UNIVERSITY OF CALIFORNIA

Santa Barbara

Facial Expression Analysis on Manifolds

A Dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Computer Science

by

Ya Chang

Committee in charge:

Professor Matthew Turk, Chair

Professor Yuan-Fang Wang

Professor B.S. Manjunath

Professor Andy Beall

September 2006

The dissertation of Ya Chang is approved.

_____

Yang-Fang Wang

_____

B.S. Manjunath

_____

Andy Beall

_____

Matthew Turk, Committee Chair

September 2006

Facial Expression Analysis on Manifolds

by

Ya Chang

Dedicated to my family

ACKNOWLEDGEMENTS

It has been a long journey since I began my Ph.D. study at University of California, Santa Barbara. I am very luck to meet many great people who made this journey enjoyable and fruitful. I would like to express my deep gratitude to those people in particular:

First I would like to thank my advisor Prof. Matthew Turk. He helped me before I could land on the States. In the past four and half years, I could not make such progress without his excellent guidance and constant support. With Matthew's effective administration, Four Eyes Laboratory is filled with free academic atmosphere that encourages me to explore the areas I am interested in. His research attitude will be my motto in my future career. I would also like to thank other members of my doctoral committee: Professor Yuan-Fang Wang, Professor B.S. Manjunath, and Professor Andy Beall, for their inspiring suggestions and critical thinking. I feel fortunate to have such a supportive committee. They made my graduate study a very rewarding and pleasant one.

I would like to thank my colleagues in Four Eyes Laboratory: Prof. Tobias Hollerer for a broader view on augmented reality; Changbo Hu as one of the first lab members who gave me many insightful advice at the early stage of thesis. Rogerio Feris and Haiying Guan joined the lab in the same year with me. We made a lot of valuable research discussion. The same thanks also go to Mathias Kolsch, Lihua Lin,

Sebastian Grange, Jae Sik Chang, Longbin Chen, Steve DiVerti, Justin Muncaster, Jason Wither, Cha Lee, and Hang-Bong Kang.

I had two wonderful summer interns at Microsoft, Redmond. Many thanks to my mentors there: Zicheng Liu and Zhengyou Zhang at Microsoft Research; Luis Ca and Jason Fuller at Windows CE Division.

Santa Barbara is the place where I met and got married with my husband: Beitao Li. He gave me unconditional support and encouragement through all the time. My parents and my brother stand behind me and love me even though they are far away.

When I look back near the end of this journey, all of the warm memory seems like it happened yesterday. The experience at Santa Barbara will become priceless treasure for my new journey. Once again, thank you!

CURRICULUM VITA

Ya Chang

September 2006

EDUCATION

Bachelor of Computer Science, University of Science & Technology, China, July 2001
Doctor of Philosophy in Computer Science, University of California, Santa Barbara, September 2006

PROFESSIONAL EMPLOYMENT

2000-2001: Research Intern, National Laboratory of Pattern Recognition, Chinese Academy of Sciences
2002-2006: Research Assistant, Four Eyes Laboratory, Computer Science Department, University of California, Santa Barbara
Summer 2004: Research Intern, Microsoft Research
Summer 2005:Software Design Engineer Intern, Windows CE Division, Microsoft

SELECTED PUBLICATIONS

Ya Chang, Matthew Turk, "Probabilistic expression analysis on manifolds," *International Conference on Computer Vision and Pattern Recognition*, Washington DC, June 2004.

Ya Chang, Marcelo Vieira, Matthew Turk, and Luiz Velho, "Automatic 3D Facial Expression Analysis in Videos", *IEEE ICCV Workshop on Analysis and Modeling of Faces and Gestures*, 2005.

Ya Chang, Changbo Hu, Rogerio Feris and Matthew Turk, "Manifold based analysis of facial expression," *Journal of Image and Vision Computing,* Volume 24, Issue 6, pp. 605-614, 2006.

PATENT

"Automatic Detection of Panoramic Camera Position and Orientation and Table Parameters," US Application Serial No**.** 11/227,046. filed jointly with R. Cutler, Z. Liu, and Z. Zhang. Microsoft Inc.

AWARDS

5-year UCSB tuition fellowship and assistantship
President's work study award, 2003
IEEE student travel award, 2004

FIELDS OF STUDY

Major Field: Computer Vision
Minor Field: Pattern Recognition, Image Procession, Statistical Learning

ABSTRACT

Facial Expression Analysis on Manifolds

by

Ya Chang

Facial expression is one of the most powerful means for people to coordinate conversation and communicate emotions and other mental, social, and physiological cues. We address two problems in facial expression recognition in this thesis: global facial expression space representation and facial expression recognition method with objective measurement.

We propose the concept of the manifold of facial expression based on the observation that the images of all possible facial deformations of an individual make a smooth manifold embedded in a high dimensional image space. To combine the manifolds of different subjects that vary significantly and are usually hard to align, we transfer the facial deformations in all training videos to one standard model. Lipschitz embedding embeds the normalized deformation of the standard model in a low dimensional **Generalized Manifold.** Deformation data from different subjects complement each other for a better description of the true manifold. We learn a probabilistic expression model on the generalized manifold. There are six kinds of universally recognized facial expressions: happiness, sadness, fear, anger, disgust, and surprise, which we explicitly represent as basic expressions. In the embedded

space, a complete expression sequence becomes a path on the expression manifold, emanating from a center that corresponds to the neutral expression. The transition between different expressions is represented as the evolution of the posterior probability of the six basic expressions.

These six kinds of basic facial expressions comprise only a small subset of all visible facial deformation. To measure the facial expression recognition rate precisely in the manifold model, we developed **Regional FACS** (Facial Action Coding System). FACS encodes facial deformation in terms of 44 kinds of Action Units (AU). By learning the AU combinations in 9 separate facial regions, the number of combinations of regional deformations is dramatically decreased compared to the number of combinations of AUs. The manifold of each facial regional can be considered as the sub-vector of the whole manifold. The experimental results demonstrate that our system works effectively for automatic recognition of 29 AUs that cover the most frequently appearing facial deformations. The FACS recognition results also lead to high recognition accuracy of six basic expression categories.

The main contributions of this thesis are: (1) A probabilistic model based on manifold of facial expression can represent facial expression analytically and globally; (2) The Regional FACS system provides a novel FACS recognition solution with objective measurement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Facial expression is one of the most powerful means for people to coordinate conversation and communicate emotions and other mental, social, and physiological cues. People began to research facial expression hundreds of years ago. Charles Darwin noted in his book *The Expression of the Emotions in Man and Animals*:

> "...the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements."

It has been an active topic in multiple disciplines: psychology, behavior science, sociology, medical science, computer science, etc. In the past decade, computer scientists noticed that facial expression plays an important role in Human Computer Interaction (HCI). Accurate and robust facial expression analysis could improve the performance of facial recognition systems that are used widely in security or surveillance systems. Automatic facial expression analysis can also serve as an efficient tool for scientist in other fields. A robust automatic facial expression system can help to make all these potential applications into reality.

We address two problems in facial expression recognition in this thesis: global facial expression space representation and facial expression recognition method with objective measurement. To explain the research goal, describe the used methods and present the experimental results, we first introduce that how facial expression is represented and classified.

Facial expressions can be classified in various ways – in terms of non-prototypic expressions such as "raised brows," prototypic expressions such as emotional labels (e.g., "happy"), or facial actions such as the Action Units defined in the Facial Action Coding System (FACS) [1]. Some psychologists claim that there are six kinds of universally recognized facial expressions: happiness, sadness, fear, anger, disgust, and surprise [2]. Existing expression analyzers [3,4,5] usually classify the examined expression into one of the basic expression categories. These six basic expression categories are only a small subset of all facial expressions expressible by the human face. For "blended" expressions, it may be more reasonable to classify them quantitatively into multiple expression categories. Considering the intensity scale of the different facial expressions, each person has his/her own maximal intensity of displaying a particular facial action. It is also useful to recognize the temporal intensity change of expressions in videos.

A key challenge in automatic facial expression analysis is to identify a global representation for all possible facial expressions that supports semantic analysis. In this thesis, we explore the space of expression images and propose the manifold of expressions as a foundation for expression analysis, using nonlinear dimensionality reduction to embed facial deformations in a low-dimensional space. The manifold model with probabilistic learning provides a global and analytical framework for facial expression analysis. Another advantage is that the manifold framework facilitates the following training of the expression recognizer. We will talk about the detail of this subject in Chapter 5.

The objective measurement of facial deformation is very essential to the research in this field. These six kinds of basic facial expressions comprise only a small subset of all visible facial deformation. FACS [1] is the mostly widely used and versatile method for measuring and describing all visually detectable facial behaviors. The system defines 44 muscle related AUs with intensity scoring, which are analog of phonemes in facial expression. Using these rules, an expression can be decomposed into the specific AUs. Although the number of AUs is relatively small, more than 7000 different AU combinations have been observed [63]. A major drawback of FACS system is that the annotation is very time consuming even for an expert because of the huge number of combinations of AUs (above 7000) and the subtlety of temporal facial deformations. It takes more than one hour to annotate 100 still images or a minute of videotape in terms of AUs and their temporal segments [1]. So it is an active research topic to build a robust automatic FACS recognition system in the recent years.

We noticed that most of AUs only affect a small sub-region of the whole face, and the number of the combinations of AUs in these sub-regions is dramatically less complex. Motivated by this combinational property of AUs, we propose a novel facial expression analysis system called **Regional FACS** to measure the facial expression recognition rate precisely in the manifold model. By learning the AU combinations in 9 separate facial regions, the number of combinations of regional deformations is dramatically decreased compared to the number of combinations of AUs. The manifold of each facial regional can be considered as a sub-vector of the

whole manifold. The experimental results demonstrate that our system works effectively for automatic recognition of 29 AUs that cover the most frequently appearing facial deformations. It serves as an intermediate layer between FACS annotation and expression recognition because accurate FACS recognition results also lead to high recognition accuracy of six basic expression categories.

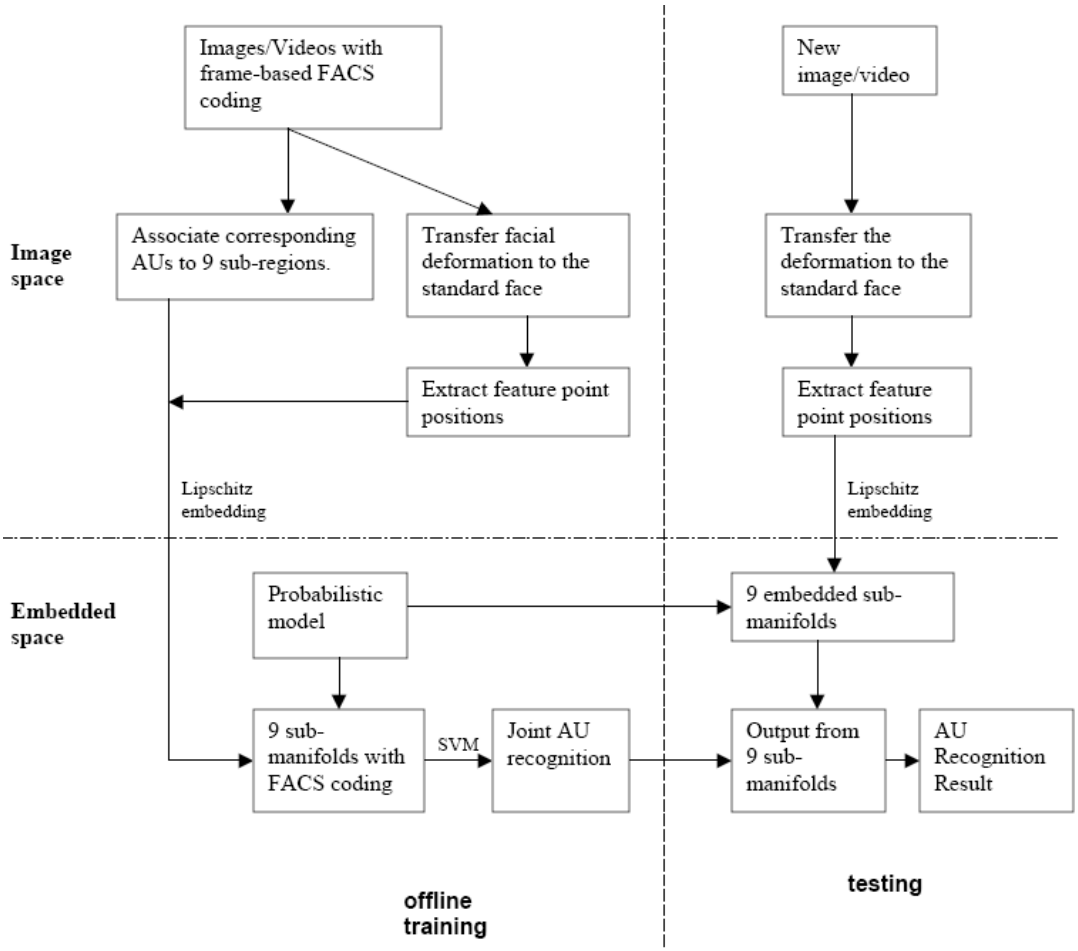


Figure 1.1: System Diagram

We developed a facial expression analysis system which is based on probabilistic manifold model and applied Regional FACS for facial expression recognition. Fig.

1.1 illustrates the overall structure of the Regional FACS system. We will give a system overview in Section 1.1 and describe the thesis outline in Section 1.2. The different modules in the system will be presented in the following chapters in more detail.

1.1 System Overview

The input of the system is static images or videos with human faces. An N-dimensional representation of the face (where N could be the number of pixels in the image or the number of parameters in a face model, for example) can be considered a point in an N-dimensional face space, and the variability of facial expression can be represented as low-dimensional manifolds embedded in this space. People change facial expressions continuously over time. Thus all images of an individual's facial expressions represent a smooth manifold in the N-dimensional face space with the "neutral" face as the central reference point. The intrinsic dimension of this manifold is much lower than N.

Non-linear dimensionality reduction has attracted attention for a long time in computer vision and visualization research [6,7]. Images lie in a very high dimensional space, but a class of images generated by latent variables lies on a manifold in this space. For human face images, the latent variables may be the illumination, identity, pose, and facial deformations.

On the manifold of expressions, similar expressions are points in the local neighborhood on the manifold. Sequences of basic emotional expressions become

paths on the manifold extended from the reference center, as illustrated in Figure 1.2. The blends of expressions lie between those paths, so they can be defined analytically by the positions of the basic paths. The analysis of the relationships between different facial expressions is facilitated on the manifold.

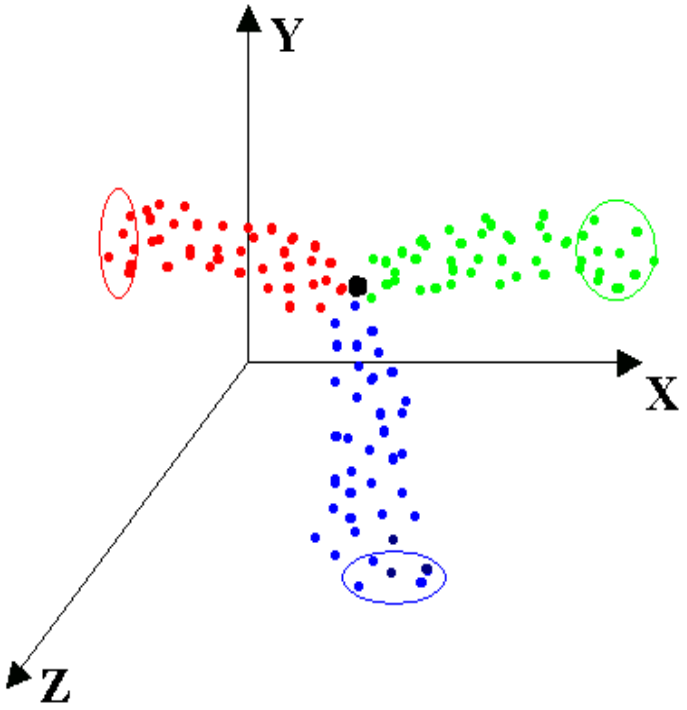


Figure 1.2: Illustration of a 3D expression manifold. The reference center is defined by the face with neutral expression. Image sequences from three different expressions are shown. The further a point is away from the reference point, the higher is the intensity of that expression.

Expression manifolds from different subjects remain difficult to align in the embedded space due to various causes: (1) subjects have different face geometries; (2) facial expression styles vary by subject; (3) some persons cannot perform certain expressions; and (4) the whole expression space is large including blended expressions, so only small portion of it can be sampled. Considering these factors,

bilinear [8] and multi-linear [9] models have been successful in decomposing the static image ensembles into different sources of variation, such as identity and content. Elgammal and Lee [10] applied a decomposable generative model to separate the content and style on the manifold representing dynamic objects. It learned a unified manifold by transforming the embedded manifolds of different subjects into one. This approach assumes that the same kind of expression performed by different subjects match each other strictly. However, one kind of expression can be performed in multiple styles, such as laughter with closed mouth, or open mouth. The matching between these styles is very subjective.

To solve this problem, we built a generalized manifold that is capable of handling multiple kinds of expressions with multiple styles. We transferred the 3D deformation from the models in the training videos to a standard model. Sumner and Popovic [11] designed a scheme for triangle meshes where the deformed target mesh is found by minimizing the transformation between the matching triangles while enforcing the connectivity. We added a temporal constraint to ensure the smooth transfer of the facial deformations in the training videos to the standard model. This model is scalable and extensible. New subjects with new expressions can be easily added in. The performance of the system will improve continuously with new data.

We built a generalized manifold from normalized motion of the standard model. It is a formidable task to learn the complete structure of the manifold of expressions in a high dimensional image space. To overcome this problem, our core idea is to embed the nonlinear manifold in a low dimensional space and recognize facial

expression from video sequences probabilistically. Lipschitz embedding [12,13] was developed to embed the manifold to a low dimensional space, while keeping the main structure of the manifold. Lipschitz embedding leads to good preservation of clusters in practical cases [14,15]. After Lipschitz embedding, the expression sequences in the gallery become paths emanating from the center, which is defined by the neutral expression. A probabilistic model was learned on the generalized manifold in the embedded space. We learn the probabilistic model of transition between those paths from the gallery videos. The probe set includes videos of random expression changes, which may not begin or end with neutral expression. The duration and the intensity of the expression are varied. The transition between different expressions is represented as the evolution of the posterior probability of the basic paths. Our empirical study demonstrates that the probabilistic approach can recognize expression transitions effectively. We also synthesize image sequences of changing expressions through the manifold model.

The objective measurement of the facial expression recognition rate has been a crucial problem in current research. FACS provides objective and precise measurement of all visible facial deformations. But there are more than 7000 combinations of AUs. It is also very hard to connect the combinations of AUs with the emotional expression in an analytical way due to the discrete nature of AUs. We developed regional FACS to use FACS to measure the accuracy of the facial expression recognition method. The face can be divided into 9 sub-regions. Each sub-region is small and does not overlap with the others. Most of AUs only affect

some specific sub-region. The possible deformation on each sub-region can be represented by much fewer combinations of AUs. The sub-manifold of some specific sub-region can be represented by only the AUs related to it.

We divide the face into nine sub-regions. The possible combination of AUs in each sub-region is dramatically reduced. The system exploits the independence of AUs affecting disconnected sub-regions at the same time. We extract features of sub-regional facial deformation and embed the high dimensional features into low dimensional sub-manifolds. The dimensionality of the sub-manifolds is determined by the number of AUs associated with them. We build a probabilistic model on the sub-manifolds for AU recognition. A Support Vector Machine is trained to combine the outputs from all sub-manifolds into final AU recognition results. We tested our system on the MMI face database [46] and the Cohn-Kanade facial expression database [30]. The experimental results demonstrate that our system works effectively for automatic recognition of 29 AUs that cover the most frequently appearing facial deformation in normal life. The FACS recognition results also lead to high recognition accuracy of six basic expression categories.

1.2 Outline

The remainder of this thesis is organized as follows.

We review the related work in Chapter 2. The relationship between our research and previous one will be revealed. We will also briefly compare our work with work in other laboratories.

Chapter 3 presents the concept of the manifold of facial expression. We also discuss how to build the manifold from training data through nonlinear dimensionality reduction and how to build generalized manifold from different subjects. The generalized manifold will improve constantly when new subjects are added to the training database.

In Chapter 4, we describe the probabilistic model on the generalized manifold. Facial feature tracking in the embedded space and facial expression recognition are performed in a cooperative manner within a common probabilistic framework. In contrast with traditional methods that consider expression tracking and recognition in separate stages, our method provides more robust results during large facial deformation.

Chapter 5 describes the idea of Regional FACS that is based on an expression manifold model. A probabilistic model is built on the sub-manifolds for AU recognition. We tested our system thoroughly on different databases. The experimental results demonstrate that our system works effectively for automatic recognition of 29 AUs that cover the most frequently appearing facial deformation in normal life. The FACS recognition results also lead to high recognition accuracy of six basic expression categories.

Chapter 6 explores real time 3D facial deformation capture and registration. We built a preliminary 3D facial expression database with real time date. The algorithms of 3D deformation transfer in Chapter 3 and expression editing in Chapter 4 are tested on our 3D facial expression database.

We review the thesis and discuss its strengths and weaknesses in Chapter 7. Facial expression has been an active research filed for many years. We also propose future research direction in the field.

Chapter 2

Related Research

The face plays an essential role in interpersonal communication. It helps to coordinate conversation and to communicate emotions and other meaningful mental, social, and physiological cues. Mehrabian [18] indicated that whether the listener liked or disliked a message depends only 7% on the spoken word, 38% on vocal utterances, while facial expressions determine 55% of this feeling. This implies that the facial expressions form the major channel in human interpersonal communication.

Over the last decade, automatic facial expression analysis has become an active research area that finds potential applications in HCI, graphics and animation, talking agents, teleconferencing and human emotion interpretation. An automatic facial expression analyzer should have the ability of bringing facial expressions into man-machine interaction as a new modality and making the interaction tighter and more efficient. Humans detect faces and interpret facial expressions in a scene with little or no effort, but automatic analysis and classification of facial expressions is a very difficult task. We still need to extract facial expression data for correct classification of facial expressions based on the facial feature tracking results. Numerous techniques have been proposed in the literature in recent years. Some surveys [16,17] gave a detailed review of existing methods on facial expression

analysis. In this section, we will survey important works, discuss their main advantages and limitations separately.

2.1 Facial Expression Data Extraction

The first step for a fully automatic facial expression analyzer is to extract the information about the encountered facial expression in an automatic way. Both the kind of input data and the representation of visual information affect the choice of the approach to facial expression information extraction.

The input data can be static images or image sequences. There are three types of face representation: holistic, analytic and hybrid. The holistic approaches fit a **template** face model to the input image or track it in the input image sequence. The analytic approaches localize the **features** of an analytic face model in the input image or track them in the input sequences. The hybrid approaches combine the above two methods to some extend.

It is a more difficult task to extract facial expression data from a video sequence because we encounter the 3-D head movement simultaneously. Feature points are more prone to drift without geometry constraints in long video sequences. Generally holistic approaches can achieve a better global result.

The holistic approaches include Active Appearance Model (AAM) [19], Point Distribution Model (PDM) [20], optical flow in facial region, etc. The analytic approaches include Facial Characteristic Points (FCP) [21], Dual-view point-based

model [22], optical flow on facial points, etc. The hybrid approaches includes labeled graph [23], potential net [24], Gabor wavelet [25], etc.

Since generally holistic and hybrid approaches use a template and analytic approaches focus on facial features, we discuss holistic and hybrid approaches in Section 2.1.1 (template-based methods) and analytic approaches in Section 2.1.2 (feature-based methods).

2.1.1 Template-based methods

Generally the template-based methods use a holistic or hybrid approach to face representation.

Cootes et al. [19] presented an Active Appearance Model that can represent both the shape and texture variability seen in a training set. An example image can be synthesized by generating a texture image and warping it using the control points. Fig. 2.1 shows the effects of varying the first two appearance model parameters of a model trained on a set of face images.

c_1 varies by ± 2 s.d.s c_2 varies by ± 2 s.d.s

Figure 2.1: First two modes of an appearance model of a face, from [19]

Hong [23] utilized a labeled graph, called General Face Knowledge (GFK), to represent the face. Each node of the graph consists of an array, which is called jet. Each component of a jet is the filter response of a certain Gabor wavelet extracted at a point of the input image. Hong use GKF to find the exact face location in an input facial image and to localize the facial features. The dense model-graph coding is very suitable for facial action coding based on the extracted deformations of the graph. However, this issue has not been discussed in [23].

Huang [20] represented face by using point distribution model (PDM). The used PDM has been generated from 90 facial feature points that have been manually localized in 90 images of 15 subjects showing six basic emotions. The PDM models the face as a whole and interacts with the estimated face region of an input image as entire. After an initial placement of the PDM in the input image, the entire PDM can be moved and deformed simultaneously. Here, a gradient-based shape parameters estimation, which minimizes the overall gray-level model fitness measure, is applied. The face should be without facial hair and glasses, no rigid head motion may be encountered, so successfulness of the method is strongly constrained.

Methods in this category also use a holistic or hybrid approach to face representation in general. Black and Yacoob [26] used local parameterized models of image motion for facial expression analysis. They utilize a planar model to represent rigid facial motions. The motion of the plane is used to stabilize two frames of the examined image sequence and the motions of the facial features are then estimated relatively to the stabilized face. Nonrigid motions of facial features within the local

facial areas of the eyebrows, eyes, and mouth are represented by affine-plus-curvature model. In their approach, the initial regions for the head and the facial features were selected by hand and thereafter automatically tracked.

Another system, which utilizes a hybrid approach to face representation, was proposed by Kimura [24]. They try to fit the Potential Net to a normalized facial image. They compute first the edge image by applying a differential filter. Then, in order to extract the external force, which is a smooth gradient of the edge image, they are applying a Gaussian filter. The filtered image is referred to as a "potential field" to which the elastic net model is placed. The net deforms further governed by the elastic force of the potential field. The facial action is encoded by the extracted deformations of the net. In their system, rigid head motions are not allowed, so the usefulness of the system is constrained.

Tian [27] located eye and mouth facial feature in by assuming the dual states of eye and mouse open and close. They also quantify the amount and orientation of furrows by using Canny edge detector. The multi-state face component model gives a good global constraint for mouse shapes and eye shapes.

2.1.2 Feature-based methods

Generally, the feature-based methods use an analytic approach to face representation.

Kobayashi and Hara [21] developed a real-time system that works with online taken images of subjects with no facial hair or glasses facing the camera while sitting at approximately 1m distance from it. They utilize a geometric face model of

30 FCPs. They first obtain a set of brightness distributions of 13 vertical lines crossing the FCPs. Then these data are given further to a trained NN for expression emotional classification. A shortcoming of the proposed face representation is that the facial appearance changes encountered in a horizontal direction cannot be modeled.

Pantic and Rothkrantz [22] are utilizing a point-based model composed of two 2D facial views, the frontal and the side view. The frontal-view face model is composed of 30 features. The side-view face model consists of 10 profile points. Fig. 2.2 shows the points on frontal and profile view. They apply multiple feature detectors for each prominent facial feature, then choose the best of the acquired (redundant) results. The system cannot deal with minor inaccuracies of the extracted facial data and it deals merely with images of faces without facial hair or glasses.

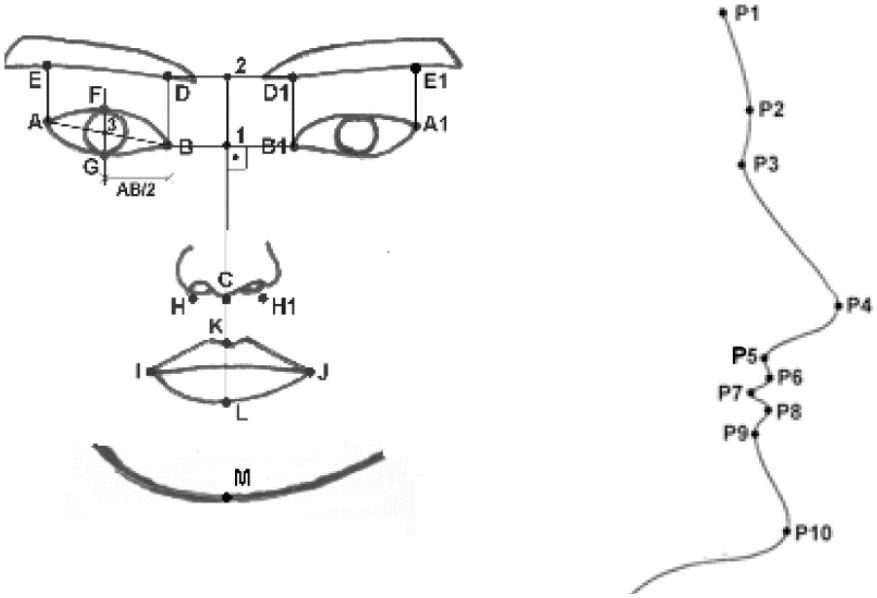


Figure 2.2: Facial points of frontal and profile view, from [22]

Yin [67] recognized the facial expression through topological analysis on the facial regions. The topographical structure of human face is analyzed based on the resolution-enhanced textures. The relationship between the facial expression and its topographic features is investigated such that facial expression can be represented by the topographic labels. Based on the observation that the facial texture and its topographic features change along with facial expressions, they compare the disparity of these features between the neutral face and the expressive face to distinguish a number of universal expressions.

Gokturk [28] utilized a deformable face model with 19 feature points. The shape vector is composed of the position of the feature points. They assume the whole shape vector is a linear combination of a small number of eigenvectors. The expression eigen-space is trained by image sequence with facial expressions from stereo camera. The facial expression in tracking stage is thus constraint.

Lee et al. [47] presented a method for modeling and recognizing human faces in video sequences. They use an appearance model composed of pose manifolds and a matrix of transition probabilities to connect them.

Zhou et al. [48] proposed a generic framework to track and recognize human faces simultaneously by adding an identity variable to the state vector in the sequential importance sampling method. The posterior probability of the identity variable is then estimated by marginalization. But their work, however, does not consider tracking and recognition of facial deformation. We were inspired by their work when we developed the probabilistic model on manifold.

2.2 Facial Expression Classification

Facial expressions can be classified in various ways – in terms of facial actions that cause an expression, in terms of some non-prototypic expressions such as "raised brows" or in terms of some prototypic expressions such as emotional expressions.

The Facial Action Coding System (FACS) [1] is probably the most known study on facial activity. FACS is designed for human observers to detect independent subtle changes in facial appearance caused by the facial muscles contractions. In a form of rules, FACS provides a linguistic description of all possible, visually detectable facial changes in terms of 44 so-called Action Units (AUs). Although the number of atomic action units is small, more than 7,000 combinations of action units have been observed. Fig. 2.3 shows the definitions of 7 individual upper face AUs and 5 combinations involving these action units. Using these rules, a trained human FACS coder decomposes a shown expression into the specific Aus that describe the expression.

Automating FACS would make it widely accessible as a research tool in the behavioral science, which is furthermore the theoretical basis of multimedia user interfaces. This triggered researchers of computer vision field to take different approaches in tackling the problem. Bartlett [64, 65] reported extensive experiments on upper and lower AU recognition and analyze spontaneous expression. Whitehill [66] investigated the correlation between different AUs and showed that this

correlation can impact recognition rate significantly because high correlation of some AUs happens naturally in prototypical expressions.

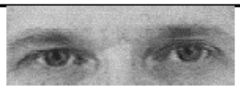
AU 1	AU 2	AU 4
		
Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together
AU 5	AU 6	AU 7
		
Upper eyelids are raised.	Cheeks are raised.	Lower eyelids are raised.
AU 1+4	AU 4+5	AU 1+2
		
Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.	Inner and outer portions of the brows are raised.
AU 1+2+4	AU1+2+5+6+7	AU0(neutral)
		
Brows are pulled together and upward.	Brow, eyelids, and cheek are raised.	Eyes, brow, and cheek are relaxed.

Figure 2.3: basic upper face AUs and AU combinations, from [27]

Most of the studies on vision-based facial expression analysis rely on Ekman's emotional classification of facial expressions. Ekman [2] defined six kinds of *basic*

emotions: happiness, sadness, surprise, fear, anger, and disgust. He indicated that the six emotions are universally associated with distinct facial expressions. Several other emotions, and many combinations of emotions have been studied but remain unconfirmed as universally distinguishable.

While the human mechanisms for facial expression interpretation are very robust, it is often very difficult for machine to determine the exact nature of the expression on a person's face. Independent of the used classification categories, the mechanism of classification applied by a particular surveyed expression analyzer is either a template-based- or a neural-network-based- or a rule-based- or statistical classification methods. We discuss the representing work using these four different methods respectively.

2.2.1. Template-based classification methods

Kimura [24] fit a Potential Net to each frame of the examined facial image sequence. The pattern of the deformed net is compared to the pattern extracted from an expressionless face and the variation in the position of the net nodes is used for further processing. They built an emotion space by applying PCA on six image sequences of three expressions - anger, happiness, and surprise–shown by a single person gradually, from expressionless to a maximum. The proposed method fails on the tests for image sequences of unknown subjects. A very small number of training examples (six sequences) and an insufficient diversity of the subjects (one person) have probably caused this.

Essa and Pentland [29] used the spatio-temporal motion-energy representation of facial motion for an observed expression. By learning "ideal" 2D motion views for each expression category, they generated the spatio-temporal templates for six different expressions–two facial actions (smile and raised eyebrows) and four emotional expressions (surprise, sadness, anger, and disgust). The Euclidean norm of the motion energy difference between the template and the observed image is used as a metric for measuring similarity. When tested on 52 frontal-view image sequences of eight subjects showing six distinct expressions, a correct recognition rate of 98% has been achieved.

2.2.2 Rule-based classification methods

Black and Yacoob [26] utilized local parameterized models of image motion to represent rigid head motions and nonrigid facial motions within the local facial areas. The motion parameters are used to derive the midlevel predicates that describe the motion of the facial features. Each midlevel predicate is represented in a form of a rule. In their method, the facial expression emotional classification considers the temporal consistency of the midlevel representation predicates. For each of six basic emotional expressions, they developed a model represented by a set of rules for detecting the beginning and ending of the expression. The method has been tested on 70 image sequences containing 145 expressions shown by 40 subjects ranged in ethnicity and age. The expressions were displayed one at the time. The achieved recognition rate was 88%. They did not give a full list of the midlevel predicates and the number of different facial actions that the method can recognize is not known.

Also, the method does not deal with blends of emotional expressions. The reason lies in the rules used for classification.

2.2.3 Neural-network-based classification methods

Hara [21] applied a 234x50x6 back-propagation neural network for classification of expression into one of six basic emotions. The units of the input layer correspond to the number of the brightness distribution data extracted from an input facial image while each unit of the output layer corresponds to one emotion category. The neural network has been trained on 90 images of six basic facial expressions shown by 15 subjects and it has been tested on a set of 90 facial expressions images shown by another 15 subjects. The average recognition rate was 85%.

Tian [27] used three-layer neural networks with one hidden layer to recognize AUs by a standard back-propagation method. Separate networks are used for the upper and lower face. These networks are trained to respond to the designated AUs whether they occur singly or in combination. When AUs occur in combination, multiple output nodes are excited. The recognition rate is 95.6% on the Cohn-Kanade Facial Expression Database [30]. We will compare our regional FACS system with Tian's algorithm in Chapter 5.

2.2.4 Statistical classification methods

Gokturk [28] classify dynamic expressions by using Support Vector Machine (SVM) [31] classification. To find a separate function that can be included from the training feature vectors from stereo tracking and generalizes well on the unknown

examples, they use SVM to find the optimal hyper-surface that not only correctly classifies the data, but also maximizes the margin of the closest data points to the hyper-surface. Having obtained the hyper-surface for each class, a test shape vector (coming from monocular face tracking) is classified. First, the location of the new data is determined with respect to each hyper-surface. For this, the learnt SVM for the particular hyper-surface is used to find the distance of the new data to that hyper-surface using the distance measure. The nearest hyper-surface represents the expression with the highest possibility. The average recognition rate is 90.8%.

In summary, capability of an automated facial expression system is related to the way of the facial expression data extraction and expression classification. All problems in this field are intriguing and none has been solved in the general case. We expect that they would remain interesting to the researchers of automated vision-based facial expression analysis for some time.

2.3 Nonlinear Dimensionality Reduction

Nonlinear dimensionality reduction has been an active research filed recently. Seung and Lee [32] suggested that an image with N pixels can be considered as a point in an N-dimensional image space, and the variability of image classes can be represented as low-dimensional manifolds embedded in image space as illustrated in Fig. 2.4.

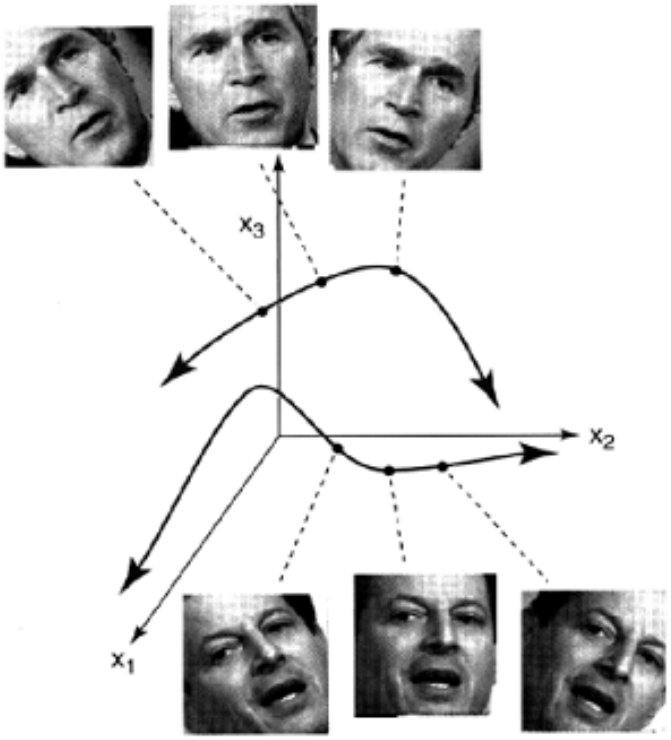


Figure 2.4: Manifolds in Visual Perception. From [7]

Tenenbaum et al. [33] introduced Isomap to find meaningful low-dimensional structures hidden in the high-dimensional data that is guaranteed to converge asymptotically to the true structure. It preserved the intrinsic geometry of the data by capturing the geodesic manifold distance between all pairs of data points. For neighboring points, input-space distance provides a good approximation to geodesic distance. For distant points, geodesic distance can be approximated by adding up a sequence of "short hops" between neighboring points. This shortest path can be computed efficiently by the Dijkstra Algorithm.

Roweis and Saul [34] showed that Locally Linear Embedding (LLE) is able to learn the global structure of nonlinear manifolds, such as those generated by images

of faces with only pose and illumination change. LLE is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embedding of high-dimensional inputs. LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima. In Lipschitz embedding [12,13], a coordinate space is defined such that each axis corresponds to a reference set $\boldsymbol{R}$, drawn from the input data set. With a suitable definition of $\boldsymbol{R}$, we can establish the bounds on the distortion for all pairs of data points in the embedding space.

Brand [70] constructed a nonlinear pseudo-invertible embedding method which effectively recovering a Cartesian coordinate system for the manifold. The objective functions are convex and their solutions are given in closed form. Chen et al. [68] also consider local data and present local discriminant embedding (LDE) for manifold learning and pattern classification. Yan et al. [69] propose a general framework: graph embedding along with its linearization and kernelization, which in theory reveals the underlying objective shared by most previous algorithms. Based on this mainframe, they develop a new supervised algorithm, Marginal Fisher Analysis (MFA) for dimensionality reduction by designing two graphs that characterize the intra-class compactness and interclass separability, respectively.

Elad and Kimmel [35] used the invariant signature of manifolds for object recognition. Lyons et al. [36] conducted a quantitative low dimensional analysis from image features for coding facial expressions. They used nonlinear non-metric multidimensional scaling of Gabor-labelled elastic graphs. Wang et al. [37]

demonstrated the importance of applying non-linear dimensionality reduction in the field of non-rigid object tracking. In fact, representing the object state as a globally coordinated low dimensional vector improves tracking efficiency and reduces local minimum problems in optimization. They learn the object's intrinsic structure in a low dimension manifold with density modeled by a mixture of factor analyzers.

Recently, researchers have applied manifold methods to face recognition [38, 39, 40] and facial expression representation [41,42,43]. We compare the performance of LLE with Lipschitz embedding in Chapter 3. The idea of geometrical distance in Isomap is applied to improve Lipschitz embedding in our system.

2.4 Facial Expression Databases

There are several publicly available facial expression databases: Cohen-Kanade facial expression database [30] provided by CMU has 97 subjects, 481 video sequences with six kinds of basic expressions. Subjects in every video began from a neutral expression, and ended at the expression apex. FACS coding of every video is also provided. The CMU PIE database [44] includes 41,368 face images of 68 people captured under 13 poses, 43 illuminations conditions, and with 3 different expressions: neutral, smile, and blinking. The Human ID database provided by USF has 100 exemplar 3D faces. The exemplar 3D faces were put in full correspondence as explained by Blanz and Vetter [45]. To our knowledge, there is no 3D expression database publicly available, so we built our own 3D database that includes 6 subjects

and 36 videos, with a total of 2581 frames. Every subject performed all six basic expressions from neutral to apex and back to neutral.

There is a relatively new MMI facial expression database [46] presented by Pantic. It includes more than 1500 samples of both static images and image sequences of faces in frontal and in profile view displaying various expressions of emotion, single and multiple facial muscle activation. It has been built as a web-based direct-manipulation application, allowing easy access and easy search of the available images. This database represents the most comprehensive reference set of images for studies on facial expression analysis to date.

The current facial expression databases are very essential to the progress in this field. They provide standard benchmark for testing and comparing different algorithms. But there is much room for further improvements.

First of all, none of these databases include spontaneous facial expression. The subjects were instructed to perform specific facial deformations which are more dramatic or faked than those appeared in normal life. A possible solution to this problem is to take videos from real-time show on TV or movies. But it is very hard to annotate these videos with large variation of face pose, illumination and to collect high quality dataset from multiple subjects. People speak and change facial expression simultaneously at most of the time. How to handle this problem is also a critical issue during collecting real facial expression data.

Second, with development in 3D morphable model [45], there has been a 3D face database [72]. But there is no 3D facial expression database when we explored this topic in [71]. It is noticed that Yin et al. [73] also published a paper on this recently.

It is a very time consuming and complicated task to build a facial expression database. There are also multiple levels of requirements for purposes of different applications. We are looking forward to the progress on this topic in the future.

Chapter 3

Manifold of facial expression

In this chapter, we will present the concept of manifold of facial expression. The manifold of expression is built through nonlinear dimensionality reduction, specifically, Lipschitz embedding. The expression manifolds of different subjects vary a lot even though they share similar structure. We develop generalized manifold by transferring the deformation of different subjects to a standard model. The training data from all subjects can complement each other in one generalized manifold.

3.1 Concept

The concept of manifold ways of preception has been proposed in Seung and Lee [7]. Since people change facial expression continuously over time, it is a reasonable assumption that all images of someone's facial expressions make a smooth manifold in the *N*-dimensional image space with the "neutral" face as the central reference point. The intrinsic dimension of the manifold is much lower than *N*. If we were to allow other factors of image variation, such as face pose and illumination, the intrinsic dimensionality of the manifold of expression would increase accordingly.

On the manifold of expression, similar expressions are points in the local neighborhood on the manifold. The basic emotional expressions with increasing intensity become curves on the manifold extended from the center. The blends of

expressions will lie between those curves, so they can be defined analytically by the positions of the main curves. The analysis of the relationships between different facial expressions will be facilitated on the manifold. More generally, the manifold of facial expression shows promise as a unified framework for facial expression analysis.

3.2 Nonlinear Dimensionality Reduction

It is a formidable task to learn the structure of the manifold of expression in a high dimensional image space. The first step is to extract the features from face images accurately and robustly.

We first apply Active Shape Model on the image sequences to reduce the variation due to scaling, illumination condition, and face pose. We seek to embed the manifold from the high dimensional feature space of AWN to a low dimensional space while keeping the main structure of the manifold. In this paper, we investigate two types of embeddings to perform this task. The first is locally linear embedding (LLE). The second is Lipschitz embedding.

3.2.1 Locally Linear Embedding

LLE is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embedding of high-dimensional inputs. LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima.

Previous approaches to this problem, based on multidimensional scaling (MDS) [49], have computed embeddings that attempt to preserve pairwise distances between data points. Some methods measure these distances in p metric space, while other methods measure these distances along the shortest paths confined to the manifold of observed inputs, such as Isomap [33]. However, LLE recovers global nonlinear structure from locally linear fits. It is based on the following geometric intuitions. If the data points are drawn from some underlying manifold uniformly, each data point and its neighbors are expected to lie on or close to a locally linear patch of the manifold. The overlapping local neighborhoods, collectively analyzed, can provide information about global geometry.

The algorithm of LLE is presented in Fig. 3.1. LLE aims to minimize the locally linear reconstruction error for every data point in the embedding space. The local geometry of the patches around every data point is characterized as the weight matrix $W_{N,N}$. It can be considered as a least squares problem to compute W subject to the constraint $\sum_{j=1}^{N} w_{i,j} = 1$. This least squares problem can be solved by the Lagrange multiplier algorithm. The low dimensional vectors Y_i represent the global internal coordinates on the manifold. Y_i are chosen to minimize the embedding cost function $\Phi(Y) = \sum_i \left| \vec{Y}_i - \sum_j w_{i,j} \vec{Y}_j \right|^2$ with fixed weight W. By the Rayleigh-Ritz theorem [50], Y_i are the smallest d eigenvectors of matrix $M = (I - W)^T (I - W)$ after discarding the bottom eigenvector.

LLE is able to learn the global structure of nonlinear manifolds. For data visualization in two and three dimensions, it works well when the data set has only one cluster. This assumption holds when the image sequences of facial expressions are from the same subject. When the data come from multiple subjects, there are many manifolds with different centers (neutral face) and stretching directions. We need to build "global coordinates" for the mixture of locally linear projection from samples to coordinate space [51], or decompose the sample data into some patches, then merge them into one global coordinates in an optimal way [52].

Input:

$\vec{X}_i, i = 1,..., N : \vec{X}_i \in R^D$

k: the number of nearest neighbors for every data point

Output:

$\vec{Y}_i, i = 1,..., N : \vec{Y}_i \in R^d$

for i=1 **to** N **do**

$[\vec{X}_{t_1},..., \vec{X}_{t_k}] = \vec{X}_i$'s k nearest neighbors .

$[w_{i,t_1},..., w_{i,t_k}] = \arg\min \sum \left| \vec{X}_i - \sum_{j=1}^{k} w_{i,t_j} \vec{X}_{t_j} \right|^2$ subject to $\sum_{j=1}^{k} w_{i,t_j} = 1$

$w_{i,j} = 0$ when $\vec{X}_j \notin [\vec{X}_{t_1},..., \vec{X}_{t_k}]$

end for

$[\vec{Y}_1,..., \vec{Y}_N] = \arg\min \sum_i \left| \vec{Y}_i - \sum_j w_{i,j} \vec{Y}_j \right|^2$

Return $\vec{Y}_i, i = 1,..., N$

Figure 3.1: Locally Linear Embedding algorithm

In our experiments, subjects were instructed to perform a series of seven kinds of facial expressions: happiness with closed mouth, happiness with open mouth, sadness, anger, surprise, fear, disgust. (The fact that these are not true emotion-driven expressions is not relevant to the analysis.) The subjects repeated the series seven times. A total of 4851 images were captured (1824 frames for the female subject, 3027 frames for the male subject). We used frontal view in this experiment because the facial feature is extracted from 2D template. If the facial features can be tracked through a 3D model, there will be fewer constraints on the input video, such as frontal view. The location of the feature points is more robust compared to the raw pixel data because it is less affected by the illumination variation.

We found that LLE is very sensitive to the selection of the number of nearest neighbors. When the data set contains just one series of seven kinds of facial expressions, every image sequence is mapped to a curve that begins from the center (neutral face) and extends in distinctive direction with varying intensity of expression as in Fig. 3.2. While there are many series, the sequences with the same kind of facial expression diverge in different directions when the number of nearest neighbors is small. The images of different expressions become mixed up easily when we increase the number of nearest neighbors as shown in Fig. 3.3.

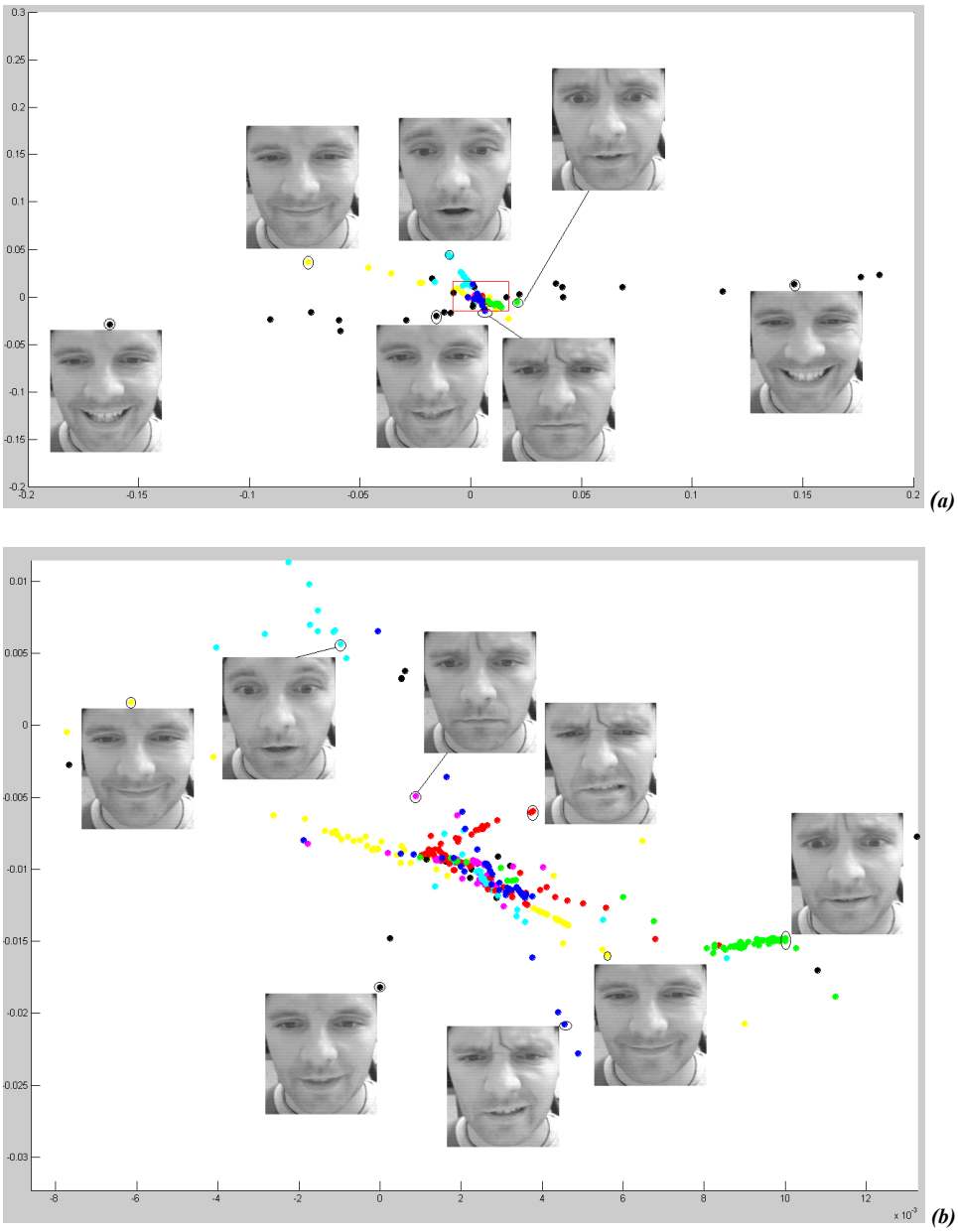


Figure 3.2: The first 2 coordinates of LLE of 478 images with the number of nearest neighbors k=9. (b) shows the enlarged red rectangular region in (a). Seven sequences for seven different expressions, which are represented by different colors: small smile: yellow; big smile: black; sadness: magenta; anger: cyan; disgust: red; surprise: green; fear: blue. The representative images are shown next to the circled points.

The reason is that LLE is an unsupervised learning algorithm. It selects the nearest neighbors to reconstruct the manifold in the low dimensional space. There are two types of variations in the data set: the different kinds of facial expressions and the varying intensity for every kind of facial expression. Generally, LLE can catch the second type of variation – an image sequence is mapped in a "line," and LLE can keep the sequences with different expressions distinctive when there is only one sequence for each expression. When the data set contains many image sequences for the same kind of expression, it is very hard to catch the first kind of variation using a small number of nearest neighbors. But with the increased number of nearest neighbors, the images of different expressions are more prone to be mixed up.

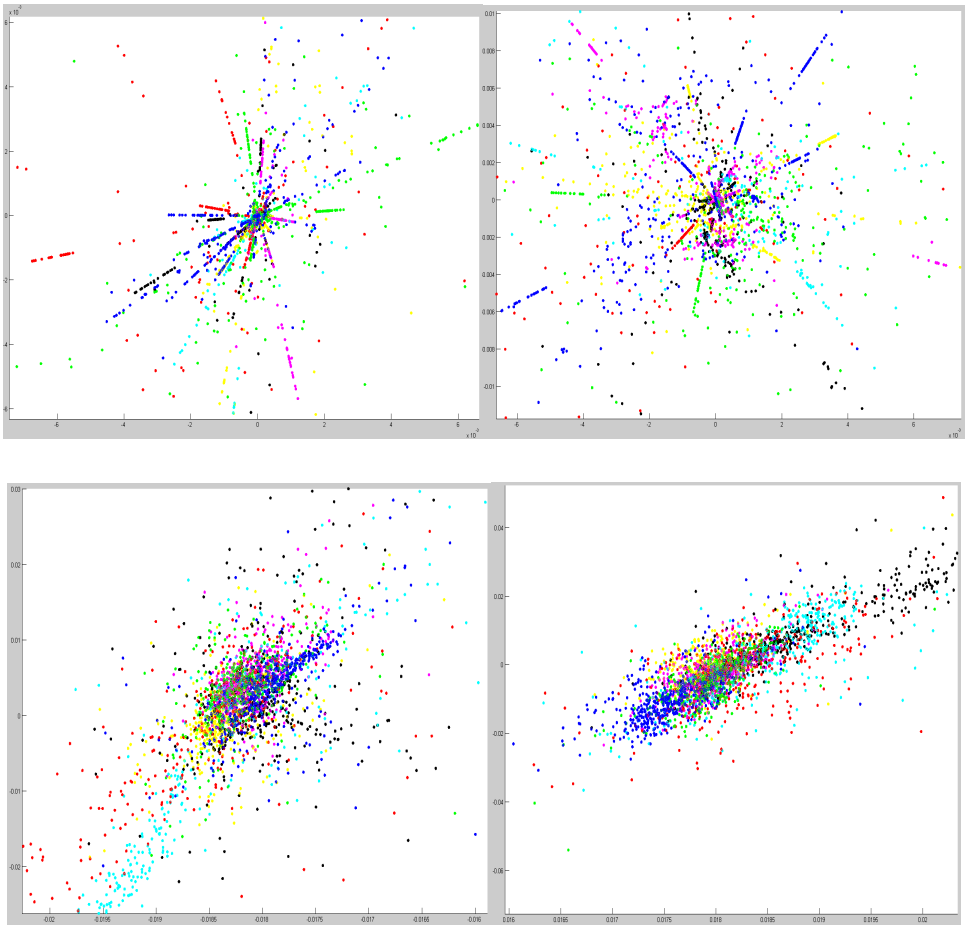


Figure 3.3: The first 2 coordinates of LLE of 3027 images (the male subject) with the number of nearest neighbors k=10,20,30,50, from left to right, up to down. The meaning of colors is the same as in Fig. 3.2.

We classify expressions by applying a k-Nearest Neighbor method in the embedding space. The result is in Table 1. It can be seen that LLE cannot achieve good expression classification results without building "global coordinates" for the mixture of locally linear projection [51], or performing some decomposing/merging processes [52].

Table 1: Comparison of classification rate on the four manifolds in Fig. 3.3.

	LLE(k=10)	LLE(k=20)	LLE(k=30)	LLE(k=50)
k_NN(k=9)	15.76%	16.60%	13.53%	16.32%
k_NN(k=13)	15.62%	17.57%	15.90%	16.47%

3.2.2 Lipschitz Embedding

Lipschitz embedding is a powerful embedding method used widely in image clustering and image search. For a finite set of input data S, Lipschitz embedding is defined in terms of a set R of subsets of S, $R = \{A_1, A_2, ..., A_k\}$. The subsets A_i are termed the reference sets of the embedding. Let $d(o; A)$ be an extension of the distance function d to a subset $A \subset S$, such that $d(o, A) = \min_{x \in A}\{d(o, x)\}$. An embedding with respect to R is defined as a mapping F such that $F(o) = (d(o; A_1); d(o; A_2); ..., d(o; A_k))$. In other words, Lipschitz embedding defines a coordinate space where each axis corresponds to a subset $A_i \subset S$ of the objects, and the coordinate values of object o are the distances from o to the closest element in each of A_i. For example, if there are only three kinds of facial expressions, the apex of each kind of expression will be mapped on the positive central part of the x-y, x-z, y-z planes, and one kind of expression with different intensity or blended expressions will be mapped on an approximately spherical surface, as illustrated in Figure 3.4. Because people can exhibit many more the three facial expressions, the

manifold of facial expression will become a super-spherical surface in k-dimensional space through Lipschitz embedding.

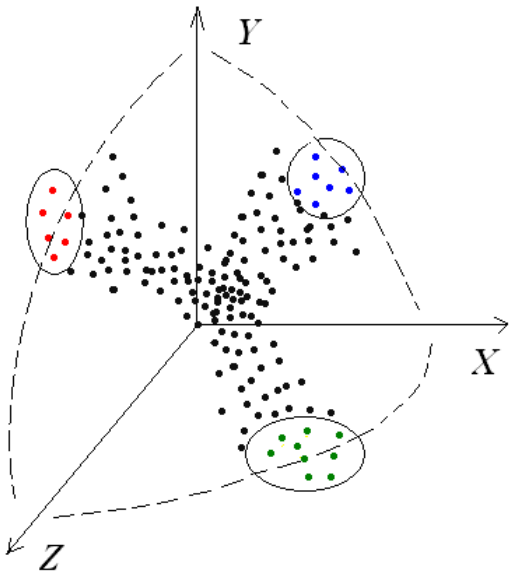


Figure 3.4: An illustration of Lipschitz embedding for k=3. The points in the circles are in the reference set which represent extreme expressions. The neutral faces are far away from every reference set.

With a suitable definition of the reference set R, the distance of all pairs of data points in the embedding space is bounded [53]. So Lipschitz embedding works well when there are multiple clusters in the input data set [14,15]. In our algorithm, we preserve the intrinsic structure of the expression manifold by combining Lipschitz embedding and the main feature of Isomap [33]. Given a video gallery covering six basic facial expressions, there are six "paths" from the neutral image to the six sets of images with the basic expressions at apex on the manifold. In Figure 1, the apex sets in 3D space are illustrated as the points within the circles. Each path is

composed of many small steps (the difference between consecutive frames). Different paths contain information on how the expressions evolve. The comparative positions between those paths correspond to the relationship between different expressions.

The distance function in Lipschitz embedding reflects the distance between points on the manifold. The crucial property that we aim to retain is proximity; i.e., which points are close to each other and which are far from each other. Due to the essential nonlinear structure of the expression manifold, the classical approaches of multidimensional scaling (MDS) [49] and PCA fail to detect the true degrees of freedom of the face data set.

For our experiments, we used six reference sets, each of which contains images of only one kind of basic facial expression at its apex. The embedded space is six dimensional. The distance function is the geodesic manifold distance. After we apply the modified Lipschitz embedding to the gallery set, there are six basic paths in the embedded space, emanating from the center that corresponds to the neutral image. The images with blended expression lie between the basic paths. For every sequence, only one image during the apex of expression is selected for the corresponding reference set. Every image is mapped to a six dimensional vector, which represents its distance to each kind of "extreme" expression. For the purpose of visualization, we can project the manifold onto its first three dimensions as shown in Fig. 3.5. One can see that the expression manifold can be considered

approximately as a spherical surface. In the embedded space, expressions can be recognized by using the probabilistic model described in Chapter 4.

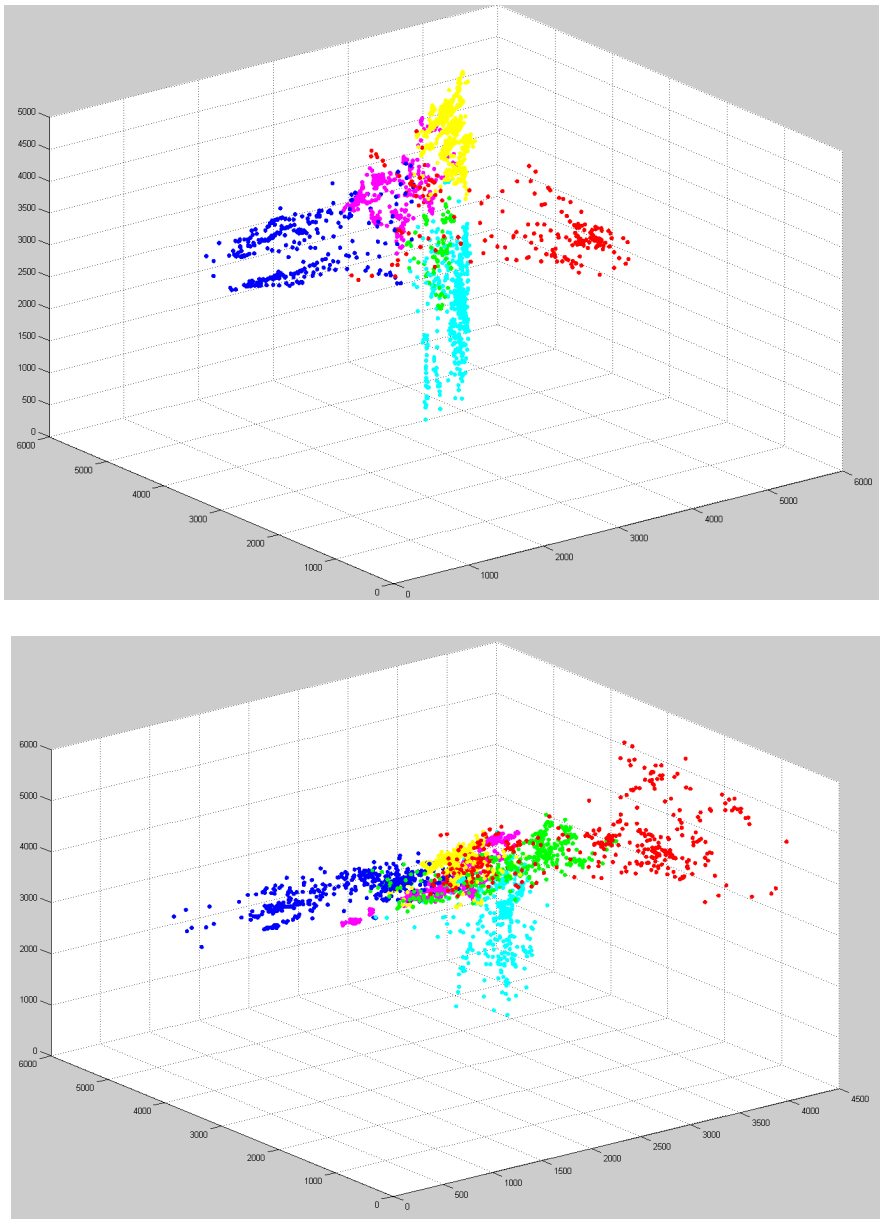


Figure 3.5: The projection on first three dimensions after Lipschitz embedding. (a) the female subject. (b) the male subject. The meaning of colors is the same as in Fig. 3.2.

3.3. Generalized manifold

The data points of a single subject make a continuous manifold in the embedding space, while the neutral face and the intensity of expression that can be displayed by different subjects vary significantly. Correspondingly, the manifolds of different subjects vary a lot in the centers, the stretching directions, and the covered regions. Because of the appearance variation across the subjects, it is very hard to align the manifolds of different subjects in a way that the images from different subjects with semantic similarity can be mapped to the near region. We can align the manifolds of different subjects directly as in Section 3.3.1, or transfer the deformation on different subjects to one standard model and build a generalized manifold as in Section 3.3.2.

3.3.1 Manifold Alignment

We learn the manifolds of expression from all subjects in the training data. To make the training data usable to every subject, the manifolds of different subjects need to be aligned because they share similar structure. In the space of Lipschitz embedding, the alignment of manifolds from different subjects can be performed in an elegant way. The k reference sets contain typical expressions. The *i*th coordinate of data points in reference set A_i is zero by definition. So the images in the reference set A_i are mapped to a compact plane region R_i. The images that represent a kind of expression from beginning to apex are mapped along the curves from the center (neutral expression) to R_i on the expression manifold. To align the manifolds of different subjects, we need only to align the corresponding region R_i of the different

manifolds to the set region one by one. The center of region R_i is $Q_i = (q_1^i, q_2^i, ..., 0_i^i, ..., q_k^i)^T$. The essential idea is to align Q_i to $(1_1, ..., 1_{i-1}, 0_i, 1_{i+1}, ..., 1_k)^T$. The manifold of expression will become a super-spherical surface with approximate radius $\sqrt{k-1}$.

A straightforward solution is to apply a linear transformation to the embedding space. A linear transformation in the k-dimensional space can be determined by the transformation of k points, i.e. the centers of the k reference regions. Unfortunately, linear transformation does not preserve the semantic similarity of data points well. Our experiments show that images of the different expressions get mixed up after such a linear transformation.

We propose a nonlinear transformation that can align the reference regions while preserving the semantic similarity of data points at the same time. The detailed algorithm is presented in Figure 3.6. The goal is to align all nonzero coordinate values of Q_i to 1. To align q_j^i of Q_i to 1, the jth member of each data point is multiplied by a scaling factor. Therefore the critical part is to design a nonlinear continuous function Φ that returns a scaling factor as 1 for the points in all other reference sets except A_i, near $1/q_j^i$ for all points in A_i, and an interpolated number according to their positions for the remaining data points. The nonlinear function Φ is defined in Figure 3.7. After every process, one nonzero coordinate value of a region center is normalized to 1 and the structure of the manifold is preserved.

When the semantic meanings of every reference set are defined in the same way for different subjects, the corresponding reference set will be aligned to the same

region by our algorithm. So the aligned manifold will map the images with semantic similarity but from different subjects in the near region.

Input:

$S = \{x_1, x_2, ..., x_n\}$, $x_i \in R^D$

Output:

$Y = \{y_1, y_2, ..., y_n\}$, $y_i \in R^k$

Select k subsets $A_1, A_2, ..., A_k$ from S.

Apply Lipschitz embedding algorithm. The result is $Y = \{y_1, y_2, ..., y_n\}$, $y_i = (y_1^i, y_2^i, ..., y_k^i)$

for i=1 **to** k **do**

  $Q_i = mean\ (y_j, if\ \ x_j \in A_i)$

end for

$Q = \{Q_1, Q_2, ..., Q_k\}$, $Q_i = (q_1^i, q_2^i, ..., 0_i^i, ..., q_k^i)$.

for i=1 **to** k **do**

  for j=1 **to** k **do**

    if $j \neq i$

      for s=1 **to** n **do**

        $scale = \Phi(y_s, Q_i, j)$;

        $y_j^s = y_j^s * scale$;

      end for

      update Q

    end if

  end for

end for

Return $Y = \{y_1, y_2, ..., y_n\}$

Figure 3.6: The alignment algorithm

Function $\Phi(y_s, Q_i, j)$

$\{Q_i = (q_1^i, q_2^i, ..., 0_i^i, ..., q_k^i), y_s = (y_1^s, y_2^s, ..., y_k^s)\}$
$b = 1;$
$v = 1;$
for t=1 **to** k **do**
  if $(t \neq i) \& (t \neq j)$
    $b = b * q_t^i;$
    $v = v * y_t^s;$
  end if
end
{if $y_s \in R_p, p \neq i$, then $v = 0, scale = 1;$
 if $y_s \in R_i$, then $v \cong b, scale \cong 1 / q_j^i$}
$scale = (v * (1 - 1 / q_j^i) / b + 1) * \exp(y_j^s / c);$
{c is an empirical constant}
Return scale

Figure 3.7: Function $\Phi(y_s, Q_i, j)$

The alignment of the two manifolds in Fig. 3.5 by nonlinear alignment is shown in Fig. 3.8. The alignment by linear alignment is shown in Fig. 3.9. We can see the clusters of different expressions are preserved well through nonlinear alignment, while images of different expressions become mixed up after linear alignment.

We apply a k-Nearest Neighbor method to classify expressions for all 4851 images on the aligned manifold of facial expression. The comparison of the classification rate between the generalized manifold drawn by nonlinear alignment and linear alignment is presented in Table 2. The experimental results further demonstrate the effectiveness of nonlinear alignment.

Table 2: Comparison of recognition rate on one generalized manifold after linear/nonlinear alignment.

	Manifold by linear alignment	Manifold by nonlinear alignment
k_NN(k=9)	94.42%	96.09%
k_NN(k=13)	94.60%	96.37%

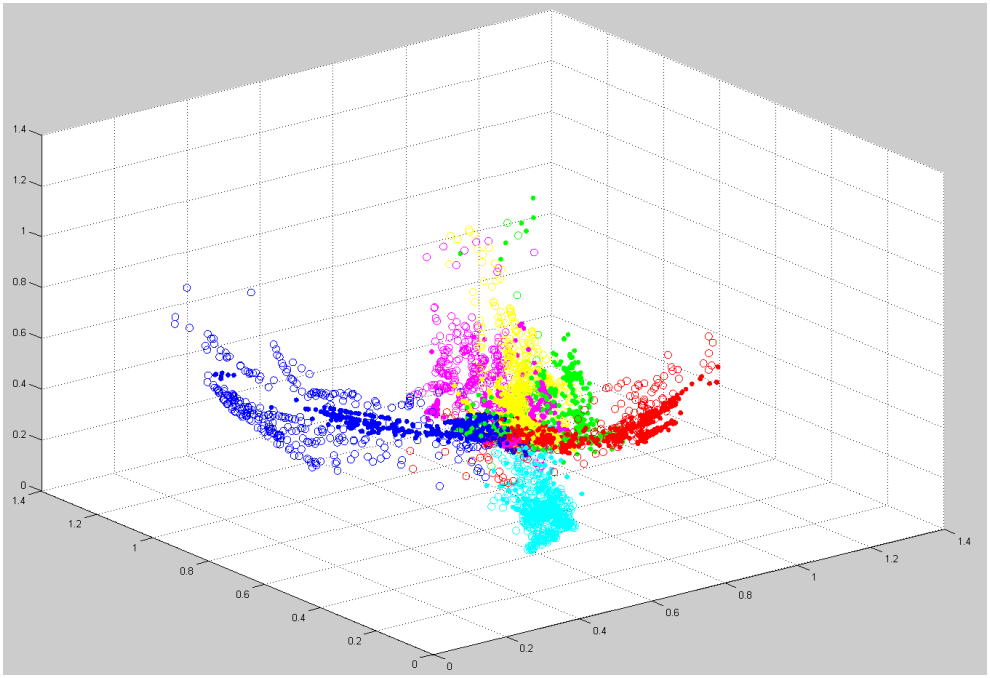


Figure 3.8: The aligned manifolds after nonlinear alignment. The points from the first manifold are represented as circles. The meaning of colors is the same as in Fig. 3.2.

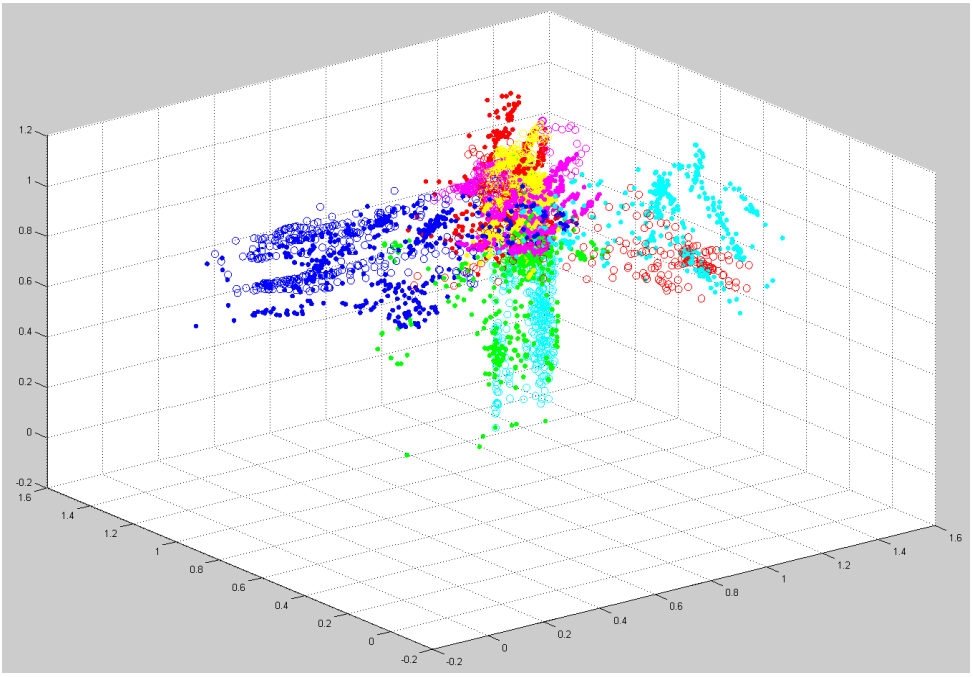


Figure 3.9: The aligned manifolds after linear alignment. The points from the first manifold are represented as circles. The meaning of colors is the same as in Fig. 3.2.

3.3.2 Deformation Transfer for Generalized Manifold

Build a generalized expression manifold by transferring deformation to a standard model has many advantages. The standard model serves as the interface between the models in the training videos and models in the testing videos. The generalized manifold, that is the expression manifold of the standard model, includes all information in the training videos. The more training data we have, the better it approximates the true manifold. We can define expression similarity on this manifold and use it to search the optimal approximation for any kind of expression. The expression synthesis will involve only the standard model and target model.

We can fit a 3D mesh model given a 2D image [45]. Sumner [11] proposed a novel method to transfer the deformation of the source triangle mesh to the target one by minimizing the transformation between the matching triangles while enforcing the connectivity. This optimization problem can be rewritten in linear equations:

$$\min_{v_1 \ldots v_n} \| c - Ax \|_2^2 \tag{3.1}$$

where the matrix norm $\| \bullet \|_F$ is the Frobenius norm, or the square root of the sum of the square matrix elements. $v_1,...,v_n$ is the vertex of the unknown deformed target mesh. x is a vector of the locations of $v_1,...,v_n$. c is a vector containing entries from the source transformations, and A is a large sparse matrix that relates x to c, which is determined by the undeformed target mesh. This classic least-square optimization problem has closed form solution as

$$Sx = b \text{, where } S = A'A, b = A'c. \tag{3.2}$$

The result is unique up to a global translation. We fix the rigid vertex, such as inner eyes corners to resolve the global position. x can be split as $x = [xf' \quad xm']'$ where xf corresponds to the fixed vertex, and xm to all the other vertices. Thus

$$\begin{aligned} c - Ax &= c - \begin{bmatrix} Af & Am \end{bmatrix} * \begin{bmatrix} xf \\ xm \end{bmatrix} \\ &= c - Af * xf - Am * xm = d - Am * xm \end{aligned}$$

Our goal is to transfer the deformation of a training subject in a video sequence to a standard face smoothly. The vertex v_i at frame t is represented as

$v_i^t, i = 1,...,n; t = 1,...,k.$ k is the length of the video. We add a constraint for temporal coherence and the optimization problem becomes

$$\min_{v_1^1 \ldots v_n^1, \ldots, v_1^k \ldots v_n^k} \sum_{t=1,\ldots,k} \| d^t - Am * xm^t \|_2^2 + \sigma \| \frac{\partial xm^t}{\partial t} \|_2^2 \tag{3.3}$$

where σ is the weight for temporal smoothing. c^t is the source transformation at frame t, $d^t = c^t - Af * xf$.

This problem can be solved in a progressive way by approximating

$$\frac{\partial xm^t}{\partial t} = xm^t - xm^{t-1},$$

where xm^0 is the vertex locations of the undeformed target mesh.

Eq. (3-3) can be rewritten as

$$\min_{v_1^1 \ldots v_n^1, \ldots, v_1^k \ldots v_n^k} \sum_{t=1,\ldots,k} \| Q * xm^t - p^t \|_2^2 \tag{3.4}$$

where

$$Q'Q = Am'*Am + \sigma I$$
$$Q'p^t = Am'*d^t + \sigma * xm^{t-1}$$

σ is chosen to guarantee $Am'*Am + \sigma I$ is symmetric positive matrix. Q always exists, while it is not needed to solve Q explicitly. Eq. (3.4) has closed solution as $Q'Q * xm^t = Q'p^t$. For efficiency, we compute and store the LU factorization of $Q'Q$ only once.

We separate the motion of the tracked source mesh into a global transformation due to head movement and a local deformation due to facial expression. The local deformation is used for facial expression (deformation) transfer.

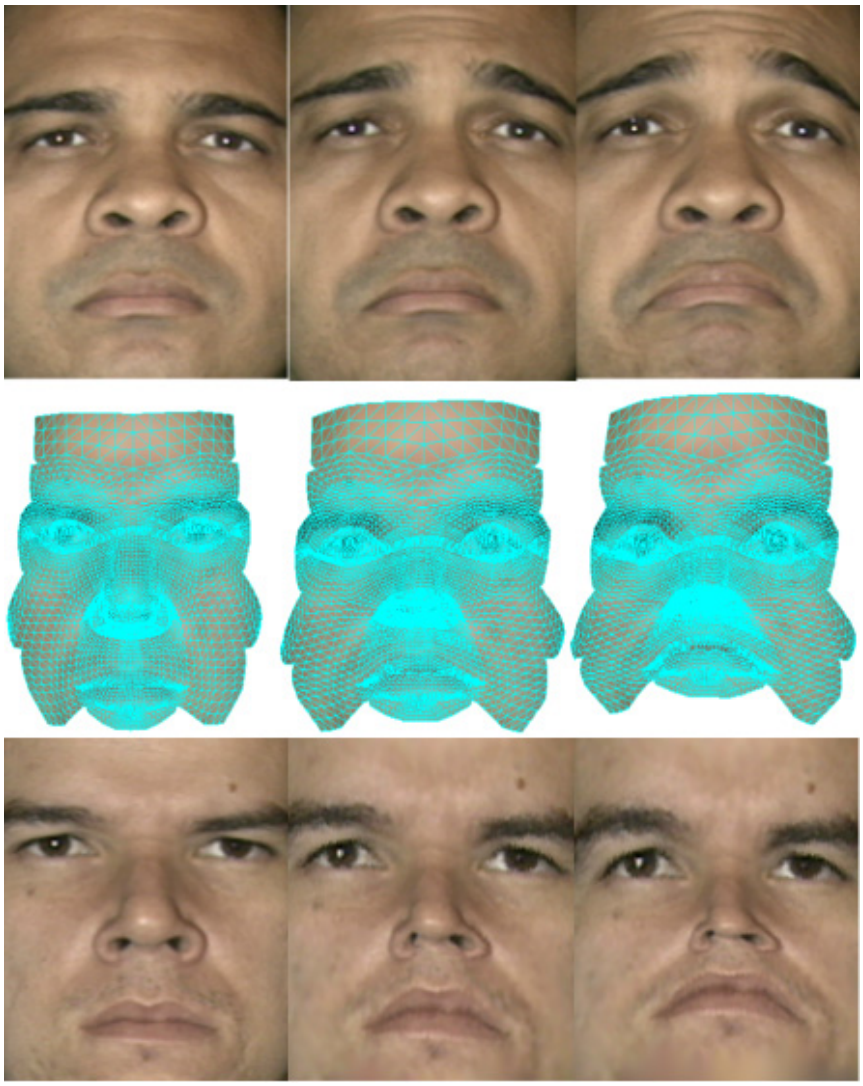


Figure 3.10: Deformation transfer with texture synthesis

Fig. 3.10 shows an example of transferring the source mesh to the target mesh with synthesized texture data. The first row is the texture image of the source video at frames 1, 12, 24. The second row is the dense mesh of the target face with transferred deformation. The first image of the third row is the texture image of the undeformed target model. The second and the third images are the corresponding synthesized faces by the deformed dense mesh.

Fig. 3.11 shows another examples of deformation transfer. The first row and second row is images of anger and the corresponding deformed standard mesh model. The first to the third column is one style of anger at frame 1, 6, and 29. The fourth to sixth column is another style of anger at frames 1, 20, and 48. The motions of the training videos are well retargeted on the standard model.

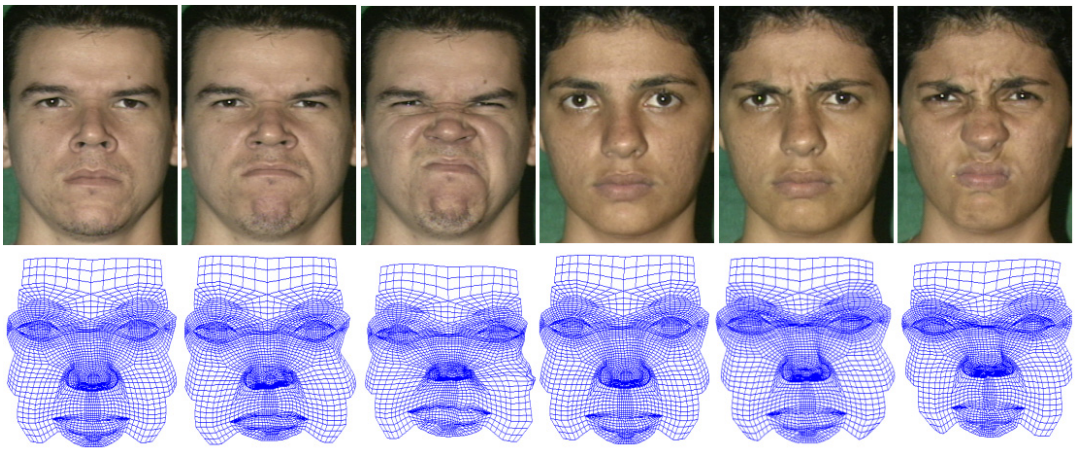


Figure 3.11: Deformation transfer from training videos

Chapter 4

Probabilistic Model

The goal of the probabilistic model is to exploit the temporal information in video sequences in order to recognize expression correctly and find the optimal replacement for expression editing.

4.1 Learning Dynamic Facial Deformation

We are interested in embedding the facial deformations of a person in a very low dimensional space, which reflects the intrinsic structure of facial expressions. From training video sequences of different people undergoing different expressions, a low dimensional manifold is learned, with a subsequent probabilistic modeling used for tracking and recognition. The goal of the probabilistic model is to exploit the temporal information in video sequences. Expression recognition is performed on the manifold constructed for each individual.

We use the Active Shape Model algorithm to detect a set of 2D facial landmarks in each image, which defines the shape of a face in each particular frame and reduces the influence of illumination variation. With a good manual initialization and separate training models prepared specifically for each expression image set, we can extract the face shape precisely. Figure 4.1 shows the facial points in our shape model. The detailed facial deformation such as wrinkles and dimpling are ignored. But the positions of the feature points still provide plenty of information to recognize

expression correctly based on our experiments. We expect better recognition results for a facial model with higher spatial resolution when more details of facial deformation can be captured.

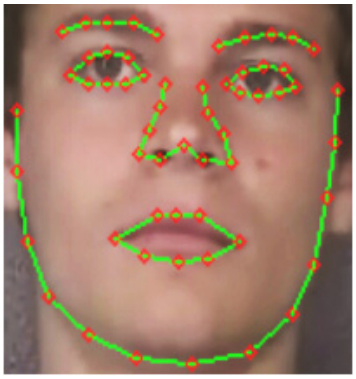


Figure 4.1: The shape model, defined by 58 facial landmarks.

The whole training dataset, comprising different video sequences of different people undergoing different facial expressions, is then specified by a set $X = \{x_1, ..., x_n\}$, where $x_i \in R^{2D}$ denotes a set of D facial points in a particular frame, and n denotes the total number of images in the training data. Unlike traditional manifold embedding methods, where data can be in any temporal order, our training images are temporally ordered according to the video sequences, thus allowing the learning of dynamics on the manifold, as we will show later.

To embed the high dimension data set $X = \{x_1, ..., x_n\}$ with $x_i \in R^{2D}$ to a space with low dimension d < 2D, we use our modified Lipschitz embedding algorithm, as described in the previous section. Our goal is to find the latent variable

$Y = \{y_1, y_2, ...y_n\}$, where $y_i \in R^d$. This latent variable encodes the knowledge of the data set and controls the data variations.

In the lower dimensional embedded space, we describe the distribution of the data using a Gaussian Mixture Model (GMM). The Expectation Maximization (EM) algorithm is used to estimate the distribution. The following equation describes the density model, where $p(y)$ is the probability that a point in the low dimensional space is generated by the model, k is the number of Gaussians, $p(\omega = i)$ constitutes the mixture coefficients and $N(\mu_i, C_i)$ describes each Gaussian distribution with mean μ_i and covariance matrix C_i:

$$p(y) = \sum_{i=1}^{K} p(\omega = i) N(\mu_i, C_i) \tag{4.1}$$

If we were to train an Active Shape Model from all the images in training data set together, the significant variation in the data set would not be modeled well and the tracking performance would be poor. Instead, we train a set of ASM models for each image cluster; that is, for each set of images corresponding to a mixture center (with a defined covariance) of the GMM in the embedded space.

We also propose a method to select and probabilistically integrate the ASM models in the ICondensation framework. We will show in Section 4.2 that online model selection allows tracking to be robust under large expression variations.

In ASM, a shape vector S is represented in the space spanned by a set of eigenvectors learned from the training data. As a result, S may be expressed as:

$$S = \overline{S} + Us \tag{4.2}$$

where $\overline{S}$ is the mean shape, U is the matrix consisting of eigenvectors and s constitutes the shape parameters, which are estimated during ASM search.

Based on the manifold representation, we can learn a dynamic model, defined as the transition probability $p(y_t \mid y_{t-1})$. Let $\omega \in \{1,..., k\}$ be a discrete random variable denoting the cluster center and let $r \in \{1,..., n_r\}$ be a discrete random variable denoting the expression class. For this work, n_r = 6, meaning that r can assume six basic expressions. We have been using the prototypical universal expressions of fear, disgust, happiness, sadness, anger and surprise, though the method does not depend on this particular grouping.

The dynamic model can be factorized in the following way:

$$\begin{aligned} p(y_t \mid y_{t-1}) &= \sum_{w_t} p(y_t \mid y_{t-1}, \omega_t) p(\omega_t \mid y_{t-1}) \\ &= \sum_{\omega_t, \omega_{t-1}} p(y_t \mid y_{t-1}, \omega_t) p(\omega_t \mid \omega_{t-1}) p(\omega_{t-1} \mid y_{t-1}) \end{aligned} \tag{4.3}$$

where $p(\omega_t \mid \omega_{t-1}) = \sum_{r_{t-1}} p(\omega_t \mid \omega_{t-1}, r_{t-1}) p(r_{t-1})$

This assumes that ω_t and y_{t-1} are conditionally independent given ω_{t-1}.

For each state of r_{t-1} (i.e., each expression class), the cluster transition dynamics $P(\omega_t \mid \omega_{t-1}, r_{t-1})$ can be learned from the training data. $P(y_t \mid y_{t-1}, \omega_t)$ is the dynamic model for a known cluster center. The dynamics in a fixed cluster are similar for each expression. Since the intra-cluster variations are much smaller than the inter-cluster variations, we approximate the truth by assuming the dynamics in a fixed cluster is the same for each expression. If each cluster contains only one point, the difference between our approximation and the truth becomes zero.

Similar to Wang et al. [37], we also model the within cluster transition as a first order Gaussian Auto-Regressive Process (ARP) by:

$$p(y_t \mid y_{t-1}, \omega_t) = N(A_{\omega_t} y_{t-1} + D_{\omega_t}, BB^T) \tag{4.4}$$

which can be represented in generative form as

$$y_t = A_{\omega_t} y_{t-1} + D_{\omega_t} + Bw_k \tag{4.5}$$

where A_{ω_t} and D_{ω_t} are the deterministic parameters of the process, BB^T is the covariance matrix, and w_k is independent random white noise.

For AR parameter learning, we use the same method as Blake and Isard [54]. Combining equations (4.3), (4.4) and (4.5), we get:

$$\begin{aligned} &p(y_t \mid y_{t-1}) = \\ &\sum_{\omega_t \omega_{t-1} r_{t-1}} p(y_t \mid y_{t-1}, \omega_t) p(\omega_t \mid \omega_{t-1}, r_{t-1}) p(r_{t-1}) p(\omega_{t-1} \mid y_{t-1}) \\ &= \sum_{\omega_t} N(A_{\omega_t} y_{t-1} + D_{\omega_t}, BB^T) \alpha(\omega_t; y_{t-1}) \end{aligned} \tag{4.6}$$

$$\text{where } \alpha(\omega_t; y_{t-1}) = \sum_{\omega_{t-1}, r_{t-1}} P(\omega_t \mid \omega_{t-1}, r_{t-1}) P(r_{t-1}) P(\omega_{t-1} \mid y_{t-1}) \tag{4.7}$$

Wang et al. [37] pointed out that the equations above model a Mixture of Gaussian Diffusion (MGD), whose mixture term is controlled by the random variable ω_t. In our work, the mixture term is also controlled by the expression recognition random variable.

4.2. Probabilistic Tracking and Recognition

In the previous section, we showed how to learn a facial expression model on the manifold as well as its associated dynamics. Now, we show how to use this representation to achieve robust online facial deformation tracking and recognition. Our probabilistic tracking is based on the ICondensation algorithm [55], which is described next, followed by expression classification. Both tracking and recognition are described in the same probabilistic framework, which enables them to be carried out in a cooperative manner.

4.2.1 ICondensation Tracking

Our object state is composed of rigid and non-rigid parts, defined by $s=(x,y,\theta,sc;y_1...y_d)$. The rigid part (x,y,θ,sc) represents the rigid face motion (position, orientation and scale), while the non-rigid part ($y_1...y_d$) is the low dimensional representation of facial deformation obtained by our modified Lipschitz embedding.

At time *t*, the conditional object state density is represented as a weighted set of samples $\{(s_t^{(n)},\pi_t^{(n)}),n=1,...N\}$, where $s_t^{(n)}$ is a discrete sample with associated weight $\pi_t^{(n)}$, where $\sum_n \pi_t^{(n)}=1$. Below we illustrate one step of a sample's evolution.

Sequential Importance Sampling Iteration:

Main Objective: Generate sample set $S_t = \{(s_t^{(n)}, \pi_t^{(n)}), n = 1,... N\}$ at time t from sample set $S_{t-1} = \{(s_{t-1}^{(n)}, \pi_{t-1}^{(n)}), n = 1,... N\}$ at time *t-1*.

Algorithm:

For each sample, n = 1 to *N*:

1) Create samples $\tilde{s}_t^n$

Choose one of the following sampling methods with a fixed probability:

(1) Generate sample from initialization prior.

(2) Generate sample from importance re-sampling, where the importance function is the posterior from time t-1;

2) Predict s_t^n from $\tilde{s}_t^n$

a) If $\tilde{s}_t^n$ was generated from the prior probability, choose s_t^n from $\tilde{s}_t^n$ adding a fixed Gaussian noise.

b) If $\tilde{s}_t^n$ was generated from the posterior probability, apply the dynamic model for prediction. For the rigid state part, we use constant prediction, adding a small Gaussian noise. For the non-rigid part, we use the MGD noise model, where the weight of each component is controlled by the cluster center distribution $p(\omega_t)$ and expression classification distribution $p(r_t)$.

3) Update the set of samples. The measurement of the sample s_t^n is $\pi_t^{(n)} = \lambda_t^{(i)} * \mathrm{M}(s_t^n)$, where $\lambda_t^{(i)}$ is the importance sampling correction term. M is the sample measurement function, described in the next subsection.

4) After all the samples are generated and measured,

normalize $\pi_t^{(n)}$ so that $\sum_n \pi_t^{(n)} = 1$ and store the sample set as $\{(s_t^{(n)}, \pi_t^{(n)}), n = 1,... N\}$

After this step, the state with largest weight describes the tracking output in each frame, consisting of face pose (x, y, θ, sc) and deformation, which is obtained by projecting ($y_1 \ldots y_d$) back to the original shape space, through a nearest-neighbor scheme.

In order to measure a sample (function M in the algorithm above), we proceed in the following way. For each mixture center in the embedded space, a specific ASM model is selected to measure image observation. This measure is given by a residual error obtained after applying one step of ASM search (we refer to Cootes et al. [19] for details on the search process). Face pose initialization is given by the sample rigid part (x, y, θ, sc) and shape initialization is computed by projecting the non-rigid part ($y_1 \ldots y_d$) of the sample back to the original shape space (using a nearest-neighbor scheme).

Once we have a residual error for each one of the mixture centers, the desired sample measurement is obtained by a weighted sum of these residuals, where the weights corresponds to the likelihood of the sample non-rigid part ($y_1 \ldots y_d$) in each Gaussian model.

This scheme allows tracking to be robust under large facial expression changes. Next we describe how to update expression classification in each frame, using a common probabilistic framework.

4.2.2 Expression Recognition

We have already showed that the distribution of the discrete random variable r (the expression recognition variable) directly affects tracking (see sample prediction and dynamic model learning). Now we show how to update the posterior probability $p(r_t \mid y_{0:t})$ in every frame to identify facial deformation.

In the ICondensation tracking, by assuming statistical independence between all noise variables, Markov property and priors of the distributions $p(\omega_0 \mid r_0)$, $p(r_0 \mid y_0)$, $p(y_t \mid y_{t-1})$ on embedded space, our goal is to compute the posterior $p(r_t \mid y_{0:t})$. It is in fact a probability mass function (PMF) as well as a marginal probability of $p(r_t, \omega_t \mid y_{0:t})$. Therefore, the problem is reduced to computing the posterior probability $p(r_{0:t}, \omega_{0:t} \mid y_{0:t})$.

$$
\begin{aligned}
& p(r_{0:t}, \omega_{0:t} \mid y_{0:t}) = \\
& p(r_{0:t-1}, \omega_{0:t-1} \mid y_{0:t-1}) \frac{p(y_t \mid r_{0:t-1}, \omega_{0:t-1}) p(r_t \mid r_{t-1}) p(\omega_t \mid \omega_{t-1})}{p(y_t \mid y_{0:t-1})} \\
& = p(r_0, \omega_0 \mid y_0) \prod_{l=1}^{t} \frac{p(y_l \mid r_l, \omega_l) p(r_l \mid r_{l-1}) p(\omega_l \mid \omega_{l-1})}{p(y_l \mid y_{0:l-1})}
\end{aligned}
\tag{4.8}
$$

By marginalizing over $\omega_{0:t}$ and $r_{0:t-1}$, we obtain:

$$
\begin{aligned}
& p(r_t = l \mid y_{0:t}) = \int_{\omega_0} \int_{r_0} \ldots \int_{\omega_{t-1}} \int_{r_{t-1}} \int_{\omega_t} p(r_0, \omega_0 \mid y_0) \\
& \prod_{l=1}^{t} \frac{p(y_l \mid r_l, \omega_l) p(r_l \mid r_{l-1}) p(\omega_l \mid \omega_{l-1})}{p(y_l \mid y_{0:l-1})} d\omega_t dr_{t-1} d\omega_{t-1} \ldots d\omega_0 dr_0
\end{aligned}
\tag{4.9}
$$

This equation can be computed by prior distributions and the product of the likelihood $\prod_{l=1}^{t} p(y_l \mid r_l, \omega_l)$. This provides us the probability distribution of the

expression categories, given the sequence of embedded deformation vectors of the standard model.

4.2.3 Experimental Results

We present the experimental results on facial deformation tracking and expression recognition.

To learn the structure of the expression manifold, we need $O(10^3)$ images to cover basic expressions for each subject and to enable stable geodesic distance computation. Since there is no database with a sufficiently large amount of subject data available, we built our own small data set for the experiments. In our experiments, two subjects were instructed to perform a series of six kinds of prototypical facial expressions, representing happiness, sadness, anger, surprise, fear, and disgust. The subjects repeated the series seven times for the gallery set. We realized the difference between posed expression and spontaneous expressions in terms of amplitude and dynamics [56]. The probe set includes a long sequence (more than 10^4 frames) where the subject can change his/her expression randomly. To simplify the problem, we assume constant illumination and near frontal view pose. The sequences were recorded at 30 fps and stored at a resolution of 320x240. All results were obtained on a Xeon 2.8GHz CPU. The complete process, including alignment, embedding, and recognition, runs at 5 fps.

To generate the shape sequence from the training data set, we trained ten ASM models for different kinds of deformations. We manually select the model in this offline stage to robustly track facial deformation along the video sequences. The

shape space dimension is 90. We used our modified Lipschitz algorithm to obtain a space with dimensionality d=6.

Tracking

We also quantitatively analyze the performance of our tracker with a standard ASM tracker. Figure 4.2 shows a precision comparison, considering as ground truth a manual labeling of eye corners and lip corners. The same images were used to train both trackers. The difference is that our method automatically splits this data to train a set of models, which are probabilistically selected during tracking. This allows more robust performance under large facial expression changes.

We verified that our probabilistic method is able to track and recognize long sequences of subjects performing subtle and large expression changes. A complete output video sequence is available at http://ilab.cs.ucsb.edu/demos/IVC-seq2.m2v. Figure 4.3 shows two frames from a tracking and recognition test using a new video sequence. The overlaid graphical bars for each expression label in the figure indicate their respective recognition probabilities.

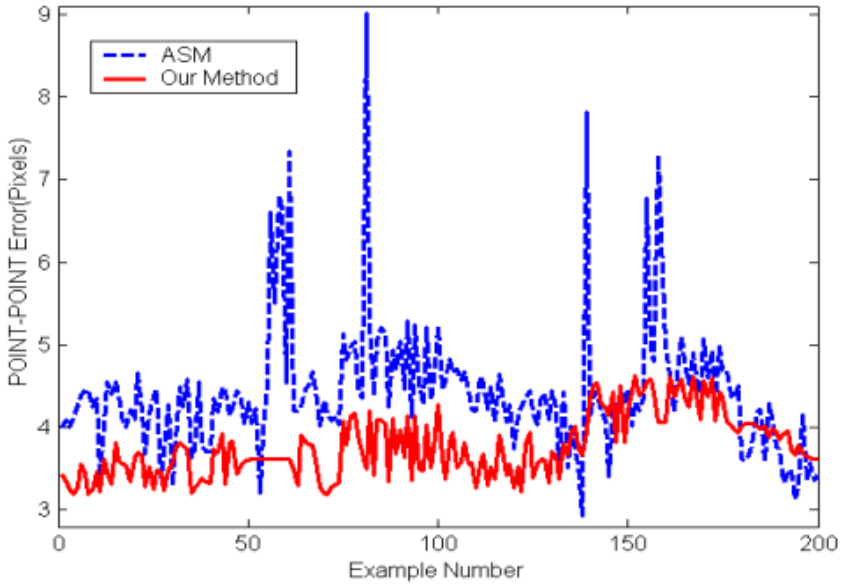


Figure 4.2: Comparison of tracking precision between an ASM tracker and our method. We have obtained considerably improvement, mainly under the presence of images with large expression changes.

Figure 4.3: Sample frames of our output tracking and recognition result in a video sequence.

Expression recognition

We visualize the learned manifold at the same time at video http://ilab.cs.ucsb.edu/demos/IVC-seq1.m2v.

The visualization of the trained manifold is shown on the right side. The embedded vector of the current frame is represented as a black point. During the expression transition, the black point "walks" from one expression path to another. The viewpoint of the manifold is changed concurrently for better visualization. Figure 4.4 shows some sample images from the available video. The first image is during a transition from fear to surprise. The second image is during a transition from anger to disgust. The third image and the fourth image are sadness and happiness respectively. The bar figures indicate the expression transition correctly. The quantitative measurement of expression recognition for every frame is not available because the output is represented as the posterior probability of basic expressions, and we do not have ground truth for this kind of representation.

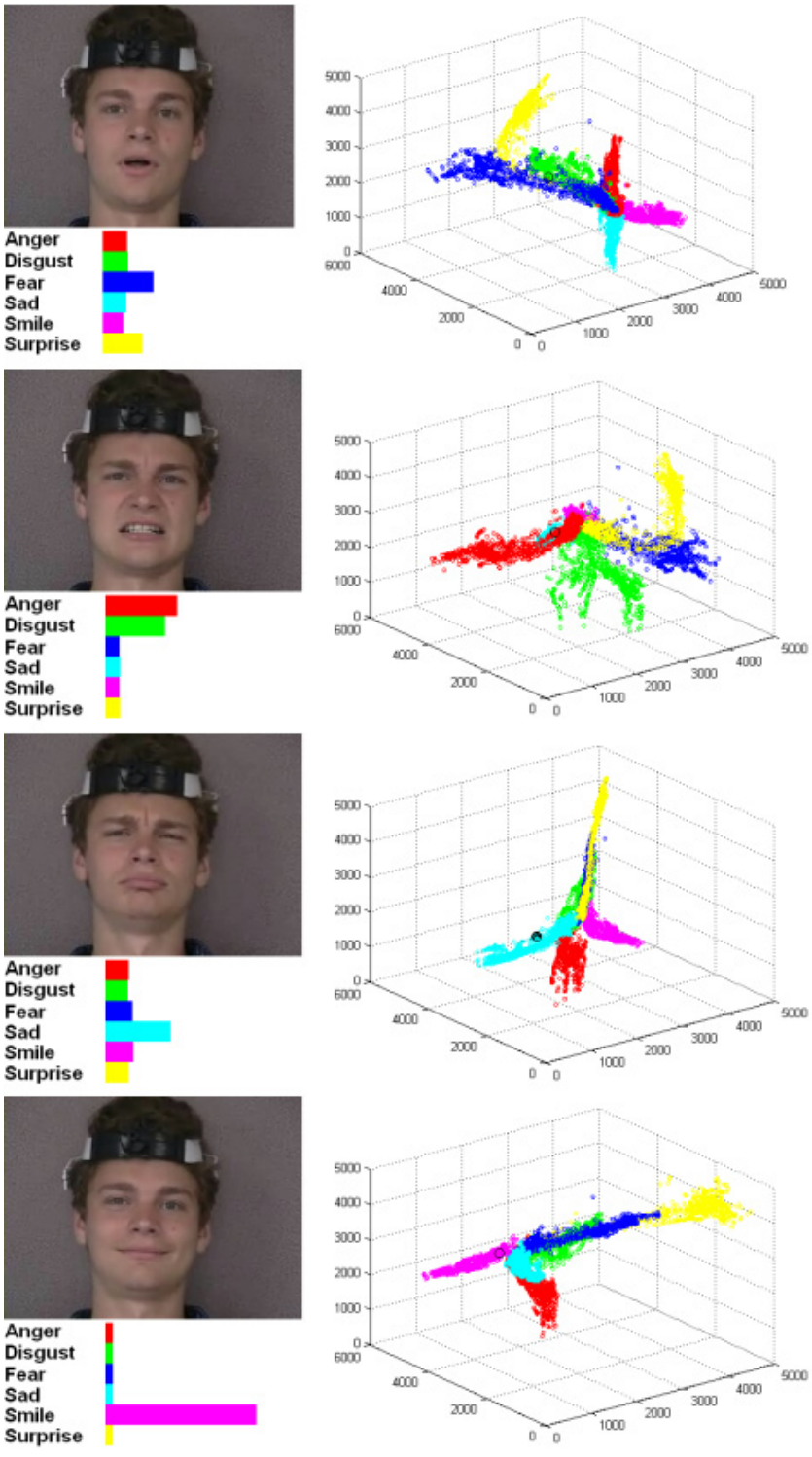


Figure 4.4: Facial expression recognition result with manifold visualization.

4.3. Synthesis of dynamic expressions

The manifold model can also be used to synthesize an image sequence with changing expressions. Given expression S, we keep the indexing $l_1,...,l_r,\ l=1,...,6,$ and r is the number of the clusters, such that:

$$p(c^{l_1} \mid S=l) < p(c^{l_2} \mid S=l)... < p(c^{l_r} \mid S=l)$$

For expression l, there are k gallery videos that begin from the neutral expression, pass the apex, and end with the neutral expression. We set the first video sequence as a template. Then we apply dynamic time warping to the following $k-1$ image sequences. Thus we have a standard time index for all k videos. For every cluster along the path l, we can measure the duration of the cluster by the range of time index of the images within the cluster. Note we compute the time range for increasing and decreasing expression separately since a cluster may cover both types of images at the same time. The time range for each cluster is $w_i, i=1,...,r.$ The average time range of all clusters is $\overline{w}$.

The algorithm for synthesizing an image sequence from expression A to expression B is listed as following. The critical part is to find a trajectory that maximizes the probability of the transitions between the clusters A_r and B_r. The optimal trajectory is computed by dynamic programming. The correlations between consecutive frames are maximized locally at the same time. To eliminate the jitter and redundancy in the image sequence, we keep a cache for recently appeared frames. If the same frame from the gallery appearances more than twice in the

passed n frames, it should be removed from the final video sequence. n is an empirical window width.

Input:

The beginning expression category: $A \in \{1,...,6\}$
The ending expression category: $B \in \{1,...,6\}$
The length of synthesized video sequence: $fnum$
The embedded vectors of r cluster centers:
$d_i \in R^6, i = 1,..., r$

Output:

Image sequence P

Function:

$floor\ (x)$: return the maximum integer no more than x .
$findnear\ (d, z)$: return the nearest z points to the embedded vector d on the learned manifold.
$GetRaw\ (d)$ **:** return the corresponding face image to the embedded vector d .
correlation(x, y): return the correlation between two images.
$T = floor\ (fnum\ / \overline{w})$;
$n_1 = A_r$; {the cluster with strongest expression A }
$n_T = B_r$; {the cluster with strongest expression B }
$[n_2,..., n_{T-1}] = \arg\max(P(c^{n_T} \mid c^{n_{T-1}})...P(c^{n_2} \mid c^{n_1}))$;
$count = 0$;
for i = 1 **to** $T - 1$
$betweenc = (w_{n_i} + w_{n_{i+1}}) / 2$;
$fbegin = findnear\ (d_i, 1)$;
$fend = findnear\ (d_{i+1}, 1)$;
$count = count + 1$;
$P_{count} = GetRaw\ (fbegin)$;

for $j = 1$ **to** $betweenc - 1$

$dist = (fend - fbegin)/(betweenc - j)$;
$candi = findnear\ (fbegin + dist, 5)$;
$comp = GetRaw\ (fbegin)$;

for k = 1 **to** 5

$candi\_im(k) = GetRaw(candi(k))$;
corr(k)=correlation(comp, candi_im(k));
if $(k == 1)$

```
        then   se = 1;
               max = corr(k);
             else if (corr(k) > max)
                     se = k;
                     max = corr(k);
                   end
                 end
               end
               count = count + 1;
               P_count = candi_im(se);
               fbegin = candi(se);

          end
        end
        return P_i, i = 1,..., count
```

With the manifold model, we synthesize image sequences of aligned face appearances with changing expressions. There are about 6000 images from 42 video sequences (seven for each basic expression) in each gallery set. The lengths of synthesized image sequences are around 200. Figures 4.5 and 4.6 show some selected images (every 20th frame) from the synthesized sequences. The trajectories with the maximum transition probability between clusters reflect the expression change correctly.

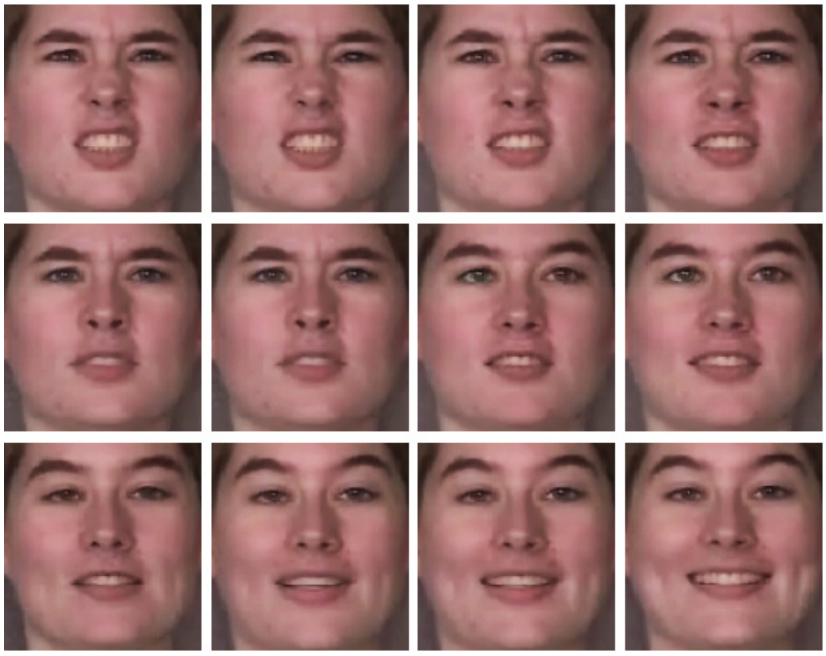


Figure 4.5: 12 frames selected from a transition from anger to happiness.

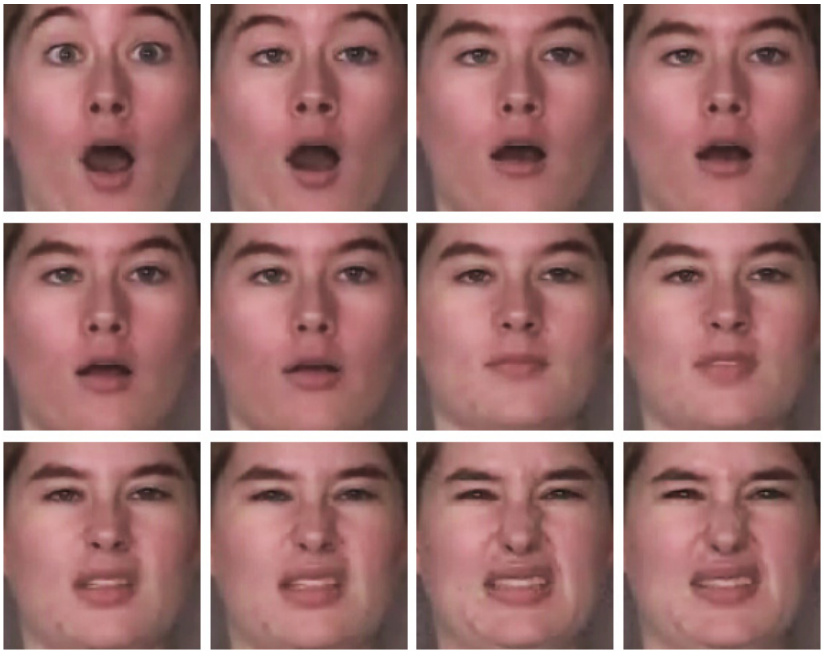


Figure 4.6: 12 frames selected from a transition from surprise to disgust.

4.4 Expression Editing

The user can define the any expression editing function F as needed. F: $R^6 \to R^6$.

$F(p(S=1),...,p(S=6)) = [q_1, q_2, ..., q_6]$

where $\sum_{i=1}^{6} q_i = 1$, q is the new likelihood of one kind of facial expression. For example, if we want to edit all sadness (S=1) videos to anger (S=2), the mapping function can be defined as

F (p (S=1), p (S=2), …, p (S=6))=

[p (S=2), p (S=1), …, p (S=6)], when p (S=1)> γ. (4.10)

This function will increase the likelihood of anger when the sadness is detected, that is, its likelihood is above a threshold γ.

The system automatically searches for the embedded vector with likelihood that is closest to the "range" expression. It first looks for the cluster whose center has the closest likelihood. In that cluster, the point closest to the embedded vector of the input frame is selected. We transfer the corresponding deformation vector back to the model in the new video. The deformation vector is blended with the deformation at the previous frame to ensure smooth editing. The synthesized 2D image uses the head pose in the real input frame and the texture information of the dense model.

Figure 4.7 is an example of expression editing. First row is from the input video of sadness. We define the expression mapping function as Eq. 4.10. The second row is the deformed dense mesh. The third row is the output: the first image is unchanged, the following images are synthesized anger faces by the expression mapping

function. The system recognized the sadness correctly and synthesized new faces with anger expression correspondingly.

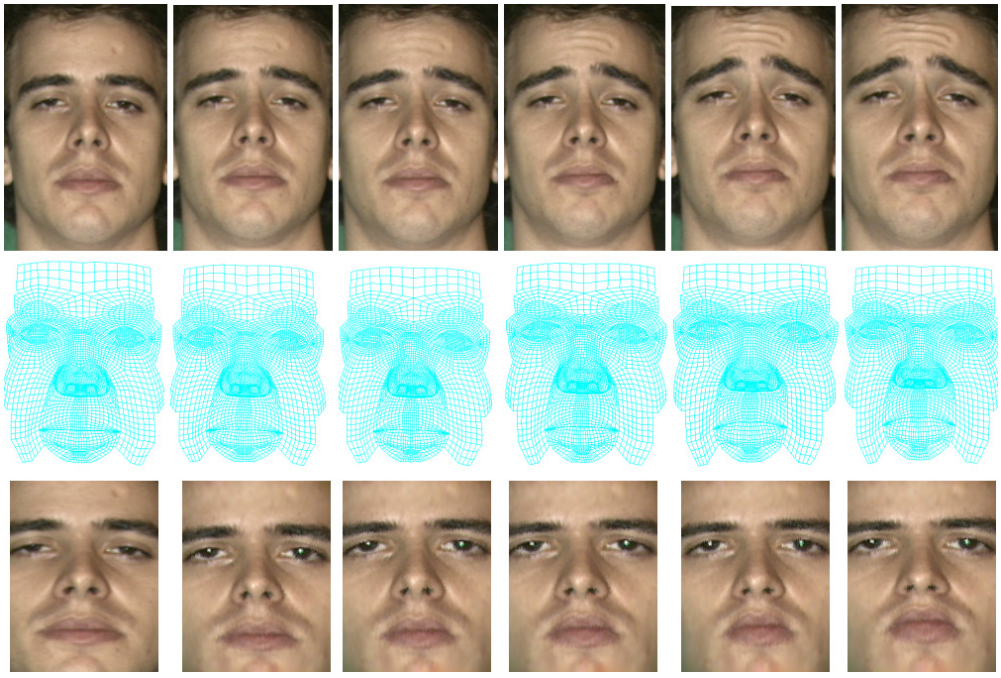


Figure 4.7: Expression editing examples

Chapter 5

Regional FACS

The expression recognition method in Chapter 4 classified the expression into six basic categories quantitatively. For some applications that require precise facial deformation, such as behavioral science, medical analysis, and expression mimic, etc., this kind of result is too coarse and subjective. FACS provides the descriptive power necessary to describe the details of facial expression. Although the number of atomic action units is relatively small, more than 7,000 different AU combinations have been observed. But AUs have inadequate quantitative definitions and, as noted, can appear in complex combinations. It takes hours for an expert to annotate a one-minute video. Furthermore it is difficult to connect AU combinations with semantic meanings. Researchers have worked on the automatic recognition of AUs. Tian [27] used multi-state feature based method to extract facial deformation and neural networks for AUs classification.

A key motivation of our research is based on the combinational property of AUs. Most AUs only affect a small sub-region of the whole face. When AUs occur in combination, they may be additive: the combination does not change the appearance of the constituent AUs; or non-additive: the appearance of the constituents does change. The former mostly happens on AUs that affect disconnected regions, such as AU2 in the forehead region and AU22 in the lip region. The latter mostly happens on closely related regions, such as combination of AU1, AU2 and AU4 on the

eye/eyebrow region as in Fig. 5.1. The additive combinations of AUs in each small region are much less complex compared to the total number of combinations. So we divide the face into nine sub-regions as illustrated in Fig. 5.2. The guideline for the subdivision scheme is that the sub-regions are small while avoiding expression wrinkles crossing the sub-region boundaries.

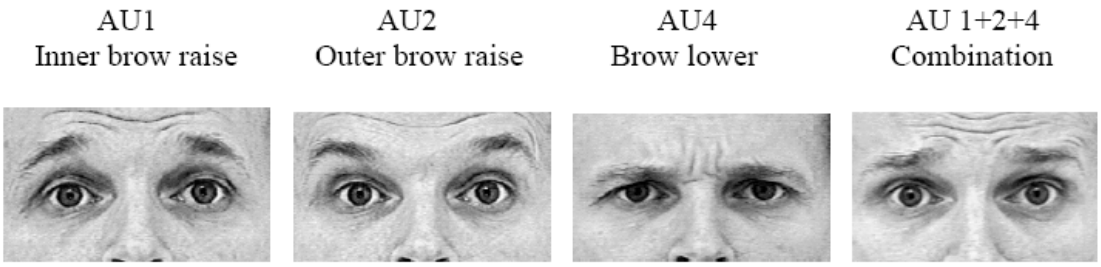


Figure 5.1: Three Action Units occur individually and in combination.

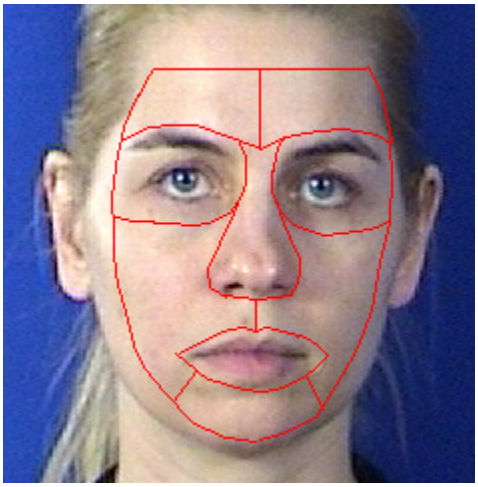


Figure 5.2: sub-region division of the face

With Regional FACS, the deformation of each sub-region can be embedded into a sub-manifold in a similar way. On the sub-manifolds, similar expressions are still points in the local neighborhood. But the dimensionality of the sub-manifolds is

objectively determined by the number of AUs associated with the sub-regions. We choose a proper reference set in Lipschitz embedding such that images with maximum intensity of the AUs are furthest from a center that corresponds to the neutral expression. The images corresponding to the combination of AUs are embedded between the new paths. Each path consists of several clusters. A probabilistic model of transition between the clusters and paths is learned through training videos in the embedded space. The likelihood of one kind of AU is modeled as a mixture density with the clusters as mixture centers. The transition between different combinations of AUs is represented as the evolution of the posterior probability of the AU paths. One of the advantages of expression manifold is that the expression data organized with manifold framework can facilitate the training of the expression recognizer greatly. With the number of combinations of AUs as large as O(10^3), it is very difficult to use traditional pattern recognition methods, such as decision tree or neural network. But the Regional FACS with sub-manifolds can achieve good recognition rate as well in this case.

A generalized manifold is built for all training subjects through deformation transfer. So the facial expression of new subjects can be recognized effectively. We train a Support Vector Machine to combine the AU recognition results from each sub-manifold. Fig. 1.1 illustrates the overall system structure. The generality of the system has been tested on the MMI face database [46] and the Cohn-Kanade database [30], both of which are FACS-coded for ground-truth.

The remainder of this chapter is organized as follows. We discuss the learning of sub-manifolds in Section 5.1. The probabilistic model built on expression sub-manifolds is described in Section 5.2. Section 5.3 presents the experiments we conducted on the databases. Section 5.4 concludes the chapter with discussion.

5.1 Learning the sub-manifolds

This section presents how to construct the sub-manifolds of facial expression.

5.1.1 Facial Feature Tracking

Table 3: Sub-regions and the associated AUs

Sub-Region	# of facets	# of vertex	Associate AUs
Forehead (left & right)	6	13	2, 4, 5
Eye (left & right)	36	47	1, 2, 4, 5, 6, 7, 42, 45
Nose	74	92	4, 9
Cheek (left & right)	30	47	6, 9, 10, 11, 12, 13, 18, 20, 30
Mouse	12	24	8, 10, 12, 14-19, 20, 22-28, 30
Chin	14	22	15, 17, 26, 27

We use a robust real time feature tracker from Nevengineering [35] to provide the 2D positions of 22 prominent feature points. The eye contour is further fitted through edge detection as in Fig. 5.3(a). A generic 3D mesh with 268 vertices and 244 facets is fitted manually at the first frame. For the following frames, this mesh is

warped by the position of the tracked features. The face is divided into sub-regions naturally through the generic mesh. The sub-regions with the associated AUs are listed in Table 3.

(a) (b)

Figure 5.3: The sample with tracked feature point and fitted mesh.

5.1.2. Lipschitz Embedding

The video sequences in the MMI face database are FACS coded in two different ways: one is FACS coded with temporal segment (that is, frame-based); the other is FACS coded with all AUs that occur in the video sequence. None of the FACS codings in the MMI database have intensity scores. In our experiments, we used the former type of videos to build the sub-manifolds of facial expression. For example, the number of AUs associated with a sub-region is $\boldsymbol{n}$. We select $\boldsymbol{n}$ reference sets. Because the frame-based FACS coding just labels the appearance/disappearance of each AU, we manually select the frames with maximum intensity of each AU for its

reference set. Thus the dimension of the embedded sub-manifold is equal to the number of reference sets, ***n***. After we apply the enhanced Lipschitz embedding to the gallery set, there are ***n*** basic paths in the embedded space, emanating from the center that corresponds to the neutral image. The images with combinational AUs lie between those basic paths.

5.2 Probabilistic FACS Recognition

The recognition of AUs is greatly facilitated on the sub-manifolds. The goal of the probabilistic model is to exploit the temporal information in video sequences. The recognition result is the joint output from each sub-manifold.

5.2.1 Learning

We apply an enhanced Lipschitz embedding to the training set. The training set contains videos with frame-based FACS coding. For one sub-manifold, the number of associated AUs is ***N***. Assume there are totally ***K*** video sequences in the training set. The embedded vector for the ***i***th image in the ***j***th video sequence is $I_{j,i} \in R^N$, $j = \{1,..., K\}$. Using a K-means clustering technique, all points are grouped into clusters $c^m, m = 1,..., r$. We compute a cluster frequency measure $T_{m1,m2} = \#(I_{j,i} \in c^{m1} \,\&\, I_{j,i+1} \in c^{m2}, j = 1,..., K)$. The prior $p(c^{m2} \mid c^{m1})$ is learned as

$$p(c^{m2} \mid c^{m1}) = \begin{cases} \delta, T_{m1,m2} = 0 \\ T_{m1,m2} * scale, otherwise \end{cases}$$

where δ is a small empirical number. Scale and δ are selected such that $\sum_{m2} p(c^{m2} \mid c^{m1}) = 1$.

The prior $p(c \mid n, n = 1,...N)$ is assigned according to the AU intensity of the cluster center, varying from 0 to 1. Since there is no intensity scoring in the training data from the MMI face database, we use the temporal segment in the following way: the first and the last frames with some AU occurrence are assigned with intensity 0.5; the frame selected as reference set is assigned with intensity 1.0; and the first and the last frames of the video sequence are assigned with intensity 0.0. We mapped the 0/1 occurrence labeling in the training set into a [0,1] region continuously as illustrated in Fig. 5.4. For example, AU1 is from frame 19 to frame 49 in one video sequence of 76 frames. The maximum intensity is at frame 30. The clusters including frame 1, 19, 30, 49, 76 are scored with intensity 0, 0.5, 1.0, 0.5, and 0 respectively. If scoring conflict happens within one cluster, we use majority vote or reduce the size of cluster. All the other clusters are assigned with an intensity score according to the distance between their cluster centers and the labeled clusters' as explained in Fig. 5.5.

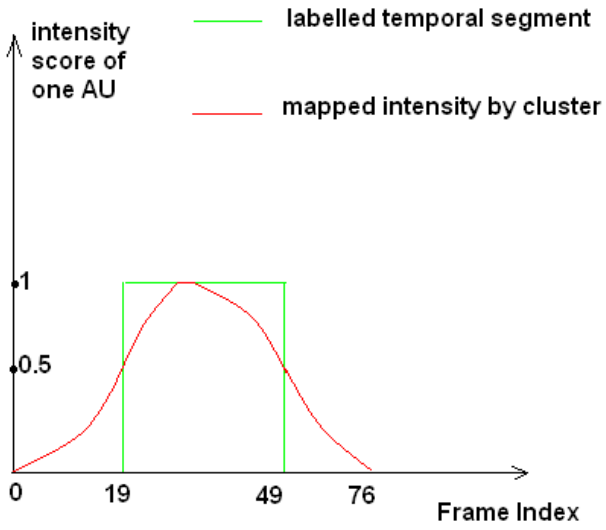


Figure 5.4: The illustration of intensity mapping from temporal segment to continuous scoring.

Function: calculate the AU intensity score for a cluster

Assume the clusters with the first and the last frame of a video sequence have centers at s_1, s_2;

The clusters with beginning and ending frames of the AU have centers at s_3, s_4;

The cluster with maximum intensity of the AU has center at s_5;

The cluster with unknown intensity score has center at *s*;

$md = \max(|s_i - s_j|); i, j = 1,...,5$

$d_1 = \min(|s - s_1|, |s - s_2|);$
$d_2 = \min(|s - s_3|, |s - s_4|);$
$d_3 = |s - s_5|;$

if($\min(d_1, d_2, d_3) > md$)

intensity = 0;

else

if ($\min(d_1, d_2, d_3) == d_1$) *intensity* = $d_1/(d_1 + d_2)$;

if ($d_2 < d_1 < d_3$) *intensity* = 0.5 - $d_2/(d_1 + d_2)$;

if ($d_2 < d_3 < d_1$) *intensity* = 0.5 + $d_2/(d_3 + d_2)$;

if ($\min(d_1, d_2, d_3) == d_3$) *intensity* = 1 - $d_3/(d_3 + d_2)$;

end

return intensity;

Figure 5.5: Algorithm of calculation of intensity score

When the image of a cluster center shows high intensity score for one AU, there is less ambiguity for the frames in the cluster to be classified with occurrence of that AU. Therefore the corresponding cluster has a higher $p(c|n)$.

By Bayes' rule,

$$p(n \mid c) = \frac{p(c \mid n)p(n)}{\sum_{n} p(c \mid n)p(n)}$$

Given a probe video sequence in the embedded space $I_t, t = 0,1,...$, the expression recognition can be represented as the evolution of the posterior probability $p(n_{0:t} \mid I_{0:t})$, where $n = 1,...N$.

We assume statistical independence between prior knowledge on the distributions $p(c_0 \mid I_0)$ and $p(n_0 \mid I_0)$. Using the overall state vector $x_t = (n_t, c_t)$, the transition probability can be computed as:

$$p(x_t \mid x_{t-1}) = p(n_t \mid n_{t-1})p(c_t \mid c_{t-1}) \tag{5.1}$$

We define the likelihood computation as follows

$$\begin{aligned} & p(I \mid c, n) \\ &= p(I \mid c)p(c \mid n) \\ &\propto \exp[-\frac{1}{2\sigma_c^2} d(I, u_c)]\, p(c \mid n) \end{aligned}$$

where u_c is the center of cluster c, σ_c is the variation of cluster c.

Given this model, our goal is to compute the posterior $p(n_t \mid I_{0:t})$. It is in fact a probability mass function (PMF) since n_t only takes values from 1 to ***N***. The marginal probability $p(n_t, c_t \mid I_{0:t})$ is also a PMF for the same reason.

Using Eq. (5.1), the Markov property, statistical independence, and time recursion in the model, we can derive:

$$
\begin{aligned}
& p(n_{0:t}, c_{0:t} \mid I_{0:t}) = p(x_{0:t} \mid I_{0:t}) \\
&= p(x_{0:t-1} \mid I_{0:t-1}) \frac{p(I_t \mid x_t) p(x_t \mid x_{t-1})}{p(I_t \mid I_{0:t-1})} \\
&= p(n_{0:t-1}, c_{0:t-1} \mid I_{0:t-1}) \frac{p(I_t \mid c_t, n_t) p(n_t \mid n_{t-1}) p(c_t \mid c_{t-1})}{p(I_t \mid I_{0:t-1})} \\
&= p(n_0, c_0 \mid I_0) \prod_{i=1}^{t} \frac{p(I_i \mid n_i, c_i) p(n_i \mid n_{i-1}) p(c_i \mid c_{i-1})}{p(I_i \mid I_{0:i-1})}
\end{aligned}
$$

By marginalizing over $c_{0:t}$ and $n_{0:t-1}$, we obtain Eq. (5.2):

$$
p(n_t \mid I_{0:t}) = \int_{c_0} \int_{n_0} \cdots \int_{c_{t-1}} \int_{n_{t-1}} \int_{c_t} p(c_0 \mid I_0) p(n_0 \mid I_0) * \prod_{i=1}^{t} \frac{p(I_i \mid n_i, c_i) p(n_i \mid n_{i-1}) p(c_i \mid c_{i-1})}{p(I_i \mid I_{0:i-1})} dc_t dn_{t-1} dc_{t-1} \ldots dn_0 dc_0 \tag{5.2}
$$

which can be computed by the priors and the likelihood $p(I_i \mid n_i, c_i), i = 1,...,t$. This gives us the probability of all associated AUs in that sub-region, which leads to the joint recognition results.

5.2.2. Joint AU Recognition Results

The total number of AUs in our training data set is 29. The output from each sub-manifold can be represented as a vector $F_i \in R^{29}$, $\boldsymbol{i}$ =1, …9. The components of $\boldsymbol{F}$ corresponding to the AUs that do not affect that sub-region are set as 0. A multi-class Support Vector Machine is trained to recognize the existence of these 29 AUs. The input is $F_i, i = 1,...9$. The output is $H \in R^{29}$. $\boldsymbol{H}$ is a vector with 0/1 components. When some AU exists in that frame, the corresponding vector is 1, and vice versa.

5.3 Experimental Results

5.3.1 Data Set

The number of videos in the MMI face database is more than 1200 with 41 subjects. Most of videos are FACS coded. As we have stated, the videos in MMI database are FACS coded in two different ways: one is FACS coded with temporal segment; the other is FACS coded with all AUs that occur in the video sequence. Considering only videos with frontal face or dual face (contain both frontal and profile view), the number of videos of the first coding type is 185 with 15 subjects.

The experiment consists of two parts. The first is frame occurrence accuracy. We use leave-one-out correlation in this part. The data set is the frame-based FACS coded 178 videos with 13 subjects (except two subjects with eyeglasses) from MMI database. For example, there are 14 videos with frame-based FACS coded that AU1 occur. We use all but one of these 14 videos to learn expression sub-manifolds and that video is used as the test data for frame occurrence accuracy for AU1. We repeat this process for all 14 videos and take the mean of accuracy as frame occurrence accuracy for AU1.

The second is video occurrence accuracy. Video occurrence is calculated based on the frame occurrence. The duration of different AUs vary significantly. So we take the mean of the duration of each AU from the training set as *length* of each AU. For example, the mean duration for AU4 is *length* = 43.5 frames. As long as AU4 occurs in more than $length*0.5 = 21.7$ frames with maximum duration less than

*length*2* = *87* frames, AU4 is recognized in this video. In this part, we select 178 videos from the first type of videos for sub-manifold learning. The test set includes 584 videos with 37 subjects in the MMI database (belong to type two and not in the training set) and 258 videos with 43 subjects in the Cohn-Kanade database. It is noted that *length* (duration) of AU for C-K database is only the half of *length* for MMI database because the videos in C-K database all stop at the apex.

The ground truth is from the labeling in both databases, which is from human experts. There are variance and discrepancy among different human experts. But we will not discuss this issue in the thesis. The experimental results are shown in Table 4. The average recognition rate is 91.12% for video occurrence, 76.67% for frame occurrence. The video occurrence rate is comparable with the results obtained by Tian et al. [14]. The frame occurrence is relatively lower for the following reasons: (1) To our knowledge, our system is able to recognize the largest number of AUs. The false positive rate is high for some AUs due to the subtlety of the related muscle movement, such as AU11, which has never been recognized in previous literature; (2) The portion of novel faces, that is 64.86%, is high in our test data set. Part of facial deformation is weakened during the deformation transfer; (3) We set the practical intensity score for AUs based on the temporal segment, which may not reflect the true intensity score. But with system tolerance, the video occurrence accuracy is still high given the results of frame occurrence recognition.

Table 4: AU Recognition Results in both databases

AU	Video Occurrence Accuracy in MMI	Frame Occurrence Accuracy in MMI	Video Occurrence Accuracy in C-K
1	0.9863	0.8444	0.9566
2	0.9897	0.8256	0.9427
4	0.9767	0.8067	0.9374
5	0.9851	0.8498	0.9436
6	0.9822	0.8360	0.9427
7	0.9573	0.8005	0.9155
8	0.9565	0.7983	N/A
9	0.9667	0.8008	0.9377
10	0.9777	0.8298	0.9310
11	0.7931	0.6215	0.6993
12	0.9646	0.8159	0.9175
13	0.8205	0.6402	N/A
14	0.8846	0.7076	0.8318
15	0.9423	0.7897	0.9204
16	0.9088	0.7441	0.8651
17	0.9516	0.7994	0.9413
18	0.8511	0.7013	0.8065
19	0.7059	0.6016	N/A
20	0.9706	0.8021	0.9251
22	0.9250	0.7539	0.8527

23	0.7442	0.6127	0.7091
24	0.8831	0.7324	0.8542
25	0.9745	0.8454	0.9354
26	0.9478	0.7930	0.9119
27	0.9862	0.8754	0.9566
28	0.9257	0.7611	N/A
30	0.9697	0.8213	N/A
42	0.7692	0.6401	N/A
45	0.9505	0.7836	0.9223

5.3.2. Recognizing Basic Expressions

There are 481 videos with 97 subjects in the Cohn-Kanade expression database. Every subject in the C-K database performed one of six kinds of basic expressions. All videos begin from neutral expression and end at the apex. All of the videos are FACS coded in the second way. It is very straightforward to learn the relationship between the six kinds of basic expressions and the FACS coding through the data in C-K database. On the other hand, there are 177 videos with 26 subjects from MMI database are coded only with basic expressions but without FACS coding.

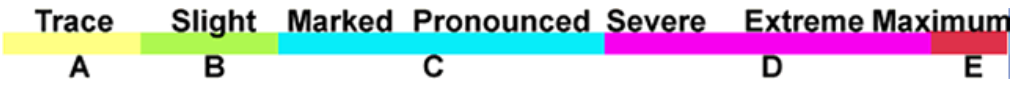


Figure 5.6: Relation between the Scale of Evidence and Intensity Scores

We trained a six-class SVM to learn the basic expression recognition given the FACS coding in a similar way with method in Section 5.3.2. Part of the Cohn-Kanade database is labeled with intensity score, that is *a, b, c, d*, and *e* as illustrated in Fig. 5.5. The intensity is mapped as a = 0.5; b = 0.6; c = 0.75; d = 0.90; e = 1.0 in vector $\boldsymbol{F}$. We test the results on those 177 videos from MMI face database. The input is the AU recognition results, and the output is one of the basic expression categories. The recognition rate is 97.74%.

5.4 Discussion

Automatic facial expression analysis is important to understand human emotion, to human computer interaction, and to computer animation. FACS is a well known and useful measurement for facial deformation. Our framework of Regional FACS for AU recognition provides a possible solution to tackle the huge number of different AUs combinations and exploit the correlation between additive/non-additive AUs. The face region is divided into nine sub-regions. Lipschitz embedding embeds the normalized deformation of the sub-regions in low dimensional sub-manifolds**.** The sub-manifold of each sub-regional can be considered as the sub-vector of the whole manifold. A SVM is learned to generate the joint AU recognition results.

Our system is capable to recognize 29 AUs which covers the most frequently appeared facial deformation in normal life. We test our system intensively on MMI face database and Cohn-Kanade database. Both of the databases are FACS coded for

ground truth. The experiments results on the probe sets demonstrate that AU can be recognized effectively. We also explore the relationship between six kinds of basic expression and FACS coding and achieve high accuracy on the database with all novel faces.

One of the main drawbacks of our algorithm is the computational speed. The deformation transfer is very time consuming. The whole processing is far from real time. We do not consider the illumination or pose change because the images/videos in both databases are high quality with well controlled illumination. The faces are all (close to) frontal. There are images with profile view for partial subjects in MMI database. Further experiments may be performed on this part of data.

Chapter 6

3D Facial Expression Analysis

There has been a lot of study on human facial expressions using either 2D static images or 2D video sequences. With 3D model, the 2D-based algorithms are capable of handing limited pose variations, but usually fail during large pose change. Although 3D modeling techniques have been extensively used for 3D face recognition and 3D face animation, barely any research on 3D facial expression recognition using 3D range data has been reported. A primary factor for preventing such research is the lack of a publicly available 3D facial expression database. The major bottleneck is real time 3D data capture and registration for facial deformation. We explore this area in this Chapter. We talked about a real time range data capture camera/projector system and the registration of the range data. The preliminary database includes 36 videos with 6 subjects.

6.1. 3D Expression Database

We build a 3D expression database by capturing real-time range data of people making different facial expressions. The range data were registered by robust feature tracking and 3D mesh model fitting.

6.1.1. Real-time 3D scanner

To construct a high quality 3D expression database, the capture system should provide high quality texture and geometry in real-time. Quality is crucial for accurate analysis and realistic synthesis. Real-time is important for subtle facial motion capture and temporal study of facial expression.

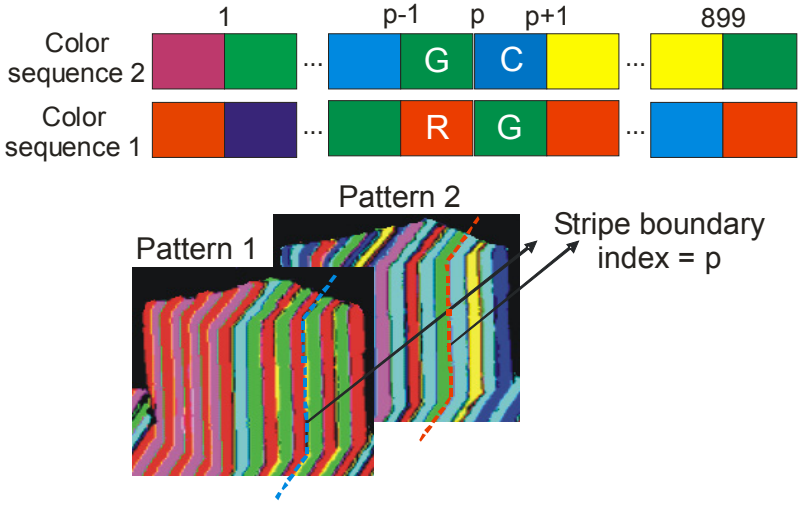


Figure 6.1: Decoding stripe transitions.

The system used for obtaining 3D data is based on a camera/projector pair and active stereo [57]. It is built with off-the-shelf NTSC video equipment. The key of this system is the combination of the color code (*b*,*s*)-BCSL [58] with a synchronized video stream.

The (*b*,*s*)-BCSL code provides an efficient camera/projector correspondence scheme. Parameter *b* is the number of colors and *s* is the number of patterns to be projected. Two patterns is the minimum, giving the best time coherence

compromise. The complementary patterns are used to detect stripe transitions and colors robustly. Our system applies six colors that can be unambiguously detected through zero-crossings: RGBCMY. In our experiments, we use a (6,2)-BCSL code that features two patterns of 900 stripes.

To build camera/projector correspondence, we project a subsequence of these two patterns onto the scene and detect the projected stripe colors and boundaries from the image obtained by a high-speed camera. The four projected colors, two for each pattern, detected close to any boundary are uniquely decoded to the projected stripe index *p* (Fig. 6.1). The correspondent column in the projector space is detected in O(1) by using (6,2)-BCSL decoding process. The depth is then computed by the camera/projector intrinsic parameters and the rigid transformation between their reference systems.

We project every color stripe followed by its complementary color to facilitate the robust detection of stripe boundaries from the difference of the two resulting images. The stripe boundaries become zero-crossings in the consecutive images and can be detected with sub-pixel precision. One complete geometry reconstruction is obtained after the projection of the pattern 1 and its complement followed by pattern 2 and its complement.

The (6,2)-BCSL can be easily combined with video streams. Each 640x480 video frame in NTSC standard is composed of two interlaced 640x240 fields. Each field is exposed/captured in 1/60 sec. The camera and projector are synchronized using genlock. For projection, we generate a frame stream interleaving the two patterns that is coded with its corresponding complement as fields in a single frame.

This video signal is sent to the projector and connected to the camera's genlock pin. The sum of its fields gives a texture image and the difference provides projected stripe colors and boundaries. The complete geometry and texture acquisition is illustrated in Fig. 6.2.

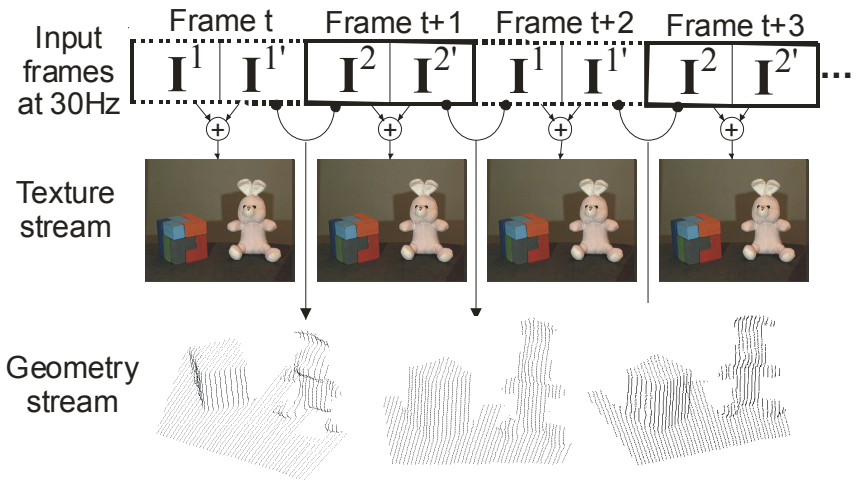


Figure 6.2: Input video frames, and the texture and geometry output streams with 30fps rate.

This system is suitable for facial expression capture because it maintains a good balance between texture, geometry and motion detection. Our videos were obtained by projecting 25-35 stripes over the face and the average resolutions are: vertical = 12 points/cm and horizontal = 1.25 points/cm (right bottom window of Fig. 24). We used one Sony HyperHAD camera and one Infocus LP-70 projector.

6.1.2. 3D Data Registration

The acquired range data need to be registered for the following analysis. The range points are first smoothed by radial basis functions (RBF). We build a coarse mesh model with 268 vertices, 244 faces for face tracking (Fig. 6.3). A generic coarse model is fitted manually at the first frame. A robust feature tracker from Nevengineering [59] provides the 2D positions of 22 prominent feature points (Fig. 6.4 (a)). The mesh's projection was warped by the 22 feature points. The depth of the vertex was recovered by minimizing the distance between the mesh and the range data [60].

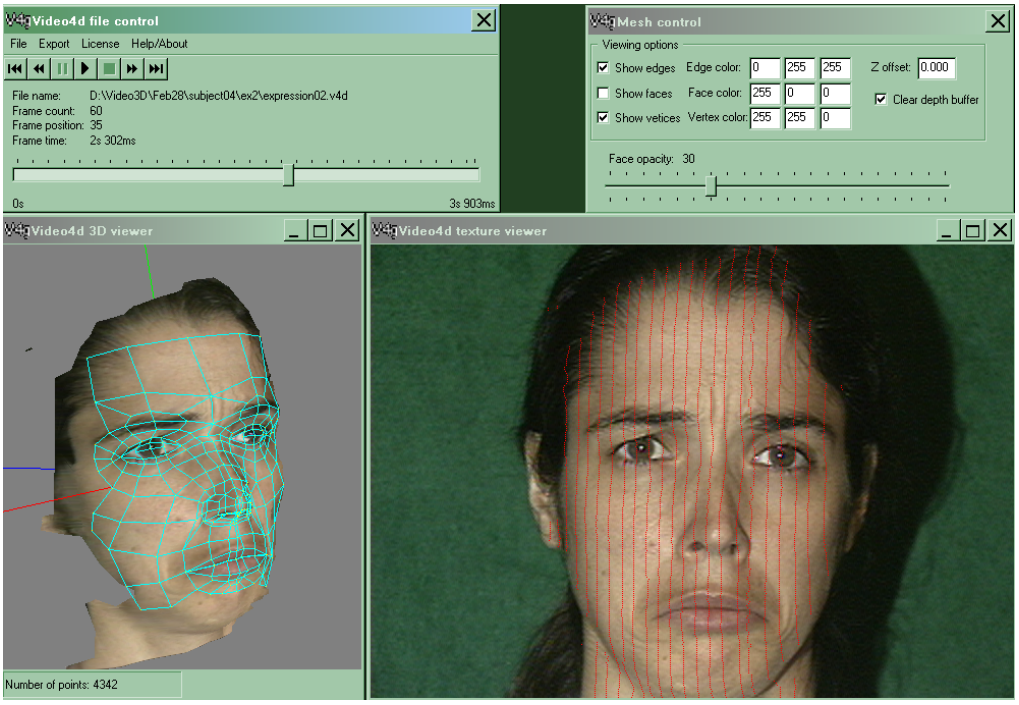
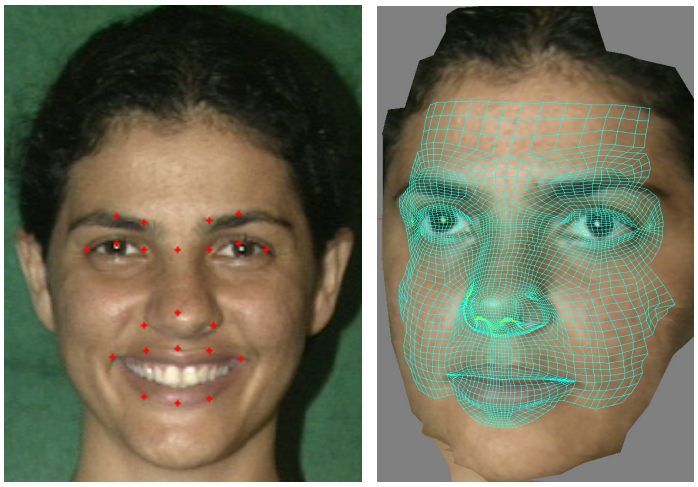


Figure 6.3: An example of 3D data viewer with fitted mesh



(a) (b)

Figure 6.4: (a) The 2D tracking results. (b) The dense mesh model

An example of the 3D viewer is shown in Fig. 24. The left bottom window shows the range data with the fitted mesh. The right bottom window is the texture image with the projected 3D points. Fig. 6.4 (a) shows the texture image with the 22 tracked feature points. Fig. 6.4 (b) shows the dense mesh with 4856 vertices and 4756 faces. The dense model is used for the synthesis of new expressions.

We collected 3D training videos from 6 subjects (3 males, 3 females). Every subject performed six kinds of basic expressions. The total number of frames in the training videos is 2581. We use Magic Morph morphing software to estimate the average of the training faces, and we use that average as the standard model. The standard model only contains geometrical data, no texture data. It will approach the "average" shape of human faces when the number of training subjects increases.

Fig. 6.5 includes some examples of the mesh fitting results. Images in each row are from the same subject. The first column is the neutral expression. The second

and third columns represent large deformation during the apex of expressions. We change the viewpoints of 3D data to show that the fitting is very robust. A supplementary video is available at http://ilab.cs.ucsb.edu/demos/AMFG05.mpg. This video gave a snapshot of our database by displaying the texture sequences and 3D view of the range data with the fitted mesh at the same time.

Figure 6.5: Mesh fitting for training videos

Chapter 7

Summary

7.1 Major Contributions

In this thesis, we proposed a novel framework for dynamic facial expression analysis on manifolds. We now summarize our main contributions:

(1) A new representation for tracking and recognition of facial expressions, based on manifold embedding and probabilistic modeling in the embedded space. Our experimental results show that the generalized manifold methods provide an analytical way to analyze the relationship between different expressions, and to recognize blended expressions.

(2) A probabilistic expression classification method, which integrates information temporally across the video sequence. In contrast with traditional methods that consider expression tracking and recognition in separate stages, we address these tasks in a common probabilistic framework, which enables them to be solved in a cooperative manner.

(3) Regional FACS for AU recognition provides a possible solution to tackle the huge number of different AUs combinations and exploit the correlation between additive/non-additive AUs. Our system is capable to recognize 29 AUs which is the largest set that can be automatically recognized in the literature by our knowledge.

(4) The exploration on building 3D facial expression database. We used a projector/camera system to capture real time facial deformation. The range data is registered through feature point tracking and 3D mesh fitting.

7.2 Limitations

Our system performs well for the tasks it is designed for. However, it is a research project and has many practical limitations. The processing speed is one of the major concerns. All of the learning/recognition are offline and far from real time. There is no rescue strategy when the system fails, which is a necessary in a practical system.

The input videos must have near-frontal faces. A possible future research direction is to consider variation on face pose and illumination [61,62], which will add more degrees of freedom to manifold of expression. How these factors affect the intrinsic geometry of expression manifold will be a challenging topic for future study.

The databases we used or built include no spontaneous facial expression. Subjects are supervised to perform specific facial expression, which is more dramatic than expression in normal life. The model trained from those databases may not necessarily perform well with videos taken from real time. On the other side, it is very hard to have ground truth for real videos due to the low quality and the variation of pose and illumination. How to build databases of spontaneous facial expression is a critical issue for future research.

The subjects usually speak while changing facial expression at the same time. The lip shape is affected equally by both expression and vocal content. Considering that case, we need to adjust the weight of upper face and lower face to recognize six basic expressions correctly. There is no such data of talking & expression face in the databases we have experiments on, further research on this topic may help to build a more practical system.

7.3 Future Works

Facial expression has been researched systematically for more than 200 years since the age of Darwin. In this section, we talk about the possible road map for the future work.

First, as we have stated, more diversified databases are needed: databases with videos taken in normal life, databases with 3D real time range data; databases with FACS annotation and time segment, etc. They will serve different levels of requirements and different applications well.

Second, the robust facial feature tracking in 2D videos is a necessary requirement for an automatic facial expression analyzer. We tried ASM with multiple clusters, Active Wavelet Networks, and a commercial software from Nevengineering. Each of them worked well under some constraints. The improvement on facial feature tracking will benefit the research on facial expression for sure.

Third, there are many potential applications for facial expression analysis. Understanding facial expression will not only help the face recognition but also the facial expression animation. But compared to fingerprint, iris recognition, facial expression is still an immature technique. How to integrate this new technique with the existing systems is an important issue for all researchers in this field.

Facial expression analysis will be an open field for multiple fields: psychology, behavioral science, medical science, and computer science, etc. We are in anticipation of future breakthrough in all these fields that shapes the growth of research.

References

[1] P. Ekman and W. Friesen, *Facial Action Coding System: Manual*, Palo Alto: Consulting Psychologist Press, 1978.

[2] P. Ekman, *Emotion in the Human Face,* Cambridge University Press, New York, 1982.

[3] M. Black and Y. Yacoob. "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion". *Int. Journal of Computer Vision*, 25(1), pp. 23-48, 1997.

[4] I. Essa and A. Pentland, "Coding, Analysis Interpretation, Recognition of Facial Expressions", *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 757-763, July 1997.

[5] Z. Zhang, M. Lyons, M. Schuster and S. Akamatsu. "Comparison Between Geometry-based and Gabor Wavelets-based Facial Expression Recognition Using Multi-layer Perceptron". *Proceedings of Int. Conf. On Automatic Face and Gesture Recognition*, 1998.

[6] J.A. Russell, "Core Affect and the Psychological Construction of emotion", Psychological Review, 110, pp. 145-172, 2003.

[7] H. Sebastian Seung and Daniel D. Lee, "The Manifold Ways of Perception**",** *Science,* vol. 290, pp.2268-2269, 2000.

[8] J.B. Tenebaum, W.T. Freeman, "Separating Style and Content with Bilinear Models", *Neural Computation Journal*, vol. 12, pp. 1247-1283, January 1999.

[9] A. O. Vasilescu, D. Terzopoulos, "Multilinear Subspace Analysis for Image Ensembles", *Proc. Computer Vision and Pattern Recognition Conf.* , Madison, WI, 2003.

[10] A. Elgammal and C. Lee, "Separating Style and Content on a Nonlinear Manifold", In *Proc. Computer Vision and Pattern Recognition Conf.* , Washington, 2004.

[11] R. Sumner and J. Popovic, “Deformation Transfer for Triangle Meshes”, In *Proceedings of SIGGRAPH*, 2004.

[12] J. Bourgain, “On Lipschitz Embedding of Finite Metric Spaces in Hilbert Space”, *Israel J. Math.*, vol. 52, nos. 1-2, pp. 46-52, 1985.

[13] W. Johnson and J. Lindenstrauss, “Extension of Lipschitz Mapping into a Hilbert Space”, *Contemporary Math.*, vol. 26, pp.189-206, 1984.

[14] G. Hristescu and M. Farach-Colton, “Cluster-Preserving Embedding of Proteins”, technical report, Rutgers Univ., Piscataway, New Jersey, 1999.

[15] M. Linial, N. Linial, N. Tishby, and G. Yona, “Global Self Organization of All Known Protein Sequences Reveals Inherent Biological Signatures”, *J. Molecular Biology*, vol. 268, no.2, pp. 539-556, May 1997.

[16] M. Pantic, L. Rothkrantz, “Automatic Analysis of Facial Expressions: The State of the Art”, *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol 22, No. 12, Dec 2000.

[17] B. Fasel, J. Luettin, “Automatic Facial Expression Analysis: a Survey”, *Pattern Recognition*, 36, 2003.

[18] A. Mehrabian, “Communication without Words,” Psychology Today, vol. 2, no. 4, pp. 53-56, 1968.

[19] T. Cootes, G. Edwards and C. Taylor. "Active Appearance Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):681-685, 2001.

[20] C.L. Huang and Y.M. Huang, “Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification,” J. Visual Comm. and Image Representation, vol. 8, no. 3, pp. 278-290, 1997.

[21] H. Kobayashi and F. Hara, “Facial Interaction between Animated 3D Face Robot and Human Beings,” Proc. Int'l Conf. Systems, Man, Cybernetics, pp. 3,732-3,737, 1997.

[22] M. Pantic and L.J.M. Rothkrantz, “Expert System for Automatic Analysis of Facial Expression,” Image and Vision Computing J., vol. 18, no. 11, pp. 881-905, 2000.

[23] H. Hong, H. Neven, and C. von der Malsburg, “Online Facial Expression Recognition Based on Personalized Galleries,” Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 354-359, 1998.

[24] S. Kimura and M. Yachida, “Facial Expression Recognition and Its Degree Estimation,” Proc. Computer Vision and Pattern Recognition, pp. 295-300, 1997.

[25] M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding Facial Expressions with Gabor Wavelets,” Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 200-205, 1998.

[26] M. Black, Y. Yacoob, “Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion” IJCV 25(1), pp. 23-48, 1997.

[27] Y. Tian, T. Kanade, J. Cohn, “Recognizing Action Units for Facial Expression Analysis” IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 23, NO. 2, Feb 2001.

[28] S. Gokturk, J. Bouguet, C. Tomasi and B. Girod, “Model-Based Face Tracking for View-Independent Facial Expression Recognition”. Proc. Int'l Conf. Automatic Face and Gesture Recognition, 2002.

[29] I. Essa and A. Pentland, “Coding, Analysis Interpretation, Recognition of Facial Expressions,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 757-763, July 1997.

[30] T. Kanade, J. Cohn, Y. Tian, “Comprehensive Database for Facial Expression Analysis,” Proc. Int'l Conf. Automatic Face and Gesture Recognition, 2000.

[31] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[32] H. Sebastian Seung and Daniel D. Lee, “The Manifold Ways of Perception**”,** *Science,* vol. 290, pp.2268-2269, 2000.

[33] J. Tenenbaum, V. de Silva and J. Langford. “A Global Geometric Framework for Nonlinear Dimensionality Reduction”. *Science*, vol. 290, pp. 2319-2323, 2000.

[34] S. T. Roweis and L. K. Saul. “Nonlinear Dimensionality Reduction by Locally Linear Embedding”. *Science*, vol. 290, pp. 2323--2326, 2000.

[35] A. Elad, R. Kimmel, “On bending invariant signatures for surfaces”, *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 25, Issue: 10, Oct. 2003, pp. 1285 – 1295.

[36] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyob, “Coding Facial Expressions with Gabor Wavelets”, *Proceedings of Int. Conf. On Automatic Face and Gesture Recognition*, 1998.

[37] Q. Wang, G. Xu and H. Ai. “Learning Object Intrinsic Structure for Robust Visual Tracking". *Proceedings of Int. Conf. on Computer Vision and Pattern Recognition*, June 16-22, 2003.

[38] K. Lee, J. Ho, M. Yang and D. Kriegman. “Video-based Face Recognition Using Probabilistic Appearance Manifolds”. *Proceedings of Int. Conf. on Computer Vision and Pattern Recognition*, June 16-22, 2003.

[39] M.-H. Yang. “Face Recognition Using Extended Isomap”. *International Conference on Image Processing, 2002.*

[40] Douglas Fidaleo and M. Trivedi, “Manifold analysis of facial gestures for face recognition”. *ACM SIGMM Multimedia Biometrics Methods and Application Workshop*, Nov. 8, 2003.

[41] Y. Chang, C. Hu, and M. Turk, "Manifold of facial expression," *Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, Oct. 17, 2003.

[42] Y. Chang, C. Hu, M. Turk, "Probabilistic expression analysis on manifolds," *Proceedings of Int. Conf. on Computer Vision and Pattern Recognition*, Washington DC, June 2004.

[43] C. Hu, Y. Chang, R. Feris, M. Turk, "Manifold based analysis of facial expression," *Proceedings of IEEE Workshop on Face Processing in Video*, Washington, D.C., June 2004.

[44] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 12, pp. 1615 – 1618, December, 2003.

[45] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces", In *Proceedings of SIGGRAPH 99*.

[46] M. Pantic, M.F. Valstar, R. Rademaker and L. Maat, ″Web-based database for facial expression analysis", *Proc. IEEE Int'l Conf. on Multimedia and Expo* (ICME'05), Amsterdam, The Netherlands, July 2005.

[47] K. Lee, J. Ho, M. Yang and D. Kriegman. "Video-based Face Recognition Using Probabilistic Appearance Manifolds". *Proceedings of Int. Conf. on Computer Vision and Pattern Recognition*, June 16-22, 2003.

[48] S. Zhou, V. Krueger and R. Chellappa. "Probabilistic Recognition of Human Faces from Video". *Computer Vision and Image Understanding,* 91(1), 2003.

[49] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, London, 1994.

[50] R.A. Horn, C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.

[51] S. Roweis, L. K. Saul, and G. E. Hinton, "Global Coordination of Local Linear Models", *Proceedings of Neural Information Processing System*, 2002.

[52] M. Brand, "Charting a Manifold", *Proceedings of Neural Information Processing System*, 2002.

[53] N. Linial, E. London, and Y. Rabinovich, "The Geometry of Graphs and Some of Its Algorithmic Applications", *Combinatorica*, vol. 15, pp. 215-245, 1995.

[54] A. Blake and M. Isard. "Active Contours". Springer-Verlag, Cambridge University press, 1998.

[55] M. Isard and A. Blake. “ICondensation: Unifying Low-level and High-level Tracking in a Stochastic Framework”. *Proceedings of European Conference on Computer Vision*, 1998.

[56] Q. Zhang, Z. Liu, B. Guo, H. Shum, “Geometry-Driven Photorealistic Facial Expression Synthesis”, *SIGGRAPH Symposium on Computer Animation*, 2003.

[57] M. B. Vieira, L. Velho, A. Sá and P.C. Carvalho, "A Camera-Projector System for Real-Time 3D Video", *IEEE International Workshop on Projector-Camera Systems (Procams)*, 2005.

[58] A. Sá, P.C. Carvalho and L. Velho, “(b,s)-BCSL: Structured Light Color Boundary Coding for 3D photography”, *Proceedings of 7th International Fall Workshop on Vision, Modeling, and Visualization*, 2002.

[59] www.nevengineering.com.

[60] T.W. Sederberg, and S.R. Parry, “Free-Form Deformation of Solid Geometric Models”, In *Proceedings of SIGGRAPH 86*, pp. 151-159.

[61] Y. Li, S. Gong, and H. Liddell. “Recognizing Trajectories of Facial Identities Using Kernel Discriminate Analysis”, *Proceedings of British Machine Vision Conference,* 2001.

[62] S.Z. Li, R. Xiao, Z. Li, and H. Zhang, “Nonlinear Mapping from Multi-View Face Patterns to a Gaussian Distribution in a Low Dimensional Space”, *Proceedings of IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS),* 2001.

[63] K. Scherer and P. Ekman, Handbook of Methods in Nonverbal Behavior Research. Cambridge, UK: Cambridge Univ. Press, 1982.

[64] Smith, E., Bartlett, M.S., amd Movellan, J.R. (2001). “Computer recognition of facial actions: An approach to co-articulation effects”. *Proc. of the 8th Joint Symposium on Neural Computation*.

[65] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior”, *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* 2006.

[66] Jacob Whitehill and Christian W. Omlin, "Local versus Global Segmentation for Facial Expression Recognition", *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* 2006.

[67] L. Yin and J. Loi and W. Xiong, "Facial Expression Representation and Recognition Based on Texture Augmentation and Topographic Masking", *Proc. of ACM Multimedia,* New York, NY, Oct., 2004.

[68] H. Chen, H. Chang, and T. Liu, "Local Discriminant Embedding and Its Variants", *Proceedings of Int. Conf. on Computer Vision and Pattern Recognition*, 2005.

[69] S. Yan, D. Xu, B. Zhang and H. Zhang, "Graph Embedding: A General Framework for Dimensionality Reduction", *Proceedings of Int. Conf. on Computer Vision and Pattern Recognition*, 2005.

[70] M. Brand, "Charting a Manifold", *Neural Information Processing Systems (NIPS)*, December 2002.

[71] Y. Chang, M. Vieira, M. Turk and L. Velho, Automatic 3D Facial Expression analysis in Videos, *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Beijing, 2005.

[72] http://marathon.csee.usf.edu/HumanID

[73] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew Rosato, "A 3D Facial Expression Database For Facial Behavior Research", *7th International Conference on Automatic Face and Gesture Recognition (FG2006)*, IEEE Computer Society TC PAMI. Southampton, UK, April 10-12 2006. p211-216.