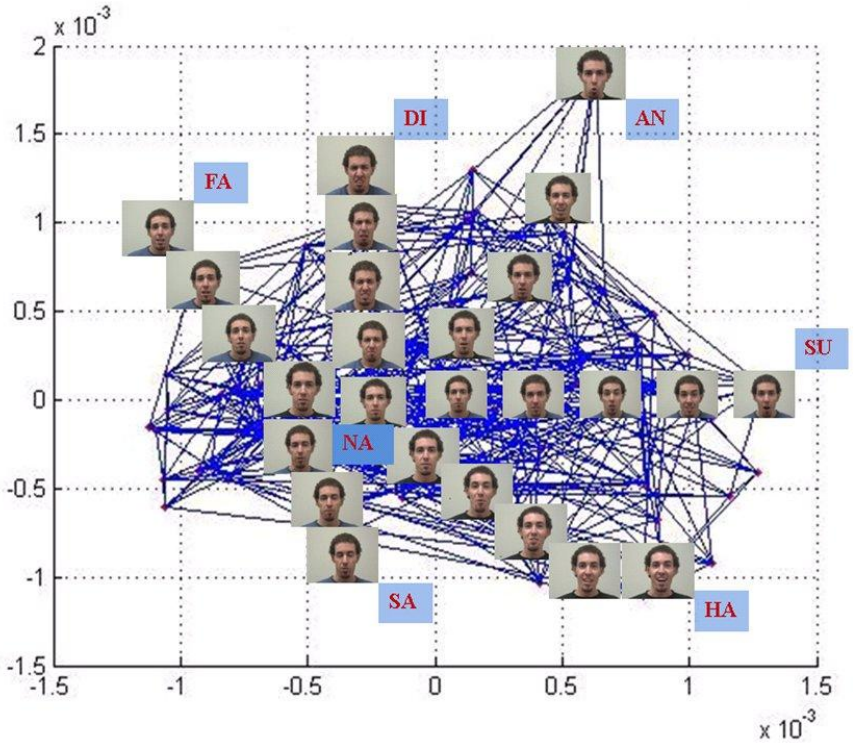
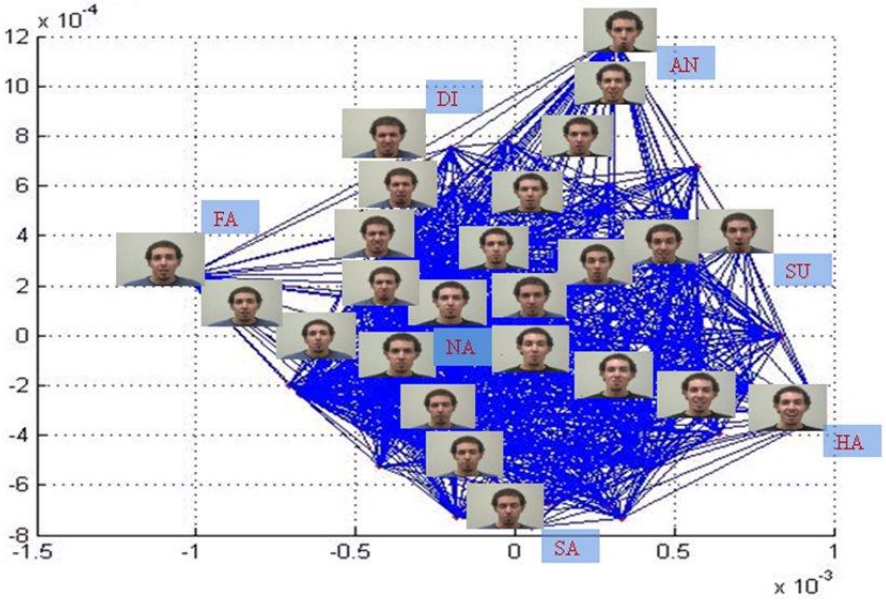


Figure 6.5 Dimensionality reduction using Isomap and D-Isomap, (a) (c) (e) shows the results using Isomap with k is 7, 12, and 20 respectively, (b) (d) (f) shows the results using discriminative Isomap



(a)



(b)

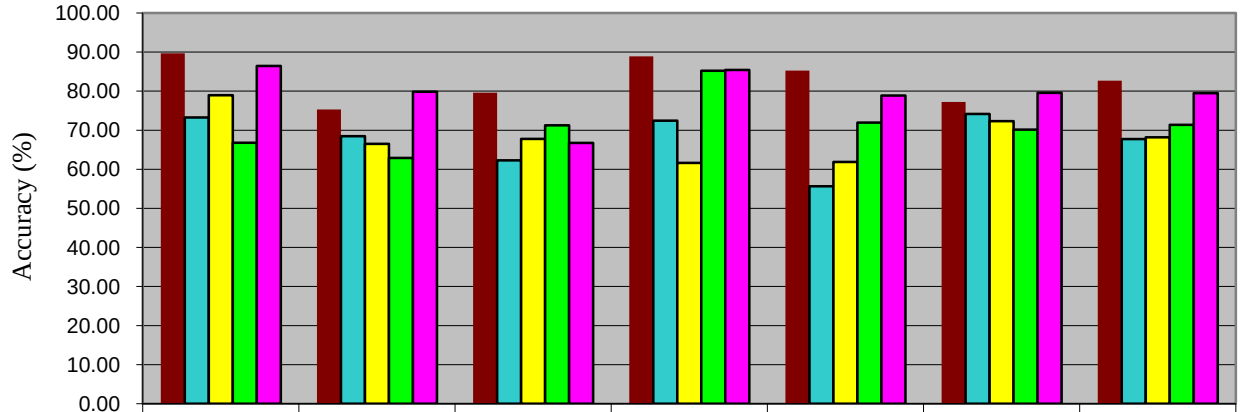
Figure 6.6 Labeled class centers in a 2D space based on the embedding results (a) shows the results using 700 samples (b) shows the results using 7000 samples

To evaluate the performance of our proposed method, we randomly divide these 700 sequences into 2 subsets containing 350 sequences each for training and testing. Training and testing procedure are repeated 5 times. Each time one of the 2 subsets was used as a test set and the other was used as a training set. The recognition accuracy is calculated as the ratio of the number of correctly classified samples and the total number of samples in the data set.

Table 6.1 Emotion recognition confusion matrix

	Detected						
Desired	Neutral	Angry	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**82.93**	0.97	2.71	2.38	4.04	5.32	1.65
Angry	1.03	**89.68**	1.66	2.77	1.23	1.52	2.11
Disgust	3.56	1.55	**75.32**	7.74	1.92	4.95	4.96
Fear	4.21	0.99	4.22	**79.61**	1.14	5.37	4.46
Happiness	2.79	1.01	2.36	1.89	**88.92**	1.37	1.66
Sadness	2.34	1.05	3.43	3.12	1.08	**85.27**	3.71
Surprise	4.69	1.06	3.76	7.66	1.37	4.21	**77.25**

We list the confusion matrix for emotion recognition with numbers representing correct percentages in Table 6.1. From the results we can see that features representing different expressions exhibit great diversity since the distances between different emotions are relatively high. On the other hand, the same expressions collected from different subjects are very similar due to the short distances within the same class. The overall recognition rate using proposed method is about 82.7%.



	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Overall
DIsomap	89.68	75.32	79.61	88.92	85.27	77.25	82.68
PCA	73.28	68.45	62.27	72.46	55.67	74.15	67.71
GMM	78.94	66.49	67.78	61.64	61.88	72.33	68.18
NN	66.80	62.90	71.27	85.21	71.96	70.16	71.38
FLDA	86.46	79.85	66.73	85.44	78.86	79.57	79.49

Figure 6.7 Recognition results of different classifiers

We conduct extensive experiments using different classification schemes, i.e., PCA, Gaussian Mixture Model (GMM), NN and Fisher's Linear Discriminant Analysis (FLDA). In PCA, the dimensionality is reduced by eliminating less significant features with smaller eigenvalues in the transformed domain. K-nearest neighbors (K-NN) is then used for the classification. Compared with PCA, D-Isomap can generate a smaller number of resulting clusters in the embedded space. So the intrinsic structures of the original data are recovered with lower computing costs. The GMM classifier is implemented in a modular architecture. A separate GMM is trained for each individual class. The parameters including the weights, mean and standard deviation of each

component are estimated by the Expectation Maximization (EM) algorithm. In out experiments, we try a range of *k* values, so that the distribution of the data can be modeled as the sum of *k* Gaussian functions. In NN classification, a three-layer feed-forward NN is investigated. The number of input layer neurons is equal to the dimension of the input feature set, while the output neurons correspond to the six emotion classes. Back-propagation algorithm is used to train the network. A new input is labeled the class that produces maximum output value. The applied FLDA classifier has six outputs corresponding to the six emotions. An input signal is labeled the class that gives the maximum output value. The experimental results for the performance comparison with the same dataset are drawn in Figure 6.7. From the figure we can see our proposed method achieves the best results for the final emotion recognition.

We also compare our proposed D-Isomap classifier with other two Isomap classifiers, i.e. Original Isomap and Extended Isomap. Since K-NN clustering computes Euclidean distances between all pairs of points, it is chosen for evaluating the classifiers. The parameters are empirically determined to achieve the lowest error rate by each method. In the original Isomap, the value of cluster *k* is 12. In the extended Isomap, the value of nearest neighbor *kw* used in within-class matrix is 5. The test results are summarized in Table 6.2.

Table 6.2 Comparison between three Isomap methods

	Parameters	Reduced Dimension	Standard Deviation	Recognition Results
Original Isomap	$k = 12$	40	18.3%	62.8%
Extended Isomap	$kw = 5$	35	11.7%	75.7%
D-Isomap	$k = 12$	20	9.2%	82.7%

We compute the reduced dimension of the embedding space**,** standard deviation and average rate of emotion recognition between three Isomap methods. Table 6.2 indicates that the proposed algorithm achieves better performance than the original Isomap and extended Isomap. The fact that the D-Isomap outperforms the other two is that this method can compact the data points from the same cluster on a high-dimension manifold to make them closer in the low-dimension space, and increase the distance between the data points from the different clusters. This ability could be beneficial in preserving the homogeneous characteristics for emotion classification.

6.5. Chapter Summary

In this Chapter we present a 3D EBS based emotion recognition system using active deformable information from video sequences. The overall accuracy of the system is about 82.7%. From experimental results we find that the significant features to distinguish one individual emotion from the other emotions are different. Some of the

features selected in a global scenario are redundant, and some of the other features might contribute to the classification of a specific emotion. Another observation is that there is not even a single feature which is significant for all the classes. This actually reveals that the nature of human emotion, which means that there are no sharp boundaries between emotions. One emotion might have similar patterns with some of the other emotions, and different patterns from the others. The human perception on emotion is based on the integration of different patterns.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In the emotion recognition field, current techniques for the detection and tracking of facial expressions are sensitive to head pose, clutter, and variations in lighting conditions. These issues can potentially be addressed effectively by 3D face modeling and analysis, which has not been widely studied with respect to human facial emotion recognition. The methods and algorithms developed in this dissertation attempt to solve such problems by using Gabor filtering and EBS based methods. Real 3D visual feature extraction using the 3D Gabor library and active deformation extraction from video sequences using the 3D EBS are introduced. These methods benefit from the advancement of computer vision techniques and applications in communication and information technology. The merits of this work are summarized below.

1) For the 3D Gabor feature based method, we employ the 3D Gabor library to extract the visual feature vector from expressional faces, which is the first attempt using such visual

features for face analysis. The filter's scale, orientation, and shape are specified according to the geometric pattern of the 3D facial expressions. The Gabor library is convolved with each set of the 3D data to extract the feature vector by combining the frequency and orientation information from the library at each voxel. For each training sample, the semantic ratings describing the facial expressions are constructed into a seven dimensional semantic expression vector. The IKCCA is used for learning the correlation between the testing sample feature vector and the semantic expression vector. According to this correlation, we calculate the associated semantic expression vector and perform the classification.

2) Face regions are automatically detected from input video sequences using local normalization technique and a coarse-to-fine classification strategy, which can alleviate a common problem in conventional detection and tracking methods: inconsistent performance due to sensitivity to variation illuminations such as local shadowing, noise and occlusion. Our method decreases computing time by candidate's localization with optimal adaptive correlation techniques and locates faces automatically on a single frame to make it possible to eliminate the manual initiation step from the head/face tracking algorithm.

3) Using a fiducial point detector and tracker, facial expressions can be detected and tracked automatically in real time. For detecting the fiducial point, we construct a set of 26 fiducial point detectors with the scale invariant feature. For the tracking part, we combine colour-based kernel correlation technique for the observation likelihood with DE-MC particle filtering distribution for multiple point tracking. It is achieved by forming the proposal distribution for the particle filter from a mixture of the kernel correlation in the current frame and the dynamic model predicted from the previous step. We use the M-component non-parametric mixture model for the multiple DE-MC particle filters' posterior distribution over the states of all the target points. We also adopt an adaptive modification that can reinitialize the position when point loss occurs. It improves the performance of the trackers to cope with the occlusion or disappearance cases.

4) Using elastic body modeling of the face to exhibit different facial expressions with elastic characteristics. Based on the continuity condition, the elastic property of each facial expression is found, and a complete wireframe face model is generated with the availability of some limited feature point positions. An adaptive partition of polygons is embedded in the EBS according to the surface curvature through the characteristic feature points. The subtle structural information can be expressed without giving complicated

facial features. The generic 3D face model is established so that the good parameters of the EBS can be used for emotion recognition, e.g. the appropriate physical characteristics for face deformations, control points. The D-Isomap is introduced to emotion classification due to the fact that this method can compact the data points from the same emotion class on a high-dimension manifold to make them closer in the low-dimension space, and makes the data points further from the different clusters as well. It results in a high recognition rate compared with other Isomap methods

7.2 Future Work

We perform emotion analysis on six basic emotions. However, human emotional states do not have a sharp boundary. Some of the emotions are a combination of different emotions. For instance, humans can express different kinds of surprise, sometimes combined with happiness, and sometimes with fear. For a natural human computer interface, the computer needs to recognize and analyze these situations. One proposal is to categorize emotion into a wide range of classes. Another might be giving different weights to the basic emotional elements, and humans can understand the potential components of an expression.

The emotion analysis is performed on a video signal that has only one emotion. In real life applications, the user's emotions are changing frequently. A scheme needs to be proposed to separate the variations of human emotion in one continuous interaction. Furthermore, the system should be capable of detecting the presence of the user automatically. To further improve detection accuracy, the bounding box should be smoothed and more accurately attached to the detected face. It can be achieved using overall averages of a fixed number of frames from the video sequences. Moreover, the processing time for each frame is about 0.6 second in our work by using a PC (Pentium4, 3 GHz, 2GB RAM) with Matlab 2008a, much higher than general video display time (30 frames per second). We will investigate a proper face tracking method to reduce the total computation time.

In [25], it is shown that combining prosodic audio features and static visual cues leads to a good emotion recognition rate. Since we have demonstrated, in this dissertation, that visual emotion recognition using the EBS-based method with dynamic features markedly outperformed the method using static visual features. Fusing dynamic visual features and the prosodic audio features can potentially yield further improved system performance significantly. So the fusion of the dynamic audio and visual features for emotion recognition will be one research topic worth conducting.

We use one classifier for all the six emotions. However, different classes might have different classification algorithms that can better model the data. An investigation based on these scenarios will help to improve the effectiveness of the system.

Bibliography

[1] M. Pantic, A. Pentland, A. Nijholt, and T.S. Huang, "Human Computing and Machine Understanding of Human Behavior: A Survey," *Proceeding of Eighth ACM international conference on Multimodal Interfaces*, pp. 239-248, Alberta 2006.

[2] B. Reeves, C. Nass, The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places, Cambridge Univ. Press, 1996.

[3] I. Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding* **,** vol. 91 (1-2), pp. 160 – 187, July 2003.

[4] R.W. Picard, Affective Computing, Cambridge: MIT Press, 1997.

[5] C. Darwin, The Expression of Emotions in Man and Animals, John Murray,1872, reprinted by University of Chicago Press, 1965.

[6] P. Ekman, M. O'Sullivan, "The role of context in interpreting facial expression: Comment on Russell and Fehr", *Journal of Experimental Psychology, General*, vol. 117 (1), pp. 86-88, 1988.

[7] A. J. Calder, A. D. Lawrence, and A. W. Young, "Neuropsychology of fear and loathing", *Nature Reviews Neuroscience*, vol. 2 (5), pp. 352-363, 2001.

[8] R. R. Cornelius, The Science of Emotion, Research and Tradition in the Psychology of Emotion, Upper Saddler River: Prentice Hall, 1996.

[9] P. Lang, "The emotion probe: Studies of motivation and attention," *American Psychologist*, vol. 50 (5), pp. 372–385, 1995.

[10] H.Schlosberg, "Three dimensions of emotion," *Psychological Review*, vol. 61 (2), pp. 81–88, 1954.

[11] P. Ekman, T. Dalgleish, M.E. Power, Basic emotions, Handbook of Cognition and Emotion, Wiley, Chichester, U.K., 1999.

[12] C. Calhoun, R. C. Solomon, What is an Emotion, New York: Oxford University Press, 1984.

[13] N. Ambady and R. Rosenthal, "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences," *A Meta-Analysis: Psychological Bull.*, vol. 111 (2), pp. 256-274, 1992.

[14] A.Mehrabian, "Communication with Words," *Psychology Today*, vol. 2 (4), pp. 53-56, 1968.

[15] P. Ekman, T.S. Huang, T.J. Sejnowski, and J.C. Hager, NSF Understanding the Face: A Human Face eStore, 1993.

[16] P. Ekman, W.V. Friesen, J.C. Hager, The Facial Action Coding System: A Technique for the Measurement of Facial Movement, San Francisco: Consulting Psychologist, 2002.

[17] M. Pantic, L.J. Rothkrantz, "Facial action recognition for facial expression analysis from static face image", *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, Vol. 34 (3), pp. 1449-1461, 2004.

[18] M.J. Lyons, J. Budynek, A. Plante, S. Akamatsu, “Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis”, *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition*, pp. 202-207, Grenoble, March 2000.

[19] L.C.D. Silva, S.C. Hui, “Real-time facial feature extraction and emotion recognition”, *Proceedings of 4th International Conference on Information, Communications and Signal Processing*, pp. 1310-1314, Singapore, December 2003.

[20] I. Cohen, N. Sebe, Y. Sun, M. S. Lew, T.S. Huang, "Evaluation of expression recognition techniques", *Proceedings of International Conference on Image and Video Retrieval*, pp. 184-195, Urbana, IL, July 2003.

[21] G. Guo, C.R. Dyer, "Learning from examples in the small sample case: face expression recognition", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, (3), pp. 477-488, June 2005.

[22] M. Pantic, I. Patras, “Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 36 (2), pp. 433-449, April 2006.

[23] K. Anderson, P.W. McOwan, "A real-time automated system for the recognition of human facial expressions", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 36(1), pp. 96-105, February 2006.

[24] H. Gunes, M. Piccardi, "Automatic Temporal Segment Detection and Affect Recognition from Face and Body Display", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39 (1), pp. 64-84, February 2009.

[25] Y. Wang, L. Guan, "Recognizing Human Emotional State from Audiovisual Signals", *IEEE Transactions on Multimedia*, vol. 10 (5), pp. 659 – 668, August 2008.

[26] M. Song, Z. Dong, C. Theobalt, H.Q. Wang, Z.C. Liu, H.P. Seidel, "A General Framework for Efficient 2D and 3D Facial Expression Analogy", *IEEE Transactions on Multimedia*, vol.9 (7), pp. 1384-1395, November 2007.

[27] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, T.S. Huang, "A study of non-frontal-view facial expressions recognition", *19th International Conference on Pattern Recognition*, pp. 1-4, Florida, December 2008.

[28] S. Chin, K.Y. Kim, "Emotional Intensity-based Facial Expression Cloning for Low Polygonal Applications", *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 39 (3), pp. 315-330, May 2009.

[29] L. Yin, et al., "A 3D Facial Expression Database for Facial Behavior Research", *Automatic Face and Gesture Recognition*, pp. 211 – 216, Southampton, April 2006.

[30] A. Chakraborty, A. Konar, U.K. Chakraborty, A. Chatterjee, "Emotion Recognition From Facial Expressions and Its Control Using Fuzzy Logic", *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 39 (4), pp. 726-743, July 2009.

[31] S. Bashyal, G.K. Venayagamoorthy, "Recognition of facial expressions using Gabor wavelets and learning vector quantization", *Engineering Applications of Artificial Intelligence*, vol. 21 (7), pp. 1056-1064, October 2008.

[32] L. Shen, L. Bai, "Face recognition based on Gabor features using kernel methods", *the 6th IEEE Conference on Face and Gesture Recognition*, pp. 170-175, Korea, 2004.

[33] N. Rose, "Facial Expression Classification using Gabor and Log-Gabor Filters", *the 7th International Conference on Automatic Face and Gesture Recognition*, pp. 346-350, Southampton, April 2006.

[34] Y.L. Tian, T. Kanade, J. F. Cohn, "Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity", *the Fifth IEEE Conference on Automatic Face and Gesture Recognition*, pp. 229-234, Washington, May 2002.

[35] H. Tang, T.S. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points ," *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1 – 6, Amsterdam, September 2008.

[36] M. Feng, T.R. Reed, "Motion Estimation in the 3-D Gabor Domain", *IEEE Transactions on Image Processing*, vol. 16 (8), pp. 2038-2047, August 2007.

[37] Q. Zhen, D.N. Metaxas, L. Axel, "Extraction and Tracking of MRI Tagging Sheets Using a 3D Gabor Filter Bank", *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp.711-714, New York, August 2006.

[38] B. Kepenekci, G.B. Akar, “Motion analysis using 3D Gabor kernels”, Signal Processing, Communication and Applications Conference, pp.1-4, SIU April 2008.

[39] B. Scholkopf, A.J. Smola, Learning with Kernels, Cambridge, MA: MIT Press, 2001.

[40] P.L. Lai, C. Fyfe, “Kernel and nonlinear canonical correlation analysis”, *International Journal of Neural System*, vol. 10 (5), pp. 365-377, 2000.

[41] Kenji Fukumizu, Francis R. Bach, Arthur Gretton, “Statistical Consistency of Kernel Canonical Correlation Analysis”, *Journal of Machine Learning Research*, vol. 8, 361-383, August. 2007.

[42] K. Okajima, "Two-dimensional Gabor-type receptive field as derived by mutual information maximization," *Neural Networks*, vol. 11, (3), pp. 441-447, 1998.

[43] H. Lu, Z. Wang, X. Liu, “Facial Expression Recognition Using NKFDA Method with Gabor Features”, *the 6th World Congress on Intelligent Control and Automation*, pp. 9902-9906, Dalian, June 2006

[44] P. Wu, X. Li, J. Zhou, G. Lei, “Face Expression Recognition Based on Feature Fusion”, International Workshop on Intelligent Systems and Applications, pp. 1 – 4, 2009.

[45] S. Lajevardi, M. Lech, “Facial Expression Recognition Using Neural Networks and Log-Gabor Filters”, Digital Image Computing: Techniques and Applications, pp. 77 – 83, 2008.

[46] H. Tang, T. S. Huang, “3D Facial Expression Recognition Based on Automatically Selected Features”, CVPRW, pp. 1-8, June 2008.

[47] H. Soyel, H. Demirel, “Facial Expression Recognition Using 3D Facial Feature Distances”, ICIAR, pp. 831-838, 2008.

[48] J. Wang, L. Yin, X. Wei, Y. Sun, “3D facial expression recognition based on primitive surface feature distribution”, In *Proc. Conf. Computer Vision and Pattern Recognition*, vol 2, pp. 1399-1406, 2006.

[49] C. Chen, S. Chiang, “Detection of human faces in color images” *IEE Proceedings, Vision, Image and Signal Processing*, pp. 127-130, Chicago, October 1998.

[50] R. Hsu, A. Mohamed, A. Jain, “Face Detection in Color Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 24 (5), pp. 696 – 706, May 2002.

[51] F. Tsalakanidou, S. Malassiotis, M. Strintzis, “Face Localization and Authentication Using Color and Depth Images”, *IEEE Transactions on Image Processing*, vol. 14 (2), pp. 152 – 168, February 2005.

[52] S. Phung, S. Bouzerdoum, S. Chai, “Skin Segmentation Using Color Pixel Classification: Analysis and Comparison”, *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 27 (1), pp. 148 – 154, January 2005.

[53] V. Govindaraju, “Locating Human Faces in Photographs”, *International Journal Computer Vision*, vol. 19 (2), pp. 129 – 146, August 1996.

[54] Y. Hori, T. Kuroda, “0.79-mm2 29-mW Real-Time Face Detection Core”, *IEEE Journal of Solid-State Circuits*, vol. 42 (4), pp. 790-797, April 2007.

[55] R. Verma, C. Schmid, K. Mikolajczyk, “Face Detection and Tracking in a Video by Propagating

Detection Probabilities", *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 25 (10), pp. 1215 – 1228, October 2003.

[56] D. Nguyen, D. Halupka, P. Aarabi, A. Sheikholeslami, "Real-Time Face Detection and Lip Feature Extraction Using Field-Programmable Gate Arrays", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36 (4), pp. 902-912, August 2006.

[57] K. Kollreider, H. Fronthaler, M. Faraj, J. Bigun, "Real-Time Face Detection and Motion Analysis with Application in 'Livenes' Assessment", *IEEE Transactions on Information Forensics and Security*, vol. 2 (3), pp. 548 - 558 September 2007.

[58] P. Vadakkepat, P. Lim, L. Silva, L. Jing, L. Ling, "Multimodal Approach to Human-Face Detection and Tracking", *IEEE Transactions on Industrial Electronics*, vol. 55 (3), pp. 1385 - 1393 March 2008.

[59] H. Rowley, S. Baluja, T. Kanade, "Neural network-based face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 20 (1), pp. 23-38, January 1998.

[60] R. Feraud, O. J. Bernier, J. Viallet, M. Collobert, "A fast and accurate face detector based on neural networks", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 23 (1), pp. 42-53, January 2001.

[61] C. Liu, "A Bayesian Discriminating Features Method for Face Detection", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 25 (6), pp. 725-740, June 2003.

[62] C. Garcia, M. Delakis, "Convolutional Face Finder: A Neural Architecture for Fast and Robust Face

Detection", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 26 (11), pp. 1408-1423 November 2004.

[63] K. Sung, T. Poggio, "Example-Based Learning for View- Based Human Face Detection", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 20 (1), pp. 39-51, January 1998.

[64] E. Osuna, R. Freund, F. Girosi, "Training Support Vector Machines: An Application to Face Detection," *Proceedings Conference, Computer Vision and Pattern Recognition*, pp. 130-136, San Juan, January 1997.

[65] H. Sahbi, D. Geman, N. Boujemaa, "Face detection using coarse-to-fine support vector classifiers", *Proceedings IEEE International Conference Image Processing*, pp. 925 – 928, New York, June 2002.

[66] C. Waring, X. Liu, "Face Detection Using Spectral Histograms and SVMs", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35 (3), pp. 467 – 476, June 2005.

[67] H. Schneiderman, T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition", *Proceedings Conference Computer Vision and Pattern Recognition, Santa Barbara*, pp. 45-51, CA, June 1998.

[68] H. Schneiderman, "Learning statistical structure for object detection", *Proceedings 10th International Conference Computer Analysis Images and Patterns*, pp. 434-441, Groningen, August 2003.

[69] X. Li, X. Zhou, "Automatic real-time face detection and tracking based on space-temporal mutual feedback for video sequence", *International Conference on Audio Language and Image Processing*,

pp. 1650 – 1654, Shanghai, July 2008.

[70] V. Pallavi, J. Mukherjee, A. K. Majumdar, S. Sural, "Graph-Based Multiplayer Detection and Tracking in Broadcast Soccer Videos", *IEEE Transactions on Multimedia*, vol. 10 (5), pp. 794 – 805, August 2008.

[71] Viola, Jones, "Robust Real Time Object Detection", *Proceedings 2nd International Workshop on Statistical and Computational Theories of Vision*, pp. 1-25, Vancouver, July 2001.

[72] X. Tan, B. Triggs, "Enhanced local texture feature sets for Face Recognition under Difficult Lighting Conditions", *IEEE Transactions on Image Processing* , vol. 19 (6), pp. 1635 – 1650, 2007.

[73] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 23 (6), pp. 643 – 660, June 2001.

[74] R. Ramamoorthi, "Analytic PCA Construction for Theoretical Analysis of Lighting Variability in Images of a Lambertian Object", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 24 (10), pp. 1322 – 1333, October 2002.

[75] R. Basri, D.W. Jacobs, "Lambertian Reflectance and Linear Subspaces", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 25 (2), pp. 218 – 233, February 2003.

[76] S. Li, R. Chu, S. Liao, L. Zhang, "Illumination Invariant Face Recognition Using Near-Infrared Images", *IEEE Transaction on Pattern and Machine Intelligent*, vol. 29 (4), pp. 627 - 639 April 2007.

[77] W. Chen, M. Er, S. Wu, “Illumination Compensation and Normalization for Robust Face Recognition Using Discrete Cosine Transform in Logarithm Domain”, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36 (2), pp. 458 - 466 April 2006.

[78] O. Arandjelov, R. Cipolla, “An illumination invariant face recognition system for access control using video”, *Proceedings British Machine Vision Conference*, Kingston September 2004.

[79] David G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol.60 (2), pp. 91-110 May 2004.

[80] L. Yaroslavsky, M. Eden, Fundamentals of Digital Optics, Birkhauser, Boston, 1996.

[81] B. Wu, H. Ai, C. Huang, S. Lao, “Fast rotation invariant multi-view face detection based on RealAdaboost”, *Proceedings IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 79-84, Seoul, May 2004.

[82] J. Friedman, T. Hastie, R. Tibshirani, “Additive logistic regression:A statistical view of boosting”, The Annals of Statistics, vol. 28, pp. 2000, April 2000.

[83] A. Vezhnevets, V. Vezhnevets, “'Modest AdaBoost' - Teaching AdaBoost to Generalize Better”, Graphicon-2005, Novosibirsk Akademgorodok, 2005.

[84] http://research.graphicon.ru/machine-learning/gml-adaboost-matlab-toolbox.html.

[85] M. Du, L. Guan, “Monocular Human Motion Tracking with the DE-MC Particle Filter”, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 14-19，Toulouse, May

2006.

[86] J.F. Cohn, A.J Zlochower, J.J. Lien and T. Kanade, "Feature point tracking by optical flow discriminates subtle differences in facial expression", *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 396 – 401, Nara, April 1998.

[87] M. Maghami, R. A. Zoroofi, B. N. Araabi, M. Shiva and E. Vahedi, "Kalman Filter Tracking for Facial Expression Recognition using Noticeable Feature Selection", *International Conference on Intelligent and Advanced Systems*, pp. 587 – 590, Kuala Lumpur, November 2007.

[88] Jian Huang Lai, P.C. Yuen, Wen Sheng Chen, S. Lao, M. Kawade, "Robust facial feature point detection under nonlinear illuminations", *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 168-174, Vancouver, July 2001.

[89] Ehsan Fazl Ersi, Kiana Hajebi, "Face recognition by fiducial point analysis", *IEEE CCECE*, pp. 1187 - 1190, Montral, May 2003.

[90] Michel Valstar and Maja Pantic, "Fully Automatic Facial Action Unit Detection and Temporal Analysis", *Computer Vision and Pattern Recognition Workshop*, pp. 149 - 149 New York, June 2006.

[91] I. Patras, M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features", *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 97-102, Southampton, May 2004.

[92] C. Sminchisescu, B. Triggs, "Covariance Scaled Sampling for Monocular 3D Body Tracking",

Proceedings of Interntional Conference on Computer Vision and Pattern Recognition, pp. 8-14, Hawaii, December 2001.

[93] Y. Rui, Y. Chen, "Better Proposal Distributions: Object Tracking using Unscented Particle Filter", Proceedings of Interntional Conference on Computer Vision and Pattern Recognition, pp. 786-793, Hawaii, December 2001.

[94] J. Deutscher, A. Blake, I. Reid, "Automatic Partitioning of High Dimensional Search Spaces Associated with Articulated Body Motion Capture", Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 669-676, Hawaii, December 2001.

[95] C. Hue, L. Cadre, P. P´erez, "Tracking Multiple Objects with Particle Filtering", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38 (3), pp. 791- 812, July 2002.

[96] M. Isard, J. MacCormick, "BraMBLe: A Bayesian multiple-blob tracker", *IEEE International Conference on Computer Vision*, pp. 34-41, Vancouver, August 2001.

[97] J. Vermaak, A. Doucet, P. P´erez, "Maintaining Multi-Modality through Mixture Tracking", *Ninth IEEE International Conference on Computer Vision*, pp. 1110-1116, Beijing, October 2003.

[98] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 46–53, Grenoble, March 2000.

[99] Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, and Jacqueline J. Hill, "Mind Reading: The Interactive Guide to Emotions", London, Jessica Kingsley, Publishers, 2004.

[100] E. Arnaud, E. Memin, B. Cernuschi-Frias, "Conditional filters for image sequence based tracking application to point tracking", IEEE Transactions on Image Processing, vol. 14 (1), pp. 63-79， May 2005.

[101] S. Birchfield "Elliptical Head Tracking Using Intensity Gradients and Color Histograms", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 232-237, June 1998.

[102] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions", *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 31 (1), pp. 39-58, January 2009.

[103] K. J., Bathe, Finite Element Procedures in Engineering Analysis, PrenticeHall, 1982.

[104] S. Platt, N. Badler, "Animating facial expression", *Computer Graphics* vol. 15 (3), pp.245-252, August, 1981.

[105] T. Cootes, G. Edwards, C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 23 (6), pp. 681–685, Jun. 2001.

[106] J. Peyras, A. Bartoli, and S. Khoualed, "Pools of AAMs: Towards automatically fitting any face image," *Electronic Proceedings of the 19th British Machine Vision Conference*, pp. 681–685, Leeds, September 2008.

[107] M. Song, Z. Dong, C. Theobalt, H.Q. Wang, Z.C. Liu, H.P. Seidel, "A General Framework for

Efficient 2D and 3D Facial Expression Analogy", *IEEE Transactions on Multimedia*, vol.9 (7), pp. 1384-1395, November 2007.

[108] S. Chin, K.Y. Kim, "Emotional Intensity-based Facial Expression Cloning for Low Polygonal Applications", *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 39 (3), pp. 315-330, May 2009.

[109] J.W. Hole Jr. and K. A. Koos, Human Anatomy. Dubuque, IA: Brown, 1991.

[110] M.H. Davis, A. Khotanzad, D.P Flamig, S.E Harms, "A physics-based coordinate transformation for 3-D image matching", *IEEE Transactions on Medical Imaging*, vol. 16 (3), pp. 317 -328, June 1997.

[111] C. J. Kuo, J. Hung, M.Tsai, P. Shih, " Elastic Body Spline Technique for Feature Point Generation and Face Modeling", *IEEE Transactions on Image Processing*, vol.14 (12), pp. 2159-2166, December 2005.

[112] J. B. Tenenbaum, V. de Silva, and john C. Langford, "A global geometric framework for nonlinear dimensional reduction," *Science*, vol. 290 (3), pp. 2319-2323, December 2000.

[113] M.H. Yang, "Face recognition using extended isomap," *In Proceeding of ICIP*, pp. 117-120, New York, September 2002.

[114] X. Geng, D.C. Zhan, Zhi-Hua Zhou, "Supervised Nonlinear Dimensionality Reduction for Visualization and Classification." *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol.35 (6), pp. 1098 – 1107, December 2005.

[115] Yiming Wu, Kap Luk Chan, Lei Wang, "Face recognition based on discriminative manifold learning" *International Conference on Pattern Recognition*, pp. 171 – 174, Cambridge, August 2004.

[116] D. Zhao, and L. Yang, "Incremental Isometric Embedding of High-Dimensional Data Using Connected Neighborhood Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31 (1), pp. 86-98, January 2009.

[117] T. Friedrich, Nonlinear Dimensionality Reduction with Locally Linear Embedding and Isomap, University of Sheffield, 2002.

[118] E. Kokiopoulou, and Y. Saad, "Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique," *IEEE Transactions on Pattern Analysis and MachineIntelligence*, vol. 29 (12), pp. 2143-2156, December 2007.

[119] J. Handl and J. Knowles, "An Evolutionary Approach to Multiobjective Clustering," *IEEE Transactions on Evolutionary Computation*, Vol.11 (1), pp.56-76, February 2007.

[120] S.Y. Ho and H.L. Huang, "Facial Modeling from an Uncalibrated Face Image Using Flexible Generic Parameterized Facial Models", *IEEE Transactions on Systems, Man and Cybernetics, Part B,* vol.31 (58), pp. 706-719, October 2001.

List of Publications

The publications based on the work of this thesis are listed below:

Journal Papers

- **Yun Tie**, Ling Guan, “Automatic face detection in video sequences using local normalization and optimal adaptive correlation techniques”, *Pattern Recognition*, Volume 42, Issue 9, Pages 1859-1868, September 2009
- **Yun Tie**, Ling Guan, “Human Emotional State Recognition Using Real 3D Visual Features from Gabor Library”, submitted
- **Yun Tie**, Ling Guan, “Automatic Detection and Tracking of Fiducial Points from Emotional Facial Expressions in Video Sequences”, submitted
- Ling Guan, Yongjin Wang, Rui Zhang, **Yun Tie**, Adrian Bulzacki, Muhammad Talal Ibrahim, ”Multimodal information fusion for selected multimedia applications”, *International Journal of Multimedia Intelligence and Security*, Volume 1, Number 1, Pages 5-32, 2010

Conference Papers

- **Yun Tie**, Ling Guan, "Local Normalization with Optimal Adaptive Correlation for Automatic and Robust Face Detection on Video Sequences", *Tenth IEEE International Symposium on Multimedia*, pp.160-165, San Diego, USA, December 2008
- **Yun Tie**, Ling Guan, "Automatic fiducial points detection for facial expressions using scale invariant feature", *IEEE International Workshop on Multimedia Signal Processing*, pp. 1 – 6, Rio de Janero, Brazil, October 2009
- Ling Guan, Paisarn Muneesawang, Yongjin Wang, Rui Zhang, **Yun Tie**, Adrian Bulzacki, Muhammad Talal Ibrahim," Multimedia multimodal methodologies", *IEEE International Conference on Multimedia and Expo*, pp. 1600 – 1603, NYC, USA, June/July 2009

- Ling Guan, Yongjin Wang, **Yun Tie**, “Toward natural and efficient human computer interaction", *IEEE International Conference on Multimedia and Expo*, pp 1560 – 1561, NYC, USA, June/July 2009
- **Yun Tie**, Ling Guan, "Fiducial Point Tracking for Facial Expression Using Multiple Particle Filters with Kernel Correlation Analysis”, *IEEE International Conference on Image Processing*, pp. 373-376, Hongkong, September 2010
- **Yun Tie**, Ling Guan, " Human Emotion Recognition Using Real 3D Visual Features from Gabor Library”, *IEEE International Workshop on Multimedia Signal Processing*, pp. 481-486, Saint Malo, France, October 2010