School of Electrical Engineering and Computing
Department of Computing

Discriminant Feature Extraction and Selection for Person-independent Facial Expression Recognition

Mingliang Xue

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

February 2015

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

Mingliang Xue	Date

Abstract

Automatic facial expression recognition has attracted significant research effort since 1990s due to its potentially wide applicability. There are many existing work in this field that try to analyze human expressions based on either 2D images/videos or 3D images/videos. Although facial expression recognition seems natural and simple to humans, it is still a challenging task for computer. Due to the subtlety and variability of human facial expressions, the existing methods based on 2D images/videos still have lots of limitations, such as being sensitive to changes of recording conditions, and insufficient to represent easily-confused expressions etc. In fact, human face is not a convex solid, which means some of the out-of-plane deformations are hard to record by single-view 2D frontal view images/videos. In order to represent human facial expression sufficiently, 3D data based methods for facial expression analysis gains a rapid popularity ever since the availability of the relatively cheap 3D recording devices. No matter it is based on 2D data or 3D data, an ideal facial expression recognition system should be fully automatic, person-independent, and able to work with all the expressions etc. The existing work have focused on addressing different aspects of these points, but a system with all the good characteristics still needs more future research efforts. This thesis investigates the facial expression recognition problem with emphasis on: (1) enabling the expression recognition methods to be fully automatic, namely landmark detection and face alignment; (2) extracting discriminative features to represent human facial expressions, such as what colour space is best for expression recognition, and what parts/components of human face possess the maximal information of expressions; (3) improve the recognition performance on easily-confused expressions.

Firstly, the colour spaces for facial expression recognition are investigated in detail by this thesis. Specifically, the Uncorrelated Colour Space (UCS) and Discriminant Colour Space (DCS) are derived with the purpose of expression recognition, the performance are compared with RGB colour space and gray images on Oulu-CASIA NIR&VIS facial expression database and CurtinFaces database. For feature extraction on gray images, there are mainly two categories: geometric feature-based approach and appearance-based approach. Inspired by the studies from both sides, this thesis proposes a combination of these two kinds of features as a better face representation for facial expression recognition. In order to improve the performance of easily-confused expressions, i.e. anger and sadness, a two-tiered hierarchical classification is utilized, and different features are fed to SVMs classifier in each tier. The six prototypic expressions are not mutually exclusive, and

the experiments results show that the hierarchical classification can eliminate the major confusion caused by anger and sadness to a significant extent.

Secondly, the thesis proposes a fully automatic 3D static facial expression recognition method. Unlike the majority of the existing work which rely on manually annotated landmarks to align the faces or extract expression features, we try to achieve fully automatic recognition. However, landmark detection on 3D face is still an open problem. In order to process the 3D face automatically, 5 fiducial points (four eye corners and nose tip) are detected on the range images which are rendered from the raw 3D face point cloud. Then, the faces are aligned by iterative closest points (ICP) algorithm and local depth features are uniformly sampled around 25 heuristic points generated according to the detected 5 fiducial points. To compensate the misalignment of the heuristic points and remove the redundant features, mRMR (*minimal-redundancy-maximal-relevance*) feature selection is applied before classification. The proposed method achieves the best performance among existing automatic methods, and even comparable to those approaches which require human interfere.

Thirdly, the problem of recognizing facial expression based on 3D video sequence (dynamic 3D expression analysis) is addressed. Previous work address dynamic facial expression recognition as a time-series problem, and sequential models like Hidden Markov Models (HMMs) are trained based on the feature sequences that extracted frame-by-frame. However, facial expression is inherently a spatiotemporal process, frame-by-frame feature extraction may be insufficient to measure the expression dynamics. Instead, we propose to extract 3D-DCT features around 68 detected landmarks, which are real 4D features, to represent 3D facial expressions dynamics. This is followed by a two-round mRMR feature selection to reduce the feature dimension and improve the recognition performance. In addition, a method to identify the most discriminative facial parts/components for human expressions is presented. The identification is conducted on 4D expression data, with the HOG3D features extract from local depth patch-sequences. A hierarchical classification embedded with feature selection is utilized to pick the most discriminative facial parts out with the direct goal of maximizing recognition rates. The selection result shows mouth, cheeks, and eyebrow carry most of expression related information.

This thesis implements and evaluates feature extraction and selection methods for person-independent facial expression recognition. Experiments are conducted on several challenging, publicly-available databases and evaluated in terms of both automaticity and recognition accuracy. Results demonstrate that the proposed methods outperforms existing algorithms significantly, particularly in recognizing the easily-confused expressions and identifying the most discriminative facial components.

Acknowledgements

In many ways, this thesis could not have been possible without the contributions, great and small, direct and indirect, of many individuals over the course of the PhD.

Firstly, I would like to thank my supervisor, A/Professor Wanquan Liu. He provided an unending source of motivation and support across the full spectrum, from giving me the opportunity to do a PhD and suggesting a really interesting topic to continuous guidance and help in all aspects of my research. I'd like to thank my co-supervisor, A/Professor Ling Li, who spent a lot of time discussing the research direction with me and gave me a lot of inspiration when I felt depressed. I'd also like to thank my co-supervisor, A/Professor Ajmal Mian, for his tireless enthusiasm for my research and many hours that we spent discussing my work, helping me to solve the problem I met in my research.

Secondly, I would like to thank my parents who are always my strongest backing. I must thank Dr. Patrick Peursum who brought a happy atmosphere to our lab frequently, and gave me so much help in and out my research. I must also thank my classmates, Antoni Liang, Billy Li, Gongqi Lin, Jie Peng, Ke Fan, Xiang Xu, Xin Zhang, Xiaoming Chen, Yi Zhang, without whom the time during my PhD study would be inconceivable.

Finally, special thanks to the China Scholarship Council (CSC) and Curtin International Postgraduate Research Scholarships (CIPRS) for providing me with a scholarship without which I could not have continued to study.

Contents

List of Figures

List of Tables

Publications

This thesis is based upon several works that have been published (or submitted) over the course of the author's PhD, listed as follows in chronological order:

- Mingliang Xue, Ajmal Mian, Wanquan Liu, Ling Li (2015). Automatic 4D facial expression recognition using DCT features. *IEEE Winter Conference on Applications of Computer Vision (WACV)*.
- Mingliang Xue, Ajmal Mian, Wanquan Liu, Ling Li (2014). Fully automatic 3D facial expression recognition using local depth features. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1096-1103. IEEE, 2014.
- Mingliang Xue, Wanquan Liu, and Ling Li (2014). The Uncorrelated and Discriminant Colour Space for Facial Expression Recognition." *Optimization and Control Techniques and Applications*, pp. 167-177. Springer Berlin Heidelberg, 2014.
- Mingliang Xue, Wanquan Liu, Ling Li (2013). Person-independent facial expression recognition via hierarchical classification. *IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 449-454. IEEE, 2013.
- Mingliang Xue, Wanquan Liu, Xiaodong Liu (2013). "A novel weighted fuzzy LDA for face recognition using the genetic algorithm." *Neural Computing and Applications* 22, no. 7-8 (2013): 1531-1541.
- (under review) Mingliang Xue, Wanquan Liu, Ling Li, Ajmal Mian (2015). Automatic Selection of Discriminative Expression Features in 4D Data. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chapter 1

Introduction

Facial expressions provide one of the most powerful and natural means for human beings to communicate their emotions and intentions. Studies of facial expression started with psychological work where the general principles of expression and the meaning of expressions in both human and animals were established by Darwin (1872). In his treatise, Darwin grouped various kinds of expressions into several similar categories, and cataloged the facial deformations that occur for each category of expressions. Another important milestone is Ekman's cross-cultural study (Ekman and Friesen, 1971) on the existence of universal categories of emotional expressions, which comprises of the prototypic expressions: happiness, sadness, surprise, fear, anger and disgust. This work has a significant influence on the development of automatic facial expression recognition system. Facial expression analysis did not become a major field of study in computer science until the 1990s, though there are some works that predate this.

The pioneering work of Suwa *et al.* (1978) presented a system for analyzing facial expressions from a sequence of movie frames by tracking 20 points. Although this system was designed in 1978, the study of facial expression recognition did not continue in this direction until the 1990s. This can be seen from the survey paper by Samal and Iyengar (1992), which states that *"research in the analysis of facial expressions has not been actively pursued"*. The turning point was that the relatively cheap computational power started becoming available in the 1990s. This facilitated the development of robust face detection and face tracking algorithm, which is required by automatic facial expression recognition. Meanwhile, Human Computer Interaction (HCI), face recognition, affective computing, synthetic face animation as well as virtual reality started gaining popularity. Various potential applications in these areas produced a renewed interest in the development of automatic facial expression recognition systems.

Physically, facial expressions are caused by facial muscle movements, which result in temporary facial component displacement and deformation. No matter how natural and simple it seems to humans, recognition of facial expression is a complex and challenging task for computers. Research on automatic facial expression recognition based on 2D images or video frames has become increasingly active since the 1990s. Survey papers by Pantic and

Rothkrantz (2000), Fasel and Luettin (2003) and Bettadapura (2012) perform comprehensive studies of the published works based on 2D static images or dynamic sequences. Despite the fast development of the 2D image based systems, most of them are still highly sensitive to the recording conditions of images, such as illumination, occlusions and other changes in facial appearance like cosmetic products and facial hair. Moreover, it has been pointed out by Sandbach *et al.* (2012b) that 2D images or videos cannot capture out-of-plane changes of the facial surface. Advances in structural light scanning, 3D scanner system, and stereo photogrammetry have enabled the acquisition of 3D facial structure to be a feasible task. In this situation, 3D face data could be captured and analyzed. Ever since the public availability of 3D face datasets, a wide range of 3D facial expression recognition approaches (Danelakis *et al.*, 2014) have been developed in order to perform analysis on 3D static face images and dynamic sequences.

1.1 Problem Statement

Automatic facial expression recognition deals with the classification of facial component motion and facial feature deformation into several abstract expression classes that are purely based on visual information, such as static images or video sequences. It does not try to estimate the underlying emotional state since emotions are not the only source of human facial expressions.

To facilitate possible application, there are some features that an ideal facial expression recognition system must possess:

1. Fully automatic:
   (a) This means all of the stages of the facial expression analysis are to be performed automatically, including face detection / facial landmark detection, facial expression feature extraction, classification.
   (b) This also implies the designed system should have the capability to work with image or video feeds of different resolutions, illuminations, and poses. It could be able to handle the changes caused by facial hair, glasses, makeup etc.
2. Person-independent:
   (a) A feasible system should be able to analyze expressions of 'stranger' faces, which means the person being analyzed does not necessarily exist in the training gallery.

(b) The designed system should be able to work on people of various cultures and skin colours, and also be robust to age.

3. Real-time:

   (a) As human facial expression dynamic is a spatio-temporal process, the ideal system should be able to process the input image sequence or video in real-time.

   (b) For advanced human-computer interface, real-time performance is an essential requirement.

The goal of this thesis is to address these issues in facial expression recognition, with purpose to improve the recognition efficiency and accuracy of the current state-of-art techniques based on 2D/3D images or videos.

1.2 Limitations of Current Techniques

Although many works exist on automatic facial expression analysis, there are still limitations of current techniques. From the survey papers by Pantic and Rothkrantz (2000), Fasel and Luettin (2003), Bettadapura (2012), Sandbach *et al.* (2012b), and Danelakis *et al.* (2014), it is easy to see that different research groups have focused their efforts on different aspects of the features mentioned above. Consequently, the state-of-art systems have some limitations:

1. Some of the methods rely on manually labelled facial landmarks, either aligning the faces in the preprocessing stage or extracting expression features around the landmarks, especially in 3D data based methods. Obviously, this is not practical since it can not achieve automatic recognition. However, these methods still use manually labelled landmarks to analyze facial expression because the landmark detection on 3D face is still an open problem.

2. It is not an easy task to keep a facial expression recognition system performing in real time, mainly because the complexity of the face/landmark detection algorithm. Specifically, many feature extraction methods are time-costly, which renders systems built on these features significantly slower than real-time.

3. The six prototypic expressions are not mutually exclusive. For many state-of-art systems, happiness and surprise are easy to recognize, but some pairs (anger-sadness, disgust-fear) are easily confused.

4. Currently the majority of algorithms still extract facial expression features frame-by-frame when the system input is video or image sequences. However, facial expression is inherently a spatio-temporal process. Recognition in such time-series data requires that effective features extracted should be able to represent not only the deformation of facial features, but also the relative timing of facial actions as well as their temporal evolution. In other words, it is essential to measure the dynamics of facial expressions. Clearly, frame-by-frame feature extraction is insufficient.

1.3 Contributions and Significance

This thesis makes five main contribution to the field of facial expression recognition analysis — (1) investigation into a better colour space for facial expression recognition; (2) a hierarchical classification with the purpose of improving recognition performance for easily-confused expressions such as anger and sadness; (3) a fully automatic 3D facial expression recognition method based on local depth features; (4) an automatic 4D facial expression recognition method based on 3D videos; and (5) selection of the most discriminative facial parts/components for expression recognition.

1.3.1 Investigating Colour Spaces for Facial Expression Recognition

The current state-of-art 2D facial expression recognition techniques in Pantic and Rothkrantz (2000), Fasel and Luettin (2003), and Bettadapura (2012) are mostly based on gray-scale image features, with few making use of colour image features. Considering colour information should lead to better recognition performance, several works have been conducted on colour face data and demonstrate the effectiveness of colour information in facial expression recognition. If colour information does in fact help facial expression recognition, then it is important to determine what colour space is the most effective, given the purpose of representing and recognizing facial expressions.

In fact very little research has been done on this topic and current trials of using colour information in facial expression recognition, such as Lajevardi and Wu (2012), choose an existing colour space without any learning strategy. Motivated by the progress in face recognition, we aim to explore whether learning colour spaces would also be effective in facial expression recognition since both face recognition and facial expression recognition have similar engineering intuition. To this end, we derive the uncorrelated colour space (UCS) and discriminant colour space (DCS) for facial expression recognition purpose, test

them on Oulu-CASIA NIR&VIS facial expression database and CurtinFaces database, and compare against both the RGB and gray colour spaces. Results show that UCS can improve facial expression recognition performance, and DCS does not work so well as UCS in facial expression recognition.

1.3.2 Hierarchical Classification for Easily-confused Expressions

As mentioned, not all of the six prototypic expressions are easily distinguishable from each other, so the confusions caused by the easily-confused expressions will affect the recognition performance significantly. In order to improve recognition performance, the proposed method attempts to eliminate such confusions via a hierarchical classification. It has two advantages. Firstly, the hierarchical classification can pick the distinguishable expressions out in the first tier, and then focuses on the classification of easily-confused ones in the second tier. Secondly, the hierarchical structure enables us to utilize the most appropriate features for expression recognitions in each tier.

The hierarchical classification attempts to divide and conquer the recognition problem with a structure of two tiers. In the first tier, the easily-confused prototypic expressions are considered as one class and join the remaining expressions for classification. In the second tier, another classifier, which focuses only on the expressions in the merged class, is trained to separate the images of the merged class into the prototypic expressions. Two-tiered structure has the advantage that it allows us to use the appropriate features in each tier, especially in the easily-confused expressions separation. Results show that the selected mouth and eyebrow features used in the second tier of the classification improve performance significantly.

1.3.3 Fully Automatic 3D Facial Expression Recognition

For all practical applications, facial expression recognition should be fully automatic. Although landmark detection on 3D face models remains an open problem, it is inevitable in designing a fully automatic facial expression analysis system. Unlike most of the existing works relying on manually labelled landmarks, we propose a fully automatic method, including automatic detection of the fiducial points. From the detected fiducial points, additional heuristic points are generated via interpolation and extrapolation based on the structure of the human face. These heuristic fiducial points are chosen such that they fall on parts of the human face that are significant for expressions, specifically the mouth,

cheekbones and eyebrows.

Clearly, the heuristic fiducial points are not as accurate as manually labelled landmarks. Therefore, local features are then extracted from the area surrounding all fiducial points and feature selection is applied to account for variability in the point (mis)alignment. Experimental results show that the proposed method achieves the best performance among the existing automatic approaches, and even comparable to methods relying on manually labelled landmarks.

1.3.4 Automatic 4D Facial Expression Recognition

The majority of the existing work on 4D facial expression recognition treat dynamic facial expression recognition as a time-series problem, and sequential models like Hidden Markov Models (HMMs) are trained based on the feature sequences extracted frame-by-frame. Inspired by the success of discrete cosine transform in video compression, this thesis takes a different approach and applies 3D-DCT on the local depth patch-sequences generated from the original sequences based on automatic detected landmarks. The compact low-frequency 3D-DCT coefficients are selected as the feature vector, which can represent expression dynamics without loss of the subtle information.

The significance of our method is that the extracted 3D-DCT features are really 4D expression features, which are able to describe the spatiotemporal expression evolution. Moreover, we propose a two-phase feature selection process from patch-level to face-level to reduce the feature dimensionality and mitigate expression confusion. Experimental results show that the proposed method can preserve the subtle information conveyed by easily-confused expressions and outperforms other existing methods.

1.3.5 Identifying Discriminative Facial Components for Human Expressions

It has been noted that facial expressions are conveyed by different facial parts/components. Pardàs and Bonafonte (2002) show that the eyebrows and mouth are the components that carry the maximum amount of the information relevant to expressions, and Bourel *et al.* (2001) reveal that sadness is mainly conveyed by the mouth area. Similarly, Kotsia *et al.* (2008)'s study on the effect of occlusions on facial expression recognition shows that the occlusion of mouth reduces the recognition rate by more than 50%. These insights inspire

us to ask what parts/components of human face carry the information that can best distinguish the six basic expressions. In other words, what are the most expressive parts of human face that convey the most discriminative information for facial expression.

To answer this question, we propose an automatic method to determine the most discriminative facial parts for expression recognition based on 4D data. The 4D facial expression are represented by Histogram of Oriented 3D-Gradients (HOG3D), and and a two-phase feature selection process is conducted to select the most important parts with the direct goal of maximizing the recognition rates. The significance of this method is that it is a data-driven method which can be adjusted according to the different input images, and does not need manual interfere such as the artificially-induced occlusion approaches (Pardàs and Bonafonte, 2002; Bourel *et al.*, 2001; Kotsia *et al.*, 2008) which are mentioned above.

1.4 Structure of the Thesis

This thesis is organized as follows. In Chapter 2, a review of related work in the fields of automatic facial expression recognition is presented. The framework of a facial expression recognition system is first briefly described. This is followed by the comparison of 2D images and 3D images for facial expression recognition, and a introduction of popular benchmark databases in the research of expression recognition. Secondly, a review of the 2D and 3D facial expression feature extraction methods is presented. Finally, a brief summary is presented for this chapter.

In Chapter 3, the colour spaces for facial expression recognition are investigated in detail. Specifically, the Uncorrelated Colour Space and Discriminant Colour Space are derived with the purpose of expression recognition, the performance are compared with RGB and gray colour space on Oulu-CASIA NIR&VIS facial expression database and CurtinFaces database.

In Chapter 4, a 2-tierd hierarchical classifier focusing on the recognition of easily-confused expressions is implemented. The classification is based on LBP features extracted from 2D images. In recognition stage, two different SVMs are trained in each tier of the hierarchical classifier. Comparison with the existing methods, the proposed method can eliminate the confusion between anger and sad significantly.

In Chapter 5, a fully automatic 3D facial expression recognition method is proposed.

First, 5 fiducial points (four eye corners and nose tip) are detected on the range images rendered from the raw 3D point cloud. Then, the face is aligned by ICP and local depth features are uniformly sampled around the heuristic points generated according the five fiducial points. After feature selection, the selected features are fed to SVMs classifier to accomplish expression recognition. The performance achieved by the proposed method is the best among existing automatic methods, and also comparable to those approaches which require human interfere.

In Chapter 6, the problem of 4D facial expression recognition is addressed. The proposed method extracts 3D-DCT features around 68 detected landmarks, which are real 4D features, to represent 3D facial expressions dynamics. This is followed by a two-round mRMR (*minimal-redundancy-maximal-relevance*) feature selection to reduce the feature dimension and improve the recognition performance. The proposed method is tested by conducting 6-class recognition and 3-class recognition. In both cases, the proposed method outperforms other existing methods on the same database.

In Chapter 7, a method to identify the most discriminative facial parts/components for human expressions is presented. The identification is conducted on 4D expression data, with the HOG3D features extracted from local depth patch-sequences. A hierarchical classification embedded with feature selection is utilized to pick the most discriminative facial parts out with the direct goal of maximizing recognition rates. The selection result shows mouth, cheeks, and eyebrow carry most of expression related information.

Finally, Chapter 8 provides a summary of the thesis, as well as its contributions and potential future directions.

Chapter 2

Literature Review

The research goal in this thesis is to develop new facial expression recognition techniques based on 2D/3D images or videos, with the purpose to improve the recognition efficiency and accuracy of the current state-of-art. The recognition process should be automatic, person-independent. To achieve these goals, the scope of research in this thesis involves facial landmark detection and face alignment, feature extraction and selection, and classification. Facial landmark detection, in which the landmarks of important facial components are detected, is a fundamental requirement for face alignment. In most cases, it is inevitable to design an automatic system, since the landmarks are required by following feature extraction. Moreover, the feature extraction and selection is critical to achieve effective person-independent facial expression recognition, due to the subtlety and variability of facial expressions.

Facial expressions are generated by contractions of facial muscles, which results in displaced facial components (mouth corners, eye lids, eye brows, lips etc.) and temporally deformed facial surface (wrinkles and bulges). In order to analyze facial behavior, it is necessary to measure the location of facial actions, their intensity as well as their dynamics. Before 1977, most of the facial behavior researchers were relying on the human observers to observe the face of the subjects and give their analysis. However, such visual observation may not be reliable and accurate. Ekman and Friesen (1978) questioned the validity of such observations by pointing out that the observers may be influenced by context. For the same observations, different cultural groups may have different interpretations. To accurately measure facial expressions, Ekman and Friesen (1978) developed the comprehensive Facial Action Coding System (FACS) which has become the de-facto standard. This work is of significant importance and has a large influence on the development of automatic facial expression recognition system.

As illustrated in Fig 2.1, a facial expression recognition system generally consists of five main step: face acquisition, pre-processing, feature extraction, feature selection and classification. There is a wide variety of approaches to achieve facial recognition for different purposes and depending on different assumptions, especially the methods for feature extraction. In this chapter, we will review some classic approaches and analyze their

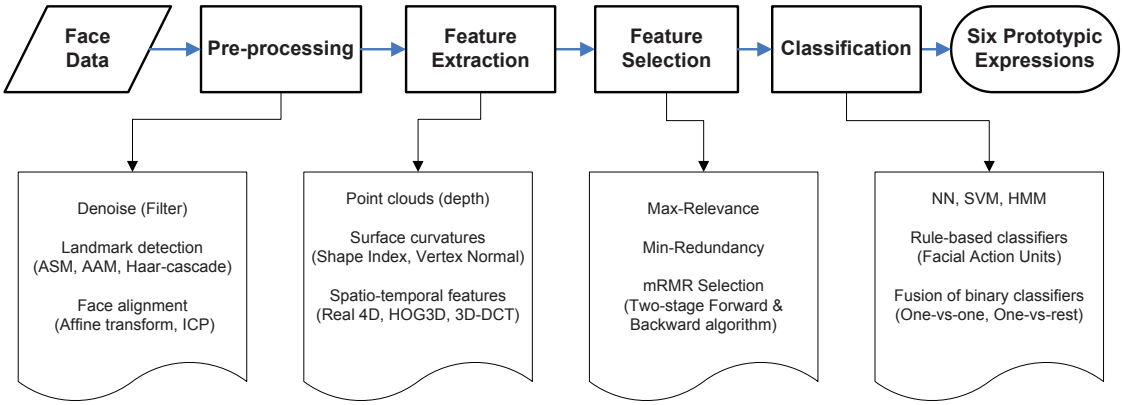


Figure 2.1: The structure of facial expression system.

advantages and disadvantages.

2.1 Expressive Face Acquisition

2.1.1 2D Facial Expression Databases

Currently, the feed of a facial expression recognition system is 2D/3D images or videos. At the beginning, the facial expression analysis approaches were proposed based on 2D images. In order to evaluate the performance of state-of-art algorithm, lots of 2D facial expression databases were constructed. The first one is the Japanese Female Facial Expression (JAFFE) Database, which contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The database was planned and assembled by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba at the Psychology Department in Kyushu University.

The most widely used 2D facial expression database is Cohn-Kanade AU-Coded Expression database. It is created for automatic facial image analysis and synthesis and for perceptual studies. Until now, there are two version available. The first version, referred to as Cohn-Kanade (CK) database, was released in 2000 for promoting research on detecting individual facial expressions. This version has several limitations, such as the lack of validated emotion label and common performance metric to evaluate new algorithms. As a consequence, the second version, referred to as the Extended Cohn-Kanade (CK+)

database (Lucey *et al.*, 2010), was created to address these concerns. It includes both posed and non-posed (spontaneous) expressions and additional types of metadata. As with the initial release, the target expression for each sequence is fully FACS coded. In addition validated emotion labels have been added to the metadata. Thus, sequences may be analyzed for both action units and prototypic expressions. Additionally, CK+ provides protocols and baseline results for facial feature tracking and action unit and emotion recognition.

There are some other database which constructed with different purpose, for example, the MMI Facial Expression Database (Pantic *et al.*, 2005) which aims to deliver large volumes of visual data of facial expressions to the facial expression analysis community. The database consists of over 2900 videos and high-resolution still images of 75 subjects. It is fully annotated for the presence of AUs in videos, and partially coded on frame-level, indicating for each frame whether an AU is in either the neutral, onset, apex or offset phase. Another popular database is the Oulu-CASIA NIR&VIS facial expression database (Zhao *et al.*, 2011), which contains videos with the six prototypic expressions from 80 subjects captured with two imaging systems, NIR (Near Infrared) and VIS (Visible light), under three different illumination conditions: normal indoor illumination, weak illumination (only computer display is on) and dark illumination (all lights are off).

2.1.2 3D/4D Facial Expression Databases

Recent advances in stereo photogrammetry and structure light scanning have made the acquisition of 3D facial structure and deformation a feasible task. The first 3D facial database that collected for facial expression recognition is BU-3DFE dataset (Yin *et al.*, 2006a), example of which can be seen in Figure 2.2. It contains static 3D facial models of 100 subjects, displaying the six prototypic expressions at four different intensity levels. The faces were captured by a 3D face imaging system (3DMD digitizer). The database was released with a set of metadata including the position of 83 facial landmarks on each facial models.

Human facial expression is inherently a spatio-temporal process, which means the static facial model is insufficient to represent facial expressions. As a consequence, BU-4DFE database (Yin *et al.*, 2008), consisting of 4D faces (3D faces changing over time), is recorded using the DI3D dynamic face capturing system. As shown in Figure 2.3, it contains sequences of the six prototypic facial expressions with each sequence lasting approximately 4 seconds. Similar as BU-3DFE dataset, each facial model of the sequence is released with 83 facial landmarks.

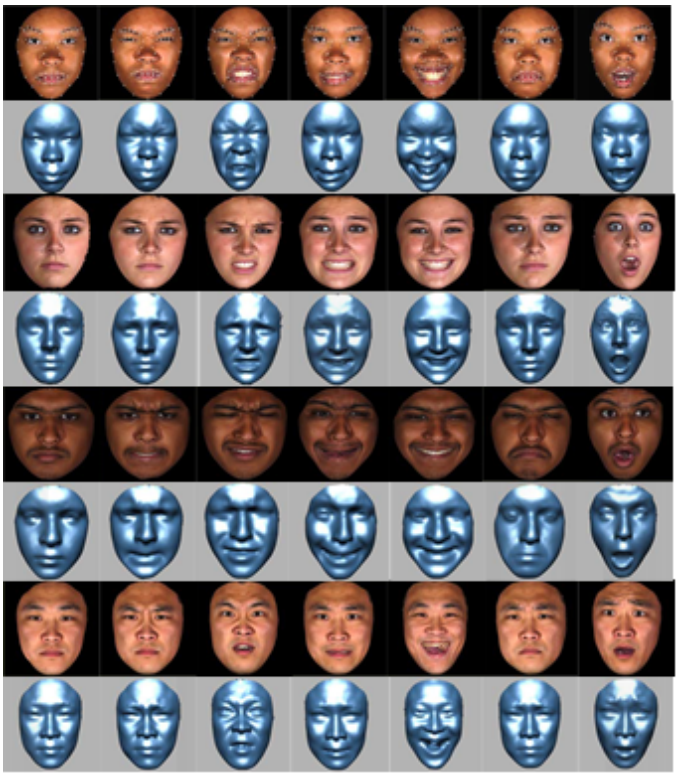


Figure 2.2: The demo images from BU-3DFE database.

2.2 2D-image/video Based Methods

There are roughly two types of approaches to extract facial features from the original 2D face images: geometric feature-based approach and appearance-based approach. Geometric feature-based methods attempt to encode the shape and location of facial components such as mouth, eyebrows, and cheeks. The facial components or facial feature points are extracted to form a vector that represents the facial geometry. Meanwhile, the appearance-based methods aim to capture the appearance changes caused by facial expressions, on either the holistic face or local regions.

2.2.1 Geometric Feature-based Methods

Generally, geometric feature-based methods detect or track the changes of facial component via a bunch of landmarks. Tian *et al.* (2001) implement a multi-state face and facial component models for tracking and modeling the various facial features, including lips, eyes, brows, cheeks, and furrows. Given an image sequence, the location of face and facial

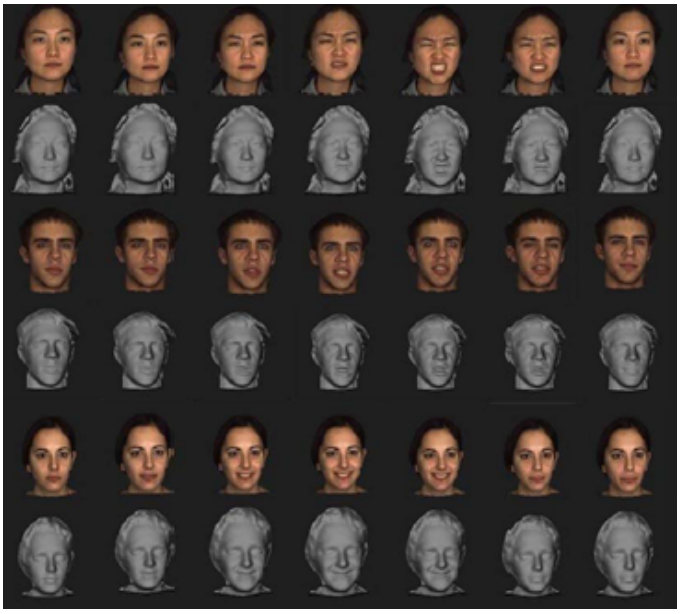


Figure 2.3: The demo images from BU-4DFE database.

features are detected automatically in the initial frame (Rowley *et al.*, 1998) and tracked in the following sequence. During tracking, detailed parametric descriptions of the facial features are extracted to describe shape, motion, and state of facial components. With these parameters as the inputs, a group of action units (neutral expression, six upper face AUs and 10 lower face AUs) are recognized whether they occur alone or in combinations.

Automatic active appearance model (AAM) is another method that widely adapted for facial landmark detection and tracking (Cootes *et al.*, 2001; Matthews and Baker, 2004; Xiao *et al.*, 2004). Cheon and Kim (2009) propose a natural facial expression recognition method that recognizes a sequence of dynamic facial expression images using the differential AAM and manifold learning as follows. First, the differential-AAM features (DAFs) are computed by the difference of the AAM parameters between a target face image and a reference image. Second, manifold learning embeds the DAFs into a continuous feature space. Finally, the distances between the input image sequence and gallery image sequences are computed in terms of directed Hausdorff distance (DHD) and then the expression by a majority voting of k-nearest neighbors (k-NN) sequences in the gallery are selected as the recognition results.

Valstar *et al.* (2005) propose a method to detect 16 facial action units using features calculated from tracked facial point data. The facial points are tracked by an improved version of Particle Filtering with Factorized Likelihoods (PFFL) (Patras and Pantic, 2004)

and spatio-temporal relations between tracked points are then used to represent action units. Then, the action units displayed in a video are classified by probabilistic actively learned SVMs.

The explicit 3D wire frame face mode (Essa and Pentland, 1997; Tao and Huang, 1999) is another type of model to track geometric facial features. The 3D wire frame model is fitted to the first frame of expression sequence by manually selecting fiducial points such as eye corners, mouth corners, and nose tip. The generic face model is warped to fit the facial features around fiducial points to analysis facial expressions.

2.2.2 Appearance-based Methods

Appearance-based features represent the texture changes of face images with different expressions. In order to process digital images, plenty of texture descriptors are proposed to analyze digital images with variety of technical purposes, such as noise filtering, argumentation, segmentation etc. The popular descriptors, such as Gabor wavelets (Daugman, 1988), Local Binary Pattern (LBP) (Ojala *et al.*, 1996, 2002), Histogram of Orientated Gradient (HoG) (Dalal and Triggs, 2005) are also widely used for facial expression analysis.

Gabor filters are widely used for facial expression recognition. The facial appearance changes are encoded by a multi-scale and multi-orientation set of Gabor filters. The Gabor filter may be applied to aligned local regions of a face (Lyons *et al.*, 1998, 1999; Tian *et al.*, 2002; Zhang *et al.*, 1998) or to the whole face (Bartlett *et al.*, 2001; Donato *et al.*, 1999; Littlewort *et al.*, 2002). In the work done by Lyons *et al.* (1999), facial expression images are encoded by a set of Gabor filters, and a grid is registered by manual labelled fiducial points. The amplitude of the Gabor transform coefficients are then sampled on the grid to form a feature vector, i.e. Labeled Graph Vector (LGV). The distances of the LGV from each facial expression cluster center are utilized for recognition. Donato *et al.* (1999) compare several techniques for recognizing facial action units on whole face image, including optical flow, local feature analysis, principal component analysis, independent component analysis and Gabor wavelet representation. The best performance were obtained using a Gabor wavelet representation and independent component analysis. It is worth noting that all the methods like in in Zhang *et al.* (1998) and Donato *et al.* (1999) have the manual stage for face alignment.

In the comprehensive study done by Shan *et al.* (2009), LBP feature for person-independent facial expression recognition is investigated empirically. Due to its tolerance against illumination variations and low computing cost, plenty of works use LBP features to represent

facial expression and then different machine learning methods, including SVM, template matching etc., are used to perform classification. Compared to Gabor wavelets, LBP features can be derived efficiently in low-dimension, while keeping characteristic expression information. Using a same classifier, LBP-based SVMs achieve slightly better performance when compared to Bartlett *et al.* (2005)' work which uses the Gabor-wavelet based SVMs.

During the past decade, HoG features (Dalal and Triggs, 2005) have received increasing research attention for the purpose of object detection. As an effective texture descriptor, it is also adapted for expression presentation. In Dahmane and Meunier (2011)'s work, HoG are used to extract the appearance features by accumulating the gradient magnitudes for a set of orientations in 1-D histograms defined over a size-adaptive dense grid, and Support Vector Machines with Radial Basis Function kernels are the base learners of emotions. Another example is presented by Orrite *et al.* (2009), in which a hierarchical decision tree is built using a bottom-up strategy by recursively clustering and merging the classes at each level. For each branch of the tree, a list of potentially discriminative HoG features is built by applying the log-likelihood maps to interest locations. This method could recognize expression states which are not present in the training set when tested on Cohn-Kanade facial expression database.

2.2.3 Geometric Feature-based vs Appearance-based Methods

The authors of Tian *et al.* (2002) and Zhang *et al.* (1998) argue that appearance features are better than geometric features, because geometric features are more sensitive to inaccurate image alignment. In addition, Lucey et al. Lucey *et al.* (2010) showed that appearance information is more important to the recognition of anger, sadness and fear. However, with the recent development on face alignment and facial feature tracking, an increasing number of expression analysis algorithms are based on geometric features. Valstar *et al.* (2005) presented a method that can detect facial action units effectively by classifying features derived from the tracked facial landmarks. They argued that the geometric features is well suited for facial expression analysis, especially with facial feature tracking. The studies from both sides indicate a possible combination of these two kinds of features as a better face representation for facial expression recognition, for example, Tian *et al.* (2002) improve the recognition rate of all facial action units by combining Gabor-wavelet coefficients and geometric features. This thesis combines the geometric features and LBP features to represent facial expressions, and achieved a significant improvement when compared to the results which use single type of features.

2.3 3D-image/video Based Method

Facial expressions reflex not only facial feature points movement but also skin surface deformation, which means the location, distance and movement of the landmarks as well as the surface shape can be used to represent expressions. A wide range of 3D facial expression recognition systems have been designed in order to perform analysis on static face models and dynamic facial image sequences. The majority of systems developed have attempted recognizing of expressions from static 3D facial expression data. However, more recent works employ dynamic 3D facial expression data for this purpose.

2.3.1 Static analysis

Several methods have been developed for the analysis of static 3D facial expressions, which use a range of features to distinguish different expressions. According the extraction method, these features can be categorized into: distance-based features, patch-based features and morphable models.

Distance-based Methods. One of the most popular methods for feature extraction in 3D static faces is the use of distances between certain facial landmarks, from which the changes caused by expression are calculated. This is similar with geometric 2D methods that track fiducial points on the face. As BU-3DFE database provides 83 facial landmark points on each of the 3D face models. These manually labelled points, as well as the distances between them, have been widely utilized for static facial expression analysis. The method developed by Soyel and Demirel (2007) uses six characteristic distances extracted from the 11 facial feature points, achieving an average recognition rate of 91.3%. Another example of using facial points in the BU-3DFE is the work done by Tang and Huang (2008). The distances between landmark points are normalized by Facial Animation Parameter Units (FAPUs). In addition, the slope of the lines connecting these points are used as an additional set of features after being divided by their norms to produce unit vectors. This method achieves an average recognition of 95.1%. Srivastava and Roy (2009) proposed to use residues as features, in which both the magnitude and direction of the displacement of the landmark points in the BU-3DFE database are encoded. A feature matrix is then formed by concatenating the different matrices in each of the three spatial directions in order to form one 2D matrix. The average rate of 91.7% is achieved by this method. Moreover, Sha *et al.* (2011) extract features by calculating the distances among all pairs of available landmark points and the surface curvature at each point in the mesh. The face was divided into triangles using a subset of the given facial landmarks, and histogram

were accumulated for each triangle of the surface curvature types. This approach obtains an average recognition rate of 83.5%.

Patch-based Methods. Facial surface patches are widely employed for feature extraction in expression recognition, because they reflect the deformation caused by expressions. The shape information of small local patches are extracted to represent facial expressions, surrounding either landmark points (Maalej *et al.*, 2010, 2011; Lemaire *et al.*, 2011), or every point in the mesh (Wang *et al.*, 2006). The curvature information is used by Wang *et al.* (2006), who fitted a polynomial patch to the local surface at each point in the mesh. The curvature features of the patches are labelled according to primitives, and achieve an average rate of 83.6%. Alternatively, patches around landmark points in the 3D mesh could be used for feature extraction. Lemaire *et al.* (2011) define patches around landmarks on facial point cloud via the fitting of Statistical Facial Feature Model (SFAM), in which three types of the variations such as shape, intensity and range value are combined linearly. (Maalej *et al.*, 2010, 2011) also found patches around landmarks in 3D mesh. The curves surrounding these points are defined on the patches, and the square root velocity function (SRVF) is calculated to capture the shape of curve. The geodesic distances between curves are computed to represent the dissimilarity, which is summed to represent the differences patches. This method achieves an average recognition rate up to 98.8% on the BU-3DFE database.

Morphable model-based Methods. An alternative approach for feature extraction is the use of morphable models. Different implementations of morphable models have been utilized to model identity, expressions, or in most cases both kinds of variations. The Statistical Facial Feature Model (SFAM) was employed as one type of morphable model by Zhao *et al.* (2010). The model was fitted to the target meshes, and the parameters of the fitting are used to extract features. The intensity and range values are used in fitting process directly, while the mean of the shape parameters is subtracted to extract displacement features. In addition, the shape index is calculated from fitting parameters, and then are encoded by local binary patterns to provide further descriptors. This approach achieves average recognition rate of 87.2% and 82.3% on the BU-3DFE database using manually labelled landmarks and automatically selected landmarks respectively. The Morphable Expression Model (MEM) used by Ramanathan *et al.* (2006) is able to model a range of different expressions for a specific individual. The corresponding points on the expressive faces of a particular subject are identified first by minimizing the value of energy function between points. Then the MEM is created based on the principal components of the expressive faces of one subject, and reconstruct a new face by performing a weighted summation of these eigne-expressions. This method achieves an average expression recognition rate of 97.0% over a custom database containing neutral faces and three expressions: happy, sad

and angry. Finally, an elastically morphable bilinear 3D model is employed by Mpiperis *et al.* (2008b). This deformable model captures variations in both identity and expression. The model was fitted to the point cloud via landmarks which are identified on both the MEM and points cloud. Once the correspondence has been established, Principal Component Analysis (PCA) was applied to find the principal components of the basic mesh deformation, i.e. eigen-meshes. Then, 3D face was modelled by a bilinear model based on these eigen-meshes, which facilities the classification of both identity and expressions. For facial expression recognition, the facial features are extracted and represented during the model fitting. This approach achieves an average recognition rates of 90.5% on the BU-3DFE database. This model is also adapted by some other works (Mpiperis *et al.*, 2008a, 2009), though the optimal parameters are obtained by different methods.

2.3.2 Dynamic analysis

Instead of using single or multiple static 3D images for expression recognition, some work has begun to utilise 3D image videos/image sequences for analysis of facial expression dynamics, especially since the release of BU-4DFE database. According to how to extract expression features from 3D image sequences, the dynamic facial expression analysis methods can be divided into two major groups: motion-based method and deformation-based methods.

Motion-based Method. The motion-based methods try to extract facial expression by tracking landmarks or critical points on 3D frames. In landmark tracking case, the local region are tracked around specific facial landmarks along 3D frames and detect temporal changes on their geometry characteristics using features such as invariant statistical moments or mesh curvatures. While in critical points tracking case, the key points are tracked along time and detect temporal changes on spatial characteristics based on these points such as distances, angles etc. The difference between these two cases is that in first case, the expression descriptors are constructed on facial regions around landmarks, while in the second case, only facial points are considered.

A typical work of landmark tracking based method is presented by Chang *et al.* (2005). A 2D semi-manual tracker is employed to track 22 landmarks with the help of wrapped mesh model projection after fitted on 3D videos data. The depth of the vertex is recovered by minimizing the distance between the model and the range data. Lipschitz embedding (Bourgain, 1985) is utilized to normalize deformation of the standard model which could be embed in a low dimensional manifold. At last, a probabilistic expression model is learned on the manifold to accomplish the expression classification. In the work

done by Tsalakanidou and Malassiotis (2009), an Active Shape Model (ASM) (Lanitis *et al.*, 1997) is implemented with 81 selected facial landmarks. The ASM is fitted to the 3D face data using the gradient feature in the neighborhood of every landmark. With the help of FACS, the extracted feature vector combines the geometric, curvature and appearance information around landmarks. A rule-based classifier is then defined for the classification of expressions and action units. Similarly, Sun *et al.* (2008) construct an Active Appearance Model (AAM) (Cootes *et al.*, 2001) to track 83 landmarks on 3D videos. Each registered vertex is assigned with one of eight primitive surface labels according to its principal curvature. Then, a set of HMMs are used for classification.

As concern to critical points tracking-based method, Berretti *et al.* (2012) propose a method based on selected key points on the nose, eyes and mouth. The facial expression conveyed in each 3D frames is represented by the distances between these points. The distances are normalized by the inner eye separation to remove the identity-related face structure information, and are then used to train a sophisticated HMM for final classification. Another example of using critical points are present by Jeni *et al.* (2012). The critical points are estimated on each 3D frame using Constrained Local Models (CLM). The normalized difference between the current shape on target expression frame and the neutral frame are used to train a SVM classifier for recognition.

Deformation-based Method. Facial deformation methods attempt to detect temporal deformation using a generic 3D face model, which has been explored in several papers. One of the first works on analysis of facial expression dynamics is proposed by Sun and Yin (2008), in which the deformable range model is adapted to each frame in the image, and its changes were tracked in order to extract geometric features. This approach achieved an average expression recognition rate of 90.4% when tested on the BU-4DFE database. Yin *et al.* (2006b) treat human face as a 3D time varying wave and propose a tracking model to estimate motion trajectories. Based on this model, a spatio-temporal descriptor, i.e. Facial Expression Label Map (FELM), is proposed. The tracking model is aligned by Iterative Closet Points (ICP), and then deformed to fit the target scan by minimizing an energy function. The combination of FELM vector and motion vector is used to represent facial expressions for classification. Another example is Sandbach *et al.* (2011)'s work which uses Free Form Deformations (FFD) (Rueckert *et al.*, 1999) to align faces and find a vector field for facial motion representation. After the frame is divided into regions by quad-tree decomposition, three types of features are extracted in each region: the distribution of vector directions, the magnitude of the motion, and the divergence and curl of the vector field. Finally, features from all the regions are concatenated as feature vector of each frame to train a HMM model for classification. Moreover, Reale *et al.* (2013) propose a real 4D feature, i.e. "Nebula", to improve expression and facial action analysis

performance. Unlike the majority of the dynamic analysis methods which extract feature frame-by-frame, this method extract expression features directly on spatio-temporal volume. The volume data is voxelized and fit to a cubic polynomial. A label is assigned based on the principal curvatures, and the angles of the least curvature are calculated. The labels angles for each feature are used to accumulate a histogram for each volume. Finally, the histograms are concatenated from all the volumes for expression representation and classification.

2.3.3 Comparison of Static and Dynamic Analysis

By comparing these three types of static analysis methods, it is easy to see that distance-based methods only use landmarks in feature extraction. The computing cost of this type of method is lower than patch-based method. However, distance-based methods may be insufficient for expression representation since facial expressions are not only reflected by facial component movement, but also the facial surface deformation. Meanwhile, the morphable model-based methods seem to be flexible and accurate enough to capture facial changes caused by expressions, but model fitting normally relies on an optimal process such as minimizing energy function. This process is much slower than real-time and also have the local minimum issues. This thesis propose a patch-based method for static 3D facial expression recognition, including fiducial points detection, patch-based depth feature extraction and feature selection to compensate the misalignment of the generated heuristic landmarks.

In general, most of the dynamic analysis methods address the facial expression recognition as a time-series problem. The frame-by-frame extracted features are used to train a sequential models like Hidden Markov Models (HMMs) for expression classification. However, facial expression is inherently a spatio-temporal process, frame-by-frame extracted features may be insufficient to capture the expression dynamics. Alternatively, the real 4D feature like "Nubula" is more suitable for spatio-temporal feature extraction. In this thesis, we also propose to extract real 4D features, such as 3D-DCT and HOG3D, to represent 3D facial dynamics rather than extract features from discrete frames.

2.4 Chapter Summary

This chapter has presented a review of existing works that are relevant to this thesis. It begins with a framework of facial expression recognition system and the FACS for expression

measurement, and introduction the benchmark database for expression recognition. We mainly focuses on the feature extraction and expression representation method, from 2D image/video based method to 3D image/video based method. For 2D image/video based method, the advantage and disadvantage of geometric feature-based and appearance-based methods are discussed, and the possible combination of these two types of feature are considered for improving the recognition performance. Then, the approaches based on 3D face data are reviewed. For 3D static facial expression analysis, three type of features are discussed and compared, especially the patch-based method which is most relevant to this thesis. The 3D dynamic expression analysis method are reviewed at last, with the brief overview of one very related work on real 4D feature extraction (Reale *et al.*, 2013). We also aim to extract spatio-temporal feature for 3D facial expression dynamic representation.

Chapter 3

Colour Space Selection for Facial Expression Recognition

Each of the existing clour spaces is proposed with a specific purpose, like RGB colour space is to describe what kind of light needs to be emitted to produce a given colour, CMYK colour space is to describe what kind of inks need to be applied in printing system so the light reflected from the substrate and through the inks produces a given color. However, the current state-of-art facial expression recognition techniques are mostly based on gray-scale image features (Pantic and Rothkrantz, 2000; Fasel and Luettin, 2003; Sandbach *et al.*, 2012b), with rarely considering colour image features. Considering the fact that different colour channels provide more information, an appropriate use of colour information may lead to better recognition performance.

Several existing results reveal that colour provides useful information for face recognition (Rajapakse *et al.*, 2004; Jones and Abbott, 2006). These work address the problem that how to extract colour features for face recognition. In order to seek a theoretically meaningful justification of colour features for face recognition, Yang and Liu (2008) proposed a discriminant colour space (DCS) for face representation and verification using discriminant analysis, while Liu (2008) derived a uncorrelated colour space (UCS) by applying principal component analysis to decorrelate the R, G and B component images. Their experimental results show that the learned colour spaces, i.e. UCS and DCS, can achieve better face recognition performance than the commonly used RGB colour space.

As concern to colour facial expression recognition, the same question comes out: which colour space is the most effective with the purpose of representing and recognizing facial expression? In fact very few research has been done on this topic and the current trials of using colour information in facial expression recognition, such as Lajevardi and Wu (2012), choose an existing colour space without learning strategy. Motivated by the research progress in face recognition that the UCS and DCS show potential performance improvement, we aim to explore in this chapter that whether these learning color spaces are also effective in facial expression recognition since both face recognition and facial recognition have similar engineering intuition. The uncorrelated and discriminant colour

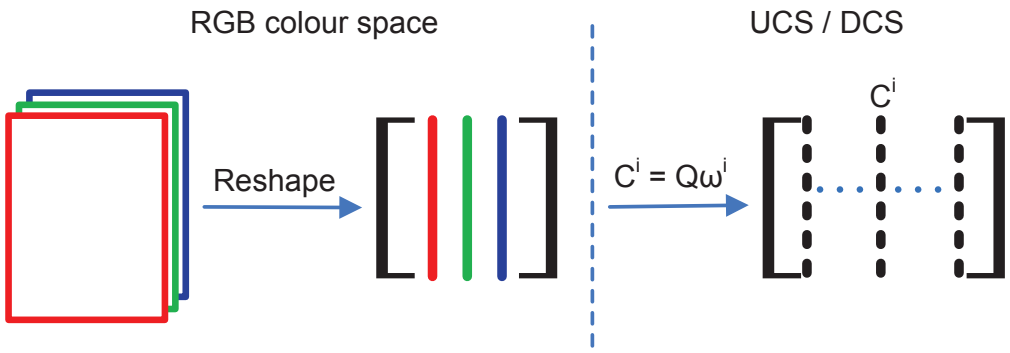


Figure 3.1: Learning colour space for facial expression recognition.

spaces are derived for facial expression recognition purpose, and tested them on Oulu-CASIA NIR&VIS facial expression database and CurtinFaces database. Some conclusions and possible research directions are given.

3.1 Learning Colour Space for FER

This chapter focuses on learning colour space to extract colour features, which is one kind of pre-processing. After pre-processing step, the face images are converted from RGB colour space into a new colour space which is learned for the expression recognition purpose. The image features are then extracted in this learned colour space for expression representation and classification.

Originally, the face images are represented in the fundamental RGB colour space, from which a number of other colour spaces are generated. Suppose $Q_{m\times n}$ is a colour image, and each of its three colour components is size of $m \times n$, we can reshape them into column vectors: $R, G, B \in R^d$, where $d = m \times n$. Consequently, the colour image can be represented by a $d \times 3$ matrix: $Q = [R\ G\ B] \in R^{d\times 3}$. Given a specific recognition task, either face recognition or facial expression recognition, the goal of learning colour space is to seek the combinations of the R, G and B colour components that can best represent colour information for the recognition purpose. Specifically, the combination can be denoted as

$$C = Q\omega = \omega^1 \cdot R + \omega^2 \cdot G + \omega^3 \cdot B \tag{3.1}$$

where $\omega = [\omega^1\ \omega^2\ \omega^3]^T$ is the weight vector, as illustrated in Figure 3.1. Thus, the task is to find the optimal weights so that C is the best representation of the image Q in term of a given criterion, such as the criterion of principal component analysis (PCA) or linear

discriminant analysis (LDA). The following section will introduce these two criterions and then give the details about how to derive the UCS and DCS.

3.2 Uncorrelated Colour Space

3.2.1 Principal Component Analysis

Principal component analysis (PCA), which was proposed by Person (1901), is mostly used as a tool in exploratory data analysis and for making predictive models. Depending on the field of application, it is also named the discrete KarhunenLove transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, eigenvalue decomposition (EVD) of XTX in linear algebra. It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables, which is the reason why PCA is also well-known as a dimension reduction method. Given a dataset of N samples $\{x_1, x_2, ..., x_N\} \subset R^n$, let us consider a linear transform that can project the original data in n-dimensional space into an m-dimensional feature space, where $m < n$. The sample x_i after projecting could be defined as

$$y_i = W^T x_i, \quad i = 1, 2, \cdots, N \tag{3.2}$$

where $W \in R^{n \times m}$ is the project matrix with orthonormal columns. In original space, the total scatter matrix is defined as

$$S_T = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T. \tag{3.3}$$

where μ is the mean of all the samples in the dataset. After projection, the total scatter matrix of $\{y_1, y_2, ..., y_N\} \subset R^m$ could be defined as $W^T S_T W$. PCA aims to maximize the determinant of the total scatter matrix of the projected samples by choosing a optimal projection matrix W_{opt}. This could be achieved by applying eigen-decomposition to the total scatter matrix S_T:

$$\begin{aligned} W_{opt} &= \arg\max_W \left| W^T S_T W \right| \\ &= [w_1 \;\; w_2 \;\; \cdots \;\; w_m] \end{aligned} \tag{3.4}$$

where $w_i, i = 1, 2, \cdots, m$ is the eigenvectors of S_T corresponding to the m largest eigenvalues. Then, the uncorrelated principal components could be derived based on these these eigenvectors.

3.2.2 Derivation of Uncorrelated Colour Space

The uncorrelated colour space (UCS) is derived from the RGB colour space via using principal component analysis to decorrelate the R, G and B colour components. Let us consider a set of N sample images $Q = \{Q_1, Q_2, ..., Q_N\} \subset R^{d\times 3}$ in RGB colour space, and a unitary column vector ω. Suppose the uncorrelated colour representation of $Q_i = [R_i\ G_i\ B_i]$ is given by

$$U_i = \omega^1 \cdot R_i + \omega^2 \cdot G_i + \omega^3 \cdot B_i = Q_i\omega \tag{3.5}$$

After converting into the uncorrelated colour space, the covariance matrix of the sample images $U = \{U_1, U_2, ..., U_N\}$ can be formulated as

$$\begin{aligned} S_T &= E[(U - EU)(U - EU)^T] \\ &= E[(Q\omega - EQ\omega)(Q\omega - EQ\omega)^T] \\ &= E[(Q - EQ)\omega][(Q - EQ)\omega)]^T \end{aligned} \tag{3.6}$$

where $E[\cdot]$ is the expectation operator. The principal component analysis criterion (Yang *et al.*, 2004) is given by

$$J(\omega) = tr(S_T) = \omega^T[E(Q - EQ)^T(Q - EQ)]\omega \tag{3.7}$$

By defining the *colour space scatter matrix*

$$L_t = E[(Q - EQ)^T(Q - EQ)] \tag{3.8}$$

the criterion can be rewritten by

$$J(\omega) = \omega^T L_t \omega \tag{3.9}$$

where ω is a unitary column vector. The ω that maximizes this criterion is the optimal weights for the UCS. Actually, ω is the eignvector of L_t. Since the colour space scatter matrix L_t is a 3×3 matrix, the uncorrelated colour space is defined by the transformation

$$[U^1\ U^2\ U^3] = [R\ G\ B] \begin{bmatrix} \omega_1^1 & \omega_2^1 & \omega_3^1 \\ \omega_1^2 & \omega_2^2 & \omega_3^2 \\ \omega_1^3 & \omega_2^3 & \omega_3^3 \end{bmatrix} = [R\ G\ B][\omega_1\ \omega_2\ \omega_3] \tag{3.10}$$

3.3 Discriminant Colour Space

3.3.1 Linear Discriminant Analysis

When the data samples $\{x_1, x_2, ..., x_N\} \subset R^n$ are given with class labels, it is better to the class label information to build a more effective method for linear transform and feature

reduction. Suppose the each samples belongs to one of the c classes $\{L_1, L_2, ..., L_c\}$, Linear discriminant analysis (LDA) (Fisher, 1936) calculates the between-class scatter S_B and within-class scatter S_W according to the class label information as follows

$$\begin{aligned} S_B &= \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \\ S_W &= \sum_{i=1}^{c} \sum_{x_j \in L_i} (x_j - \mu_i)(x_j - \mu_i)^T \end{aligned} \tag{3.11}$$

where the N_i is the number of samples in class L_i, and μ_i is the class center of L_i. The optimal projection matrix W_{opt} is chosen as the martix with orthonomal columns which maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix in the projected subspace, as follows

$$\begin{aligned} W_{opt} &= \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|}| \\ &= [w_1 \;\; w_2 \;\; \cdots \;\; w_m] \end{aligned} \tag{3.12}$$

where $w_i, i = 1, 2, \cdots, m$ is the set of generalized eigenvectors of S_B and S_W corresponding to the m largest generalized eigenvalues $\lambda_i, i = 1, 2, \cdots, m$, such as

$$S_B w_i = \lambda S_W w_i, \;\; i = 1, 2, \cdots, m. \tag{3.13}$$

After projected by the optimal projection matrix W_{opt}, the samples could achieve the maximal ratio of the between-class scatter to the within-class scatter.

3.3.2 Derivation of Discriminant Colour Space

Clearly, when learning UCS for facial expression recognition, the expression label of sample images is not utilized. However, the discriminant colour space (Yang and Liu, 2008), which applies discriminant analysis, considers the label of sample images. Let c be the number of the facial expressions, Q_{ij} be the j-th colour image in class i, where $i = 1, 2, ..., c, j = 1, 2, ..., n_i$ (n_i is the number of training samples in class i). The colour space between-class scatter matrix L_b and colour space within-class scatter matrix L_w are defined as

$$L_b = \sum_{i=1}^{c} p_i(\bar{Q}_i - \bar{Q})^T(\bar{Q}_i - \bar{Q}) \tag{3.14}$$

$$L_w = \sum_{i=1}^{c} p_i \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Q_{ij} - \bar{Q}_i)^T(Q_{ij} - \bar{Q}_i) \tag{3.15}$$

where p_i is the priori probability for class i, $\bar{Q}_i$ is the mean image of the training images in class i, $\bar{Q}$ is the mean image of all the training images. The discriminant analysis criterion becomes

$$J(x) = \frac{x^T L_b x}{x^T L_w x} \tag{3.16}$$

where x is a unitary vector. In addition, L_b and L_w are nonnegative define matrices, the criterion in equation (3.16) is a generalized Rayleigh quotient. Its optimal solution $x_1 = [x_1^1\ x_1^2\ x_1^3]^T$ is actually the generalized eigenvector of eigen-decomposition problem $L_b x = \lambda L_w x$, corresponding to the largest eigenvalue. In practise, one discriminant colour component is not enough for the colour information representation, so all three of the eigenvectors are kept to form the discriminant colour space

$$[D^1\ D^2\ D^3] = [R\ G\ B][x_1\ x_2\ x_3] \tag{3.17}$$

We should remind that in deriving DCS for face recognition, the label is the identity of persons and for facial expression recognition, the label is the type of expressions. So the UCSs are the same both for face recognitor and FER but the DCSs are different.

3.4 Experiments

In this section, the learned colour spaces are tested on Oulu-CASIA NIR&VIS facial expression database and CurtinFaces database. As to the experiment setup, there are usually two ways to generate training set and testing set for a facial expression recognition system. One is person-dependent, while the other is person-independent. In the person-dependent case, the individuals included in the testing images also show up in the training images. It means that the classifier has seen the individuals included in the testing images. However, in the person-independent case, the individuals included in the testing images never appear in the training images. The training images and testing images are both mixtures of images from different individuals. The individuals in the testing images are totally *strangers* to the classifier. In this section, both the person-dependent and the person-independent cases are conducted on each of the databases. We also obtain the recognition results of the uncorrelated colour space (UCS) and discriminant colour space (DCS), compared against with the results of RGB colour space and gray scale images. In either of the cases, we utilize Fisher's linear discriminant(FLD) to extract facial expression features, and then feed them into the nearest-neighbour (NN) classifier to obtain the recognition results.

Size	AN	DI	FE	SA	HA	SU
Training	477	488	520	530	550	503
Testing	582	404	537	407	503	449

Table 3.1: The configuration of the person-independent case on Oulu-CASIA database.

Size	AN	DI	FE	SA	HA	SU
Training	547	466	546	486	546	495
Testing	512	426	511	451	507	457

Table 3.2: The configuration of the person-dependent case on Oulu-CASIA database.

3.4.1 Results on Oulu-CASIA NIR&VIS Database

The Oulu-CASIA NIR&VIS facial expression database (Zhao *et al.*, 2011), consists of six expressions from 80 individuals between 23 to 58 years old, and almost 73.8% of the subjects are males. The images are frame sequence of a video, and oringinally digitized into 320×240 pixel arrays. In the experiment, the first 9 images of each sequence are ignored for their low expression intensity. The selected 6059 images are aligned into 64×64 pixel arrays according to the coordinates of eyes and mouth. Both person-dependent and person-independent experiment is conducted on the aligned face images.

In the person-independent case, the images of the first 40 individuals are taken as the training samples, and the last 40 individuals' images are chosen as testing images. Thus, it is guaranteed that the training and testing images are from different subjects. The configuration for the training and testing size of every expression in person-independent case is listed in table 3.1. Figure 3.2(a) gives the recognition rates of the six prototypic expressions in person-independent case. It shows that all the colour spaces achieve the best performance in the recognition of happiness, and the worst performance in the recognition of fear. It is notable that in the person-independent case, the discriminant colour space (DCS) can improve the recognition rates of fear by more than 10% when compared against RGB colour space, whereas the recognition rates of anger and surprise are much worse,

	Gray	RGB	DCS	UCS
Independent	49.5	49.9	48.6	53.0
Dependent	91.3	91.4	91.7	92.5

Table 3.3: The average recognition rates on Oulu-CASIA database.

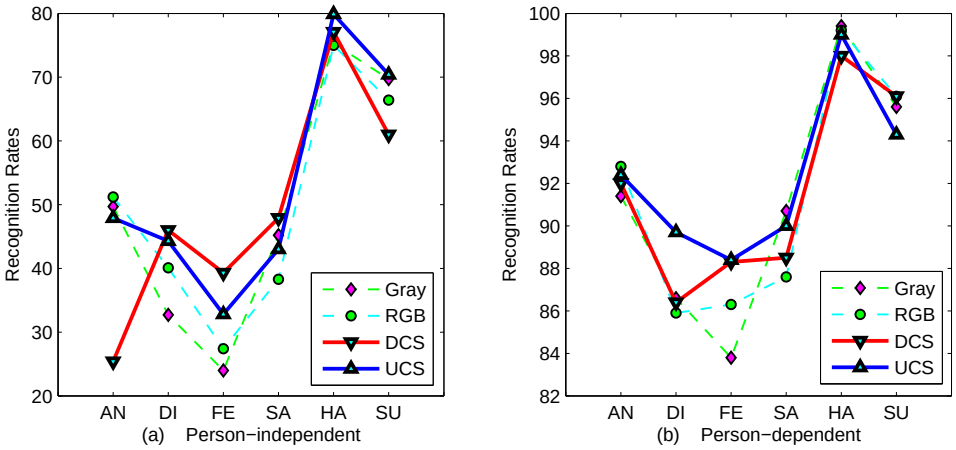


Figure 3.2: The recognition rates on Oulu-CASIA database.

even worse than gray images. The uncorrelated colour space (UCS) performs better than DCS on anger, happiness and surprise while slightly worse on disgust, fear and sadness.

In the person-dependent case, the images included in the first half of the each sequence are selected to form training set, while the latter half of the sequences serve as testing images. Table 3.2 records the configuration for the training and test size of each expression in person-dependent case, and the recognition rates of the six prototypic expressions are illustrated in figure 3.2(b). Compared with RGB colour space, the discriminant colour space (DCS) improve the recognition rates of fear and sadness, while the uncorrelated colour space (UCS) performs even better than DCS, especially in the recognition of digust.

In both cases, all the colour spaces achieve the best performance in the recognition of happiness, and the worst performance in the recognition of fear. Table 3.3 records the average recognition rates for the four colour spaces. One can see that the uncorrelated colour space (UCS) is the best colour representation for facial expression recognition, since it achieves higher average recognition rates than other colour spaces. As illustrated in figure 3.2, the discriminant colour space (DCS) fails to keep high recognition rates in anger and surprise when compared against RGB colour space, so it is not consistent enough to represent colour information in facial expression.

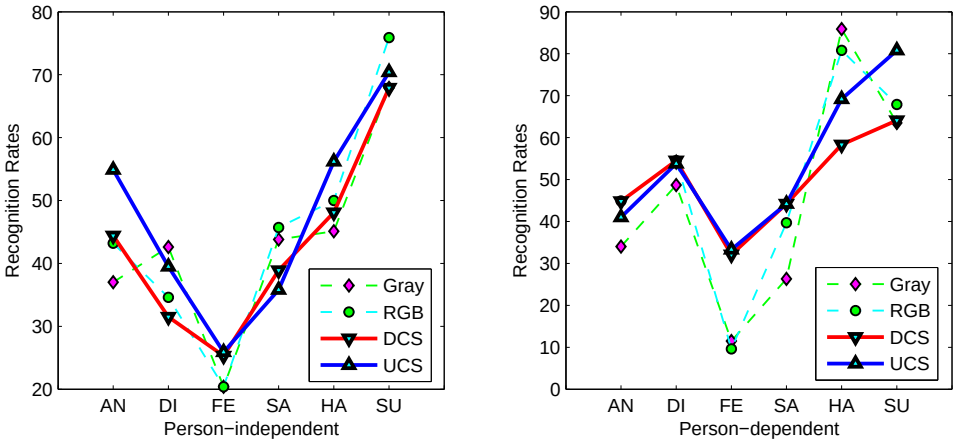


Figure 3.3: The recognition rates on CurtinFaces database.

3.4.2 Results on CurtinFaces Database

The CurtinFaces database contains over 5000 face images of 52 individuals. It was collected with a Kinect sensor and a standard Panasonic digital camera, with each Kinect capture accompanied by an image taken from the Panasonic camera at the same time. All the images are originally kept in RGB space. These images have varying facial expression, pose, illumination and occlusion, simulating a real-world uncontrolled face environment. In our experiment, we select a subset which consists of 1872 images of 52 subjects with 6 prototypic expressions and align them into 64×64 pixel arrays according to the coordinates of eyes and mouth. Every subject has 6 images in each of the 6 prototypic expression. All the colour spaces are tested on the aligned images under three different setup, namely person-independent, person-dependent and crossing image sources.

3.4.2.1 Person-independent vs person-dependent

In the person-independent case, the images of first 25 individuals are chosen to form the training set, and the images of the remainder 27 subjects go into the testing set. Thus, the training size of each expression is 150 (25×6), and the testing size is 162 (27×6). However, in the person-dependent case, we select the first 3 images of each expression (for all the 52 subject) to form training set, and the reminder 3 images of each expression are taken as testing set. That is to say, both the training and testing size are 156 (52×3).

Figure 3.3 shows the recognition rates of the experiments conducted in both person-

	Gray	RGB	DCS	UCS
Independent	42.8	45.0	42.7	47.1
Dependent	45.0	49.6	49.7	53.7

Table 3.4: The average recognition rates on CurtinFaces database.

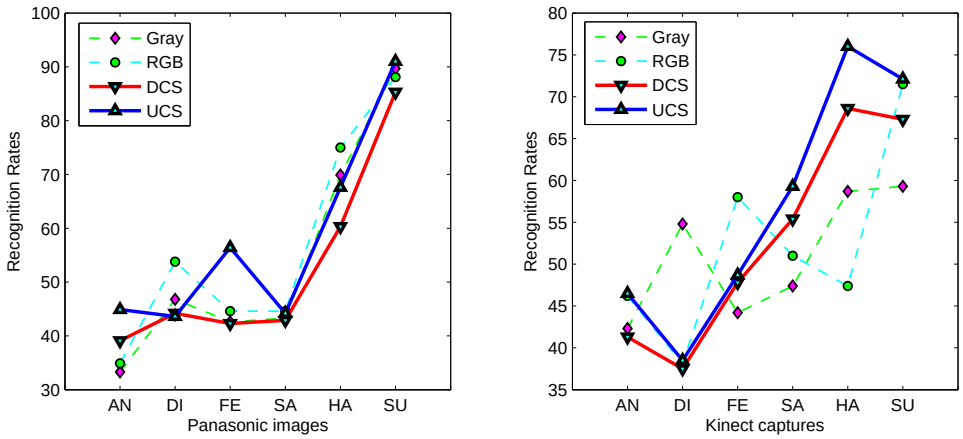


Figure 3.4: The recognition rates of crossing image sources on CurtinFaces database.

independent and person-dependent ways. Among the 6 prototypic expressions, happiness and surprise are relatively easier to recognize in both cases. Compared with RGB colour space, the uncorrelated colour space is generally more effective in colour information representation for facial expression recognition. However, in the person-dependent case, the discriminant colour space works better than RGB colour space on fear, but worse on happiness and surprise. The uncorrelated colour space achieves the highest average recognition rates, while discriminant colour space fails to show a consistent performance, as recorded in table 3.4.

3.4.2.2 Crossing image sources

In CurtinFaces database, every Kinect capture is accompanied by an image taken by the standard Panasonic camera, and both of them are colour image and represented in RGB colour space. In the last test, the facial expression recognition is conducted by crossing image sources: training on the Kinect captures and testing on the images from Panasonic camera, and vice versa. Actually, this is a specially case of person-dependent facial expression recognition, since all the subjects involved in the recognition comes out

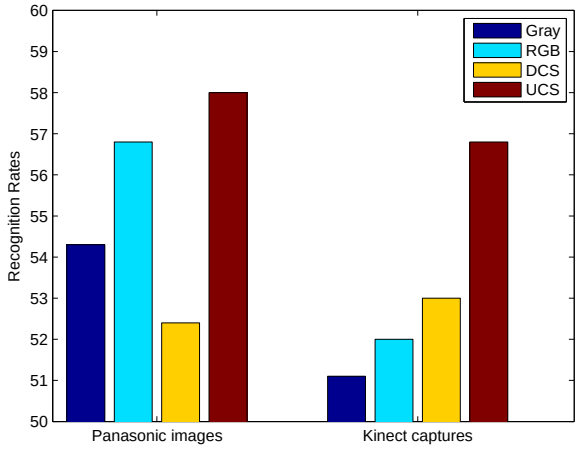


Figure 3.5: The average recognition rates of crossing image sources on CurtinFaces database.

in the training set. To the classifier, there is no *stranger* in the testing set. The only difference lies in the images source of the training and testing set: the one is standard Panasonic digital camera, the other is Kinect sensor. The images from Panasonic camera are in high-resolution, but the Kinect captures are in low-resolution.

Figure 3.4 records the recognition rates of the 6 prototypic expressions. It shows that the uncorrelated colour space is universally better than the discriminant colour space, whether trained on Panasonic images or Kinect captures. The average recognition rates, which is illustrated in figure 3.5, tell us that the discriminant colour space is slightly better than RGB colour space when the recognition system is trained on Kinect captures, but much worse when trained based on the Panasonic images. However, no matter based on which images source, the performance of uncorrelated colour space is always better than RGB and DCS. Therefore, the uncorrelated colour space is the best space to represent colour information for facial expression recognition, but discriminant colour space is not good enough to yield a consistent performance. The image source will vary a lot in a real facial expression system. The system should be able to work robust over any image source, and so it is with the colour space adopted by the system. Compared with DCS, the uncorrelated colour space (UCS) is better for colour representation, since it performs better consistently over different image sources.

3.5 Chapter Summary

Colour provides useful information in the image recognition problem. Normally, colour information is represented in RGB colour space, but there are neither theoretical nor experimental justifications for supporting it is a good representation for a specific recognition purpose. This chapter explores the colour information representation in facial expression problem, via learning the discriminant colour space and uncorrelated colour space to encode colour information.

Firstly, the experiment results reveal that the uncorrelated colour space represents colour information best for facial expression recognition since it achieves the highest recognition rates. However, the discriminant colour space fails to achieve a consistently better performance than RGB space, occasionally even worse than gray space, which is significantly different from face recognition. This reveals that DCS is not suitable for facial expression recognition. Secondly, the facial expressions contained in CurtinFaces database is much harder to be recognized than the Oulu-CASIA NIR&VIS database. The average recognition rate is above 90% on Oulu-CASIA NIR&VIS database in the person-dependent case, but only about 50% on the CurtinFaces database. The pose, illumination and occlusion varieties included in CurtinFaces database are quite challenging in facial expression recognition. Finally, the fact that DCS does not work in facial expression recognition reflects the difficulty of coulour representation in facial expression. This is may caused by the ambiguity of some easily-confused facial expressions. Therefore, human beings' identities and expression types can not be treated similarly in pattern recognition.

In addition, the training based on Panasonic images achieved better performance than based on Kinect captures, since Panasonic images are in high-resolution. This is important issue for real life facial expression recognition. For example, there are expression samples of high-resolution available in training stage, but the test samples are captured by a surveillance camera from long distance, the face resolution could be very low. Our experiment on this issue reveal that even in person-dependent setup, recognizing expression by cross image source with different resolution is still very different.

Chapter 4

Easily-confused Expression Recognition via Hierarchical Classification

The possibility of enabling computers to recognize and analyze the information conveyed by facial expression has attracted significant research interest over the last few years. It has given rise to a number of methods for automatic facial expression recognition (Pantic and Rothkrantz, 2000; Fasel and Luettin, 2003). Though much progress has been made, robust and effective facial expression recognition remains difficult due to the subtlety and variability of facial expressions. These studies conducted either person-independent or person-dependent facial expression recognition in the experiments. Obviously, a feasible facial expression recognition system should be able to work person-independently. The comparisons of the two cases by Cohen *et al.* (2003) reveal that the person-independent case is much more difficult than the person-dependent case.

(Ekman and Friesen, 1971) made a cross-cultural study on the existence of universal categories of emotional expressions, which is referred to as the prototypic expressions consisting of happiness, sadness, surprise, fear, anger and disgust. It has been shown that the six prototypic expressions are not mutually distinguishable. The developed systems (Cohen *et al.*, 2003; Kotsia and Pitas, 2007; Kotsia *et al.*, 2008; Michel and El Kaliouby, 2003) show that there are often confusions between anger and disgust while some other work (Aleksic and Katsaggelos, 2006; Sebe *et al.*, 2007) show sadness is often confused with anger. In order to achieve highly accurate recognitions for all the expressions, many researchers attempted to eliminate such confusions. The attempts generally involve two vital steps: feature extraction and expression classification.

There are two common approaches to extract facial features from the original face images: geometric feature-based approach (Gu and Ji, 2004; Valstar *et al.*, 2005) and appearance-based approach (Bartlett *et al.*, 2003; Shan *et al.*, 2009). In Tian *et al.* (2002) and Zhang *et al.* (1998), the authors argued that appearance features are better than geometric

features, because geometric features are more sensitive to inaccurate image alignment. In addition, Lucey *et al.* (2010) showed that appearance information is more important to the recognition of anger, sadness and fear. However, with the recent development on face alignment and facial feature tracking, an increasing number of expression analysis algorithms are based on geometric features. Valstar *et al.* (2005) presented a method that can detect facial action units effectively by classifying features derived from the tracked facial landmarks. They argued that the geometric features is well suited for facial expression analysis, especially with facial feature tracking. The studies from both sides indicate a possible combination of these two kinds of features as a better face representation for facial expression recognition.

Normally, a classifier is designed based on the extracted features for the classification of the six prototypic expressions, in which all the expressions are treated equally and evenly. Out of the six expressions, happiness and surprise are the easiest to recognize (Michel and El Kaliouby, 2003; Pardàs and Bonafonte, 2002). The remaining four expressions are more subtle ones that are often confused with each other. This fact prompts us to divide and conquer the problem of recognizing the 6-classes expressions by a hierarchical classification. The expressions that are commonly confused are merged into one class, which is joined by the remaining prototypic expression classes to form the first tier of classification. It is expected that the first tier classification will perform well since the expressions that easily confused have been merged together. In the second tier, the prototypic expressions in the merged class are separated by an earmarked classifier. The hierarchical classification provides an opportunity to utilize different features to obtain the best performance in each tier.

In this chapter, a hierarchical SVM classifier is designed to improve the performance of person-independent facial expression recognition. This hierarchical classifier enables us to divide and conquer the recognition problem in two tiers. The easily-confused expressions are merged into one class in the first tier and then separated in the second tier. We also propose to utilize different kinds of features in each tier, because classification of different expressions is targeted in each tier. The combined features of LBP and displacement are used for facial expression description in the first tier since it yields the best performance among possible features. In the second tier, the landmarks on mouth and eyebrows are selected to represent expressions since mouth and eyebrows are proven highly related to those easily-confused expressions (Bourel *et al.*, 2001; Kotsia *et al.*, 2008). The experimental results obtained by applying the hierarchical classifier on the CK+ dataset (Lucey *et al.*, 2010) demonstrate the satisfactory performance of the proposed method.

4.1 Facial Expression Representation

Effective facial feature extraction from face images plays an important role in facial expression recognition. The appearance features and the geometric features are the common features utilized in facial expression representation. Shan *et al.* (2009) compared the LBP features with the Gabor features for facial expression recognition using different classifiers, and studied their performances over various resolutions. The comparison results revealed that the LBP features are effective and efficient for facial expression recognition, even for low resolution face images. For the geometric features, the displacement of facial landmarks reflects the facial component motion, which widely used for expression representation.

4.1.1 Local Binary Pattern (LBP)

The appearance features model the appearance changes of faces, mainly caused by different facial expressions. The original LBP operator was proposed by Ojala *et al.* (1996) as a powerful means of texture description. As illustrated in Figure 4.1, LBP operator labels the image pixels by thresholding a 3×3 neighbourhood of each pixel with the center value and considering the results as one binary number. The histogram of the LBP labels accumulated over a local region is then used as a texture descriptor. The binary numbers, which is called Local Binary Patterns, encode the local texture primitives including corners, edges, spots etc, so the histogram of LBP could be used as a texture representation.

Originally, the LBP operator consider the 3×3 neighbourhood to label the central pixel, which can not capture dominant features in large scale. Thus, Ojala *et al.* (2002) extended the operator by using circular neighbourhoods and bilinear interpolation of the pixel values. This allows the extended LBP operator $LBP_{P,R}$ to use any radius and number of pixels, where (P, R) denotes the neighbourhood of P uniformly spaced sampling points on the circle of radius R. Actually, $LBP_{P,R}$ produces 2^P different output binary patterns, with certain patterns containing more information than others (Ojala *et al.*, 2002). Thus, it is more effective to use a subset of the 2^P local binary patterns, i.e. uniform patterns, to describe the texture information. A local binary pattern is called uniform if it contains no more than two bitwise transitions from 1 to 0 or vice versa when the binary pattern string is considered as circular. It has been noted by (Ojala *et al.*, 2002) that uniform patterns hold nearly 90% of all patterns in the $(8, 1)$ neighbourhood and about 70% in the $(16, 2)$ neighbourhood. Hence, in the computation of the extended LBP operator $LBP^{u2}_{P,R}$, the

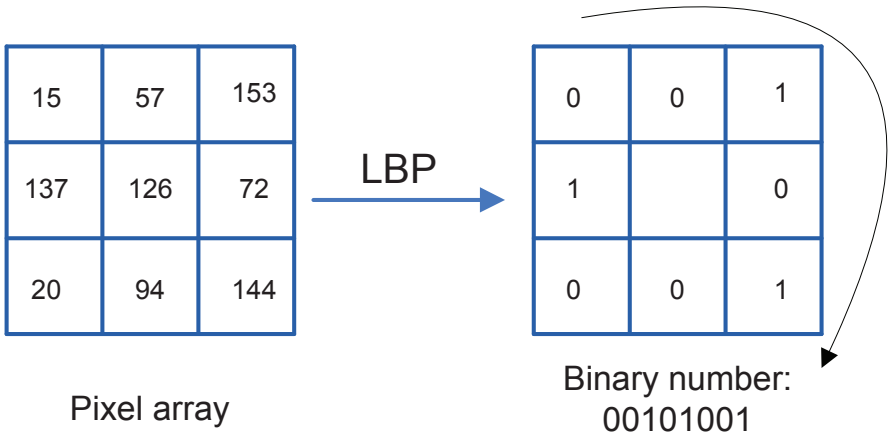


Figure 4.1: The demo of extract binary number from pixel array.

histogram has a separate bin for every uniform pattern, and all non-uniform patterns are assigned to a single bin. The number of resulted histogram bins are much less than 2^P. For example, $LBP_{8,1}$ has 256 bins, but $LBP^{u2}_{8,1}$ only has 59 bins.

The $LBP^{u2}_{P,R}$ operator is adopted by this chapter for facial appearance feature representation. It can encode the appearance information about the distribution of the local micro-patterns caused by facial expressions. Therefore, it can serve intuitively as the appearance feature representation in facial expression analysis.

4.1.2 Displacement of Facial Landmarks

Geometric features have been widely exploited in facial representation (Tian *et al.*, 2001; Pantic and Rothkrantz, 2004), where locations and displacements of facial components are extracted to represent the face geometry. Recently, Lucey *et al.* (2010) manually label some key frames in video sequences and use a descent AAM fitting algorithm (Matthews and Baker, 2004) to get the landmarks of the remaining frames. Their work shows that facial expression recognition benefits a lot from fusion of both shape and appearance features.

Assuming that the facial components have been labelled with N landmarks, the coordinates of the landmarks could be denoted as $p_i = (x_i, y_i), i = 1, ..., N$. The face images could then be represented by the location information vector $P = [p_1\, p_2 \ldots p_i \ldots p_N]$, which is the concatenation of all the landmarks p_i. The location information vectors encapsulate the shape and position of the facial components which are affected by the expressions.

Furthermore, in facial expression recognition based on image sequences, the facial movements can be measured by the geometrical displacement of corresponding facial feature points between the current frame and the initial frame. The displacement vector D can hence be derived from concatenating the displacements of all the landmarks:

$$D = [\Delta x_1\ \Delta y_1\ \Delta x_2\ \Delta y_2\ ...\ \Delta x_i\ \Delta y_i\ \ ...\ \Delta x_N\ \Delta y_N] \tag{4.1}$$

where $\Delta x_j, \Delta y_j$ is the x, y coordinate displacement of the j-th landmarks respectively. The displacement information encodes the motion of the landmarks from a neutral face to faces with expressions of different intensity. Both the displacement and location information are utilized in our method because they are directly related to facial expressions.

4.2 The Proposed Method

As mentioned before, the six prototypic expressions are not mutually distinguishable, so the confusions caused by the easily-confused expressions will affect the recognition performance significantly. In order to improve the recognition performance, the proposed method attempts to eliminate such confusions via a hierarchical classification. It has two advantages. Firstly, the hierarchical classification can pick the distinguishable expressions out in the first tier, and then focuses on the classification of easily-confused ones in the second tier. Secondly, the hierarchical structure enables us to utilize the most appropriate features for expression recognitions in each tier. This section provides the details of the proposed hierarchical classification method.

4.2.1 Feature Extraction

In the proposed method, both the appearance-based feature (LBP features) and the geometric feature (Displacement and Location features) are extracted from the face images.

LBP features: Each face image is first aligned and then divided into 42(6×7) blocks, and the 59-bin $LBP^{u2}_{8,1}$ operator is used to extract texture features form each block, which is a trade-off between the recognition performance and computational complexity. The 59-bin LBP histogram derived from each of the 42 blocks are concatenated to a 2478(59×42) dimension vector to represent a face image.

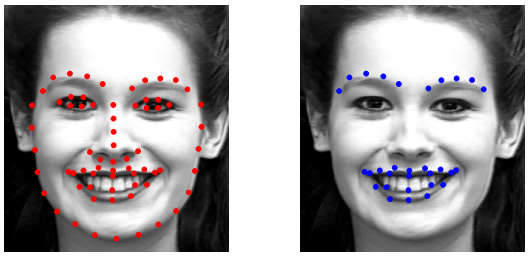


Figure 4.2: (Left)Landmarks used in displacement feature extraction. (Right)The selected landmarks of Mouth and Eyebrow.

Displacement features: In the proposed method, 68 landmarks which are tracked by Lucey *et al.* (2010) are utilized to extract the displacement information, as illustrated in Fig. 4.2 (Left). For each face image, the x, y coordinates displacement of the landmarks are obtained by subtracting the landmark locations of the neutral image from the corresponding landmarks locations in current image. The displacement feature is represented by a vector with the length of 136 which is formed by concatenating the x, y coordinate displacements.

MEb(Mouth and Eyebrow) features: The mouth and eyebrows are the most important parts for facial expression recognition. It has been shown by Pardàs and Bonafonte (2002) that the mouth and eyebrows possess the maximum amount of information related to facial expressions, with the mouth carrying more information than the eyebrows. In Kotsia *et al.* (2008), the authors show that occlusion of the mouth leads to inaccuracies in the recognition of anger, fear, happiness and sadness, whereas the occlusion of the eyes and brows leads to a dip in the recognition accuracy of disgust and surprise. Another occlusion research by Bourel *et al.* (2001) has demonstrated that sadness is mainly conveyed by the mouth. Thus, we select the landmarks on the mouth and eyebrows (see Fig.4.2 (Right)) and utilize both the location and displacement of these landmarks to form the MEb (Mouth and Eyebrow) features for distinguishing anger and sadness confusion which are most commonly confused.

4.2.2 Hierarchical Classifier Design

Since the six prototypic expressions are not evenly distinguishable, we attempt to divide and conquer the recognition problem by a hierarchical classification. The hierarchical classification has a structure of two tiers. In the first tier, the easily-confused prototypic expressions are considered as one class and join the remaining expressions for classification.

In the second tier, another classifier, which focuses only on the expressions in the merged class, is trained to separate the images of the merged class into the prototypic expressions. The design of the 2-tiered structure allows us to use the appropriate features in each tier.

Support vector machines (SVMs) have been proven powerful in facial expression classification (Valstar *et al.*, 2005; Bartlett *et al.*, 2003). It also achieves the best performance according to a comprehensive study (Shan *et al.*, 2009), so we adopt SVMs as the classifiers for facial expression recognition in this paper. SVMs attempt to find the hyperplane that maximizes the margin between the positive and negative observations for a specified class. Given a training set of labelled examples $\{(x_i, y_i), i = 1, ..., k\}$ where $y_i \in \{-1, 1\}$, a testing example x is labelled by the following function:

$$f(x) = sgn(\sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b) \tag{4.2}$$

where α_i are Lagrange multipliers of a dual optimization problem that determine the classification hyperplane, $K(\cdot, \cdot)$ is a kernel function, and b is the threshold parameter of the hyperplane.

SVMs make binary decisions. However, there are six classes in facial expression recognition, each representing one of the prototypic expressions (anger, disgust, fear, sadness, happiness and surprise). In this paper, we use LIBSVM (Chang and Lin, 2011a) for the training and testing of SVMs, which achieves the multi-classes classification according to the one-against-rest technique. With regard to the parameter selection, we carry out coarse-to-fine grid search in a 5-fold cross-validation. The parameter which yields the best cross-validation accuracy is selected for the decision function.

As illustrated in Fig.4.3, a hierarchical SVM classifier with two tiers is designed for the six prototypic expressions recognition. Firstly, we merged two of the six prototypic expressions (anger and sadness), which are the most commonly confused expressions, into one class. Together with the remaining four prototypic expressions, there are 5 classes in the first-tier classification. A 5-classes SVM classifier is used for this tier. The performance of the first-tier classification should be much better than directly classifying the six expressions since the major confusion has been removed by merging anger and sadness images together. For the images categorized as the merged class (anger and sadness), a 2-classes SVM is trained in the second tier to separate them into anger and sadness.

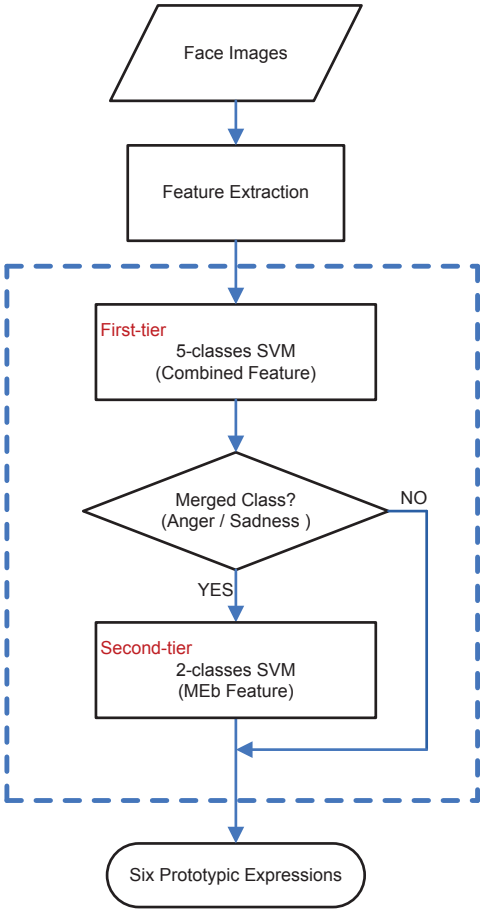


Figure 4.3: The flow chart of the hierarchical classification.

4.2.3 Feature Selection in Each Tier

The first-tier classification plays such an important role in the whole recognition procedure that it should perform as accurately as possible. We trained the 5-classes SVM classifier based on the LBP feature, the displacement feature and the combined LBP and displacement feature separately. The results show that the combined feature yields the best performance in the classification of the 5 classes (*4 prototypic expressions and one merged class*). Thus, the combined feature is selected to represent facial expressions in the first tier.

After the first tier, the merged class flows into the second-tier classification, in which the images are separated into anger and sadness. The features highly related to these two expressions are adopted because the second-tier SVM only focuses on the classification of anger and sadness. We choose the displacement feature as the expressions representation

at first, but approximately 20% sadness images are still misclassified as anger.

The inveterate confusion of anger and sadness evokes us to extract new feature since all the features used so far can not provide significant distinctions in representing these two expressions. Since the mouth and eyebrows possess the maximum amount of information related to the facial expressions (Pardàs and Bonafonte, 2002), especially that the sadness is mainly conveyed by the mouth (Bourel *et al.*, 2001), we attempt to extract features of mouth and eyebrows to discriminate anger and sadness. As illustrated in Fig.4.2 (Right), the landmarks of mouth and eyebrows are selected and both the location and displacement of these landmarks are used to form the MEb (Mouth and Eyebrows) feature for the classification of anger and sadness in the second tier. The recognition result reveals that the selected MEb feature could separate anger and sad better than all the other features.

4.3 Experiments

4.3.1 Experiment Settings

Experiments have been conducted by applying the proposed method on the extended Cohn-Kanade (CK+) dataset (Lucey *et al.*, 2010), which is comprised of 593 image sequences of 210 individuals. The expression in each sequence began with a neutral face and ended at the peak intensity. For all the 593 sequences, each image was AAM tracked with 68-points landmarks. However, only 327 of the 593 sequences carry the prototypic expression labels. The original images in the CK+ dataset are digitized into either 640×490 or 640×480 pixel arrays with 8-bit grey scale or 24-bit colour values.

In our experiments, only the images from the labelled 327 sequences are chosen to test the proposed method. The first 5 images of each sequence are ignored due to their low expression intensity. The selected images are aligned and resized into 110×150 pixel grey scale arrays automatically according to the location of eyes and mouth and then split into training images and testing images.

4.3.2 Person-dependent vs Person-independent

There are two ways to generate training set and testing set for a facial expression recognition system. One is person-dependent, while the other is person-independent. In the

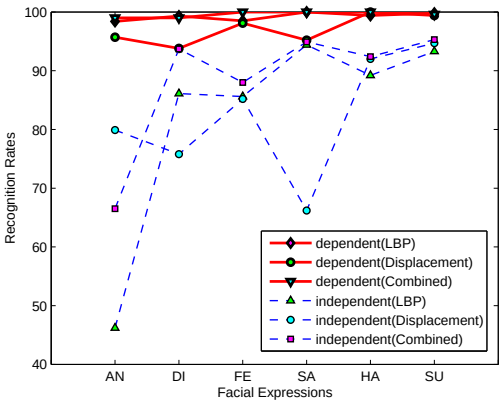


Figure 4.4: The comparison of person-dependent and person-independent facial expression recognition.

person-dependent case, the individuals included in the testing images also show up in the training images. It means that the classifier has seen the individuals included in the testing images. However, in the person-independent case, the individuals included in the testing images never appear in the training images. In this section, both the person-dependent and the person-independent cases are conducted on the selected images from CK+ dataset. The LBP features, displacement features and combined features are extracted as the expression representation.

Fig. 4.4 illustrates the performances of the person-dependent and person-independent expression recognition based on the LBP feature, displacement feature and the combined feature. The comparison shows that the confusions are very little in the person-dependent case, especially based on the combined feature. However, in the person-independent case, the expression recognition problem becomes much more difficult. The classification performances of the six prototypic expressions decrease significantly, especially in the recognition of anger and sadness. The confusion matrices of person-independent recognition based on the LBP feature and displacement feature are recorded in Table 4.1 and Table 4.2 respectively. It can be seen that 47.5% of the anger images are misclassified as sadness in the LBP feature based recognition while 31.3% of the sadness images are confused as anger in the displacement feature based classification. Even based on the combined feature, the confusion of anger and sadness is still as high as 25.0% (as shown in Table 4.3).

	AN	DI	FE	SA	HA	SU
AN	**46.2**	2.5	0.4	**47.5**	1.5	1.9
DI	1.6	**86.1**	2.4	9.1	0.0	0.8
FE	0.0	3.7	**85.6**	10.2	0.0	0.5
SA	1.5	1.5	1.5	**94.4**	0.0	1.0
HA	0.8	1.3	6.1	1.3	**89.2**	1.3
SU	0.2	1.2	1.6	3.5	0.2	**93.3**

Table 4.1: Confusion matrix of person-independent recognition based on the LBP feature.

	AN	DI	FE	SA	HA	SU
AN	**79.9**	6.8	0.0	11.7	0.0	1.7
DI	9.2	**75.8**	0.0	15.1	0.0	0.0
FE	1.4	0.0	**85.2**	13.0	0.0	0.5
SA	**31.3**	0.0	2.0	**66.2**	0.0	0.5
HA	0.0	1.1	0.6	6.3	**92.0**	0.0
SU	0.0	0.9	0.9	2.6	0.9	**94.7**

Table 4.2: Confusion matrix of person-independent recognition based on the displacement feature.

	AN	DI	FE	SA	HA	SU
AN	**66.5**	7.8	0.0	**25.0**	0.0	0.6
DI	2.0	**93.7**	0.0	4.0	0.0	0.4
FE	3.2	2.8	**88.0**	6.0	0.0	0.0
SA	0.5	2.5	0.5	**94.9**	0.0	1.5
HA	0.6	1.9	3.2	0.6	**92.4**	1.3
SU	0.0	0.9	1.6	2.1	0.0	**95.3**

Table 4.3: Confusion matrix of person-independent recognition based on the combined feature.

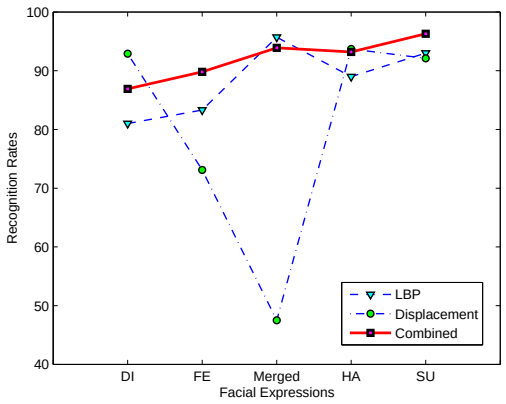


Figure 4.5: The first-tier feature selection.

4.3.3 Results of the Proposed Method

The proposed hierarchical classification only focuses on the difficult person-independent expression recognition. In the proposed method, a 5-classes SVM classifier is trained in the first tier classification since anger and sadness have been merged into one class. The recognition performances based on the LBP feature, displacement feature and combined feature are illustrated in Fig. 4.5. Obviously, the combined feature yields the best classification performance. Thus, it is selected as the expression representation in the first tier.

The first-tier classification categorizes the images into 4 prototypic expressions plus the merged class. In order to finish the prototypic expression recognition, the images in the merged class need to be separated into anger and sadness in the second tier. We first utilize the displacement feature in the second-tier classification. The confusion matrix of the hierarchical classification is recorded in Table 4.4. Compared to the result in Table 4.3, the hierarchical classification improves the recognition rate of anger from 66.5% to 86.0%. However, the recognition rate of sadness goes down to 75.8%, and the major confusion still lies between anger and sadness, with 21.7% of the sadness images misclassified as anger.

In order to better distinguish anger and sadness, we extract the MEb feature to be used in the second tier. Table 4.5 records the recognition result. It can be seen that the confusion between anger and sadness decreases to 14.6% while the sadness recognition rate reaches 93.4%. Although the rate for anger drops to 77.8%, the overall recognition rate reaches 89.6%, which shows that the selected MEb feature could separate anger and sad better than all the other features. Finally, we compare the proposed method with

	AN	DI	FE	SA	HA	SU
AN	**86.0**	7.6	0.0	6.4	0.0	0.0
DI	6.7	**86.9**	0.0	6.0	0.0	0.4
FE	1.4	2.8	**89.8**	6.0	0.0	0.0
SA	21.7	0.0	0.5	**75.8**	0.0	2.0
HA	0.2	0.8	2.1	3.0	**93.2**	0.6
SU	0.5	0.0	0.7	2.6	0.0	**96.3**

Table 4.4: The confusion matrix of the hierarchical classification (2-tier based on the displacement feature).

	AN	DI	FE	SA	HA	SU
AN	**77.8**	7.6	0.0	14.6	0.0	0.0
DI	1.6	**86.9**	0.0	11.1	0.0	0.4
FE	2.8	2.8	**89.8**	4.6	0.0	0.0
SA	4.0	0.0	0.5	**93.4**	0.0	2.0
HA	0.2	0.8	2.1	3.0	**93.2**	0.6
SU	0.7	0.0	0.7	2.3	0.0	**96.3**

Table 4.5: The confusion matrix of the hierarchical classification (2-tier based on the MEb feature).

several state-of-art methods in table 5.4. It shows that the proposed method achieves the best performance in the recognition of anger, fear and sadness.

Methods	AN	DI	FE	SA	HA	SU	AvgRate
Lucey *et al.* (2010)	75.0	94.7	65.2	68.0	**100.0**	96.0	83.2
Rudovic *et al.* (2012)	71.3	90.8	79.0	90.5	92.6	96.6	86.8
Zhong *et al.* (2012)	71.4	95.3	81.1	88.0	95.4	**98.3**	88.3
The Proposed	**77.8**	86.9	**89.8**	**93.4**	93.2	96.3	**89.6**

Table 4.6: The comparison with the state-of-art methods.

4.4 Chapter Summary

In this chapter, a hierarchical classification approach is proposed for person-independent facial expression recognition. Due to the difficulty in distinguishing anger and sadness, they are combined into one class and join the other four prototypic expressions in the first tier of classifications and then separated in the second tier. The hierarchical structure of the proposed method provides us with the opportunity to fuse different kinds of feature into the classification, which can enhance the recognition performance.

The experiment results on CK+ dataset show that the hierarchical SVM classifier improves the recognition performance for facial expression significantly, especially in reducing confusions between anger and sadness. We only test the proposed method on CK+ dataset because it is the only available dataset with facial landmark information. It is interesting that the selected mouth and eyebrow feature separates anger and sadness better than the displacement feature. This suggests that discriminative information of the prototypic expressions is conveyed by different facial components.

Chapter 5

Fully Automatic 3D Facial Expression Recognition

Prior research has mainly focused on 2D images or videos (Pantic and Rothkrantz, 2000) but recently 3D facial expression recognition (Sandbach *et al.*, 2012b) has gained popularity due to its invariance to pose and illumination. This chapter focuses on static 3D facial expressions. Feature extraction from face data plays a crucial role in recognition systems. Many methods analyze 3D faces by extracting local patches around manually labeled landmarks to achieve good expression recognition performance (Wang *et al.*, 2006; Maalej *et al.*, 2011). In Wang *et al.* (2006), local patches around landmarks are approximated with polynomial surfaces and principle curvatures are extracted by eigenvalue decomposition. They reported an average accuracy of 83.6%. More recently, Maalej *et al.* (2011) proposed to define a patch in terms of concentric geodesic rings that follow the curvature of the facial surface and calculated the Riemannian distance between corresponding rings of patches on the test face against a reference scan face. They report impressive results of 98.8% average accuracy. Surprisingly, this performance is even better than the human performance (Yin *et al.*, 2006a) on the same BU-3DFE database.

For all practical applications, facial expression recognition must be fully automatic. It is easier to manually label the expression, which is the required final outcome of the process, than to manually label multiple landmarks on a face. However, the above methods define patches around *manually* labeled landmarks that can be consistently located across faces and expressions. It is still an open problem to automatically recognize expressions without manual reference landmarks. Lemaire *et al.* (2013) attempt automatic facial expression recognition by extracting whole-face differential mean curvature maps (DMCM) features that can capture facial surface deformations caused by expressions without using facial landmarks. They reported an average recognition rate of 78.1%. Since they used the entire face, their features include face regions that are not relevant to particular expressions. In contrast, patch-based approaches work well as they can be located specifically on important landmarks such as the mouth, cheeks and eyes and different set of features can be defined for each patch.

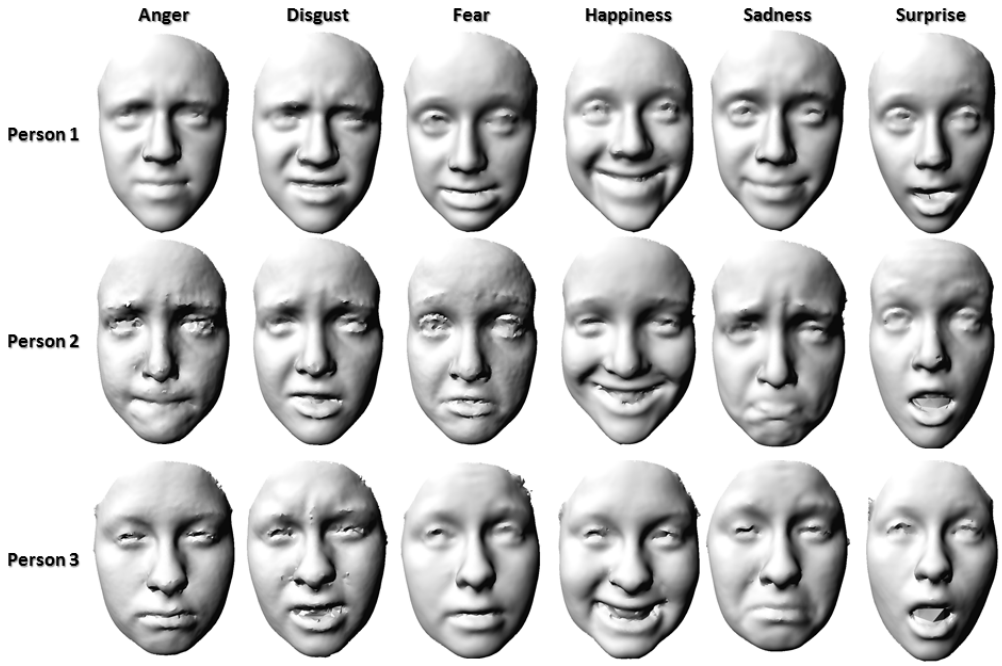


Figure 5.1: Facial expression examples from the BU-3DFE database.

We propose a fully automatic facial expression recognition algorithm based on depth features extracted from local patches. In order to define local patches without human intervention, we detect the nose tip and four eye corners automatically as five fiducial landmark points. From these, another 25 heuristic landmarks are generated and local depth features are extracted from patches around all the 30 landmarks. Then, mutually exclusive features which jointly have the largest characterizing power are selected from the extracted depth features using mRMR (maximum Relevance Minimum Redundancy) (Peng *et al.*, 2005). Feature selection is a critical step as depth features contribute differently to each type of expression. Moreover, it also takes care of errors in landmark locations. We seek to use a similar approach to that of Maalej *et al.* (2011) due to their very high recognition rates, however their concentric geodesic rings cannot be segmented to facilitate feature selection. Hence we instead utilize a discrete sampling of the depth patch as our features. Finally, the selected features are fed to a SVM classifier for expression classification.

5.1 Pre-processing

The raw 3D faces in BU-3DFE (Yin *et al.*, 2006a) are noisy and have minor pose variations as shown in Figure 5.1. As illustrated in Figure 5.2, we preprocess the faces before feature

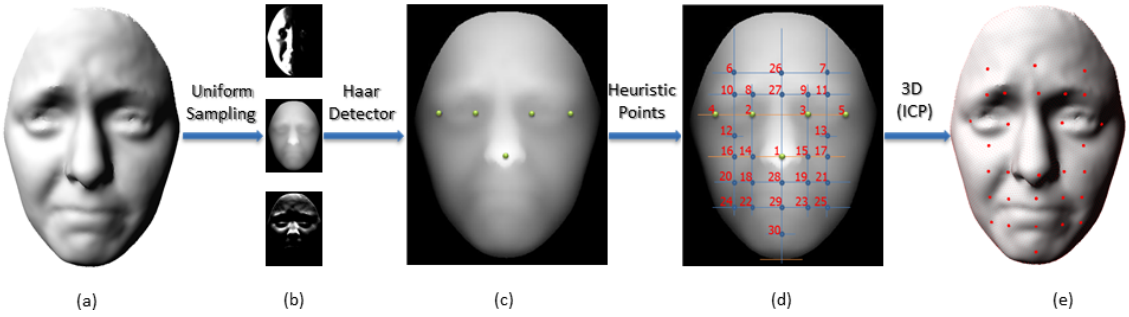


Figure 5.2: Pre-processing of a 3D face. (a) Original 3D face; (b) Range image and its x and y gradients rendered from 3D face; (c) Detected 5 fiducial points; (d) Generating heuristic points on range image; (e) Locating heuristic points on 3D face.

extraction. At first 5 fiducial points are automatically detected, followed by registration of facial point clouds. Heuristic points are generated for feature extraction. In fiducial point detection, the nose tip and four eye corners are located by a Haar detector and AdaBoost classifier (Viola and Jones, 2004) which enables the proposed method to be fully automatic. Then the 3D facial point clouds are aligned and registered according to a T-area located on each face using the five fiducial points. We use only the T-area since it is not very sensitive to noise. On each of the registered faces, 30 heuristic points are generated based on the 5 fiducial points to extract depth features.

5.1.1 Realtime Fiducial Points Detection

Automatic landmark detection on 3D faces is still an open problem due to the significant topology changes caused by expressions, such as opening mouth in surprise. We notice that features vary very little around some points when expression changes, such as the four eye corners and nose tip. Our realtime detection method detects these five points on a 3D facial surface. These five fiducial points and their relative distances are used to generate another 25 heuristic points on the face.

We train a Haar-cascade classifiers (Mian, 2011), which are based on the AdaBoost algorithm used for face detection (Viola and Jones, 2004). Given a 3D face as shown in Figure 5.2(a), the surface is uniformly sampled by a grid in the x, y-plane, and the depth information (z-direction) is encoded in a range image. The resulting range image and its x and y gradients (see Figure 5.3) are used to train the Haar cascade classifiers. For each point, the detector returns several candidate locations. The facial structure and relative location relationships between eyes and nose tip are utilized to remove the outliers and identify the

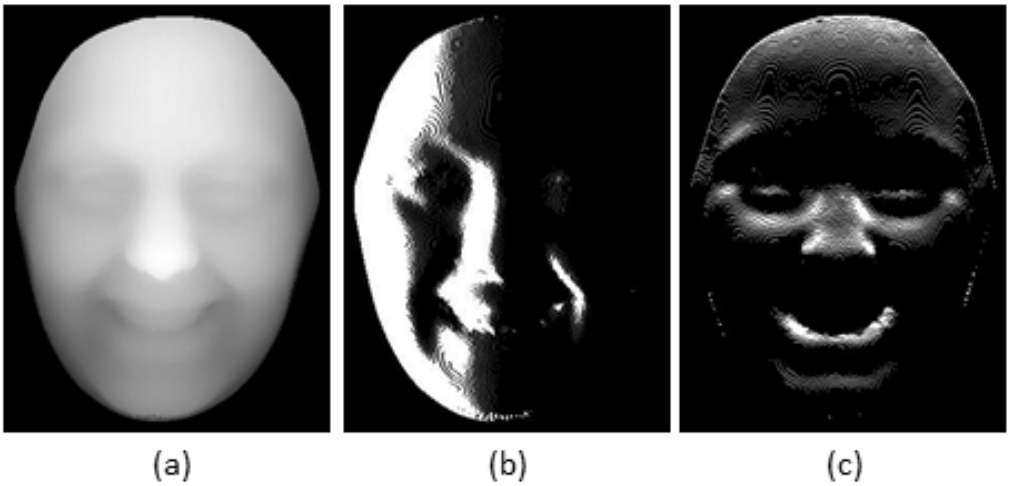


Figure 5.3: Example range image and its x and y gradients.

Table 5.1: Detection time of fiducial points.

Feature Point	Detection time
Nose	69.27 ms
Left eye	27.09 ms
Right eye	27.38 ms
Total	123.74 ms

correct eyes and nose clusters. Take Figure 5.4(a) as an example, the candidate locations of nose may fall on the lips and eyebrows, but nose is supposed to be in the central area of the face. Thus, only the central candidates accounts for the final decision. In 5.4(b), when detecting the right eye corners, some candidate locations are on the left side of the face, which are definitely the outliers. Note that the detector in Mian (2011) is in fact trained to find the eye outer corners and horizontal face scale. We extend this work to additionally localize the inner eye corners as well and thus detect five points rather than three as reported in Mian (2011). The process is illustrated in Figure 5.4.

We run our detection on all the 2500 faces in BU-3DFE database and the average detection time is recorded in Table 5.1. The total detection time for one face is less than 130 ms. Furthermore, since the BU-3DFE database provides manually labeled ground truth locations for the four eye corners, the distance from the detected location to the corresponding ground truth is calculated and illustrated in Figure 5.5, which shows that 90% of the detection errors are less than 4mm.

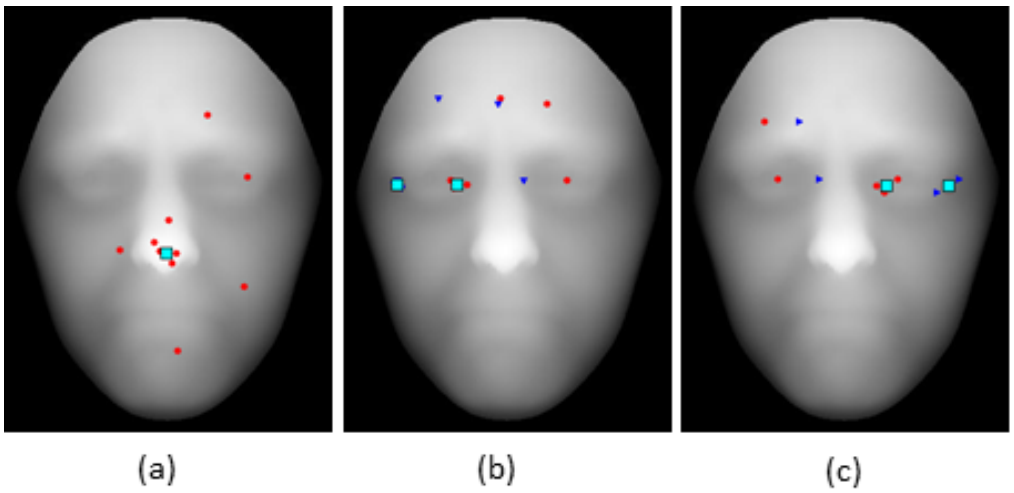


Figure 5.4: Demonstration of fiducial point detection. Small dots are candidates and large dots represent the final detections.

5.1.2 Registration

Minor pose variations exist in the BU-3DFE, which will affect the effectiveness of the feature extraction. Thus, it is necessary to register the faces against a 'standard' face. In the proposed method, all the faces are registered to the first female's neutral face in the BU-3DFE database by the Iterative Closest Points (ICP) (Besl and McKay, 1992) algorithm. The whole faces are not suitable for rigid registration due to the nonrigid facial surface deformations and topology changes caused by expressions. Therefore, we crop out a T-area from the face surface using a binary mask generated with the five fiducial points (see Figure 5.6). The T-area mainly covers the nose and forehead regions, which are relatively stable against facial expressions. Thus, T-ares is suitable for pose correction. Technically, the point clouds from T-area of two faces are fed to the ICP algorithm to calculate rotation matrix and translation vector, which are then used to register the corresponding two faces.

5.1.3 Heuristic Point Generation

The nose tip and eye corners are suitable to serve as fiducial points, but not representative enough to extract expression features. Thus, we generate another 25 heuristic points for expression feature extraction, as illustrated in Figure 5.7. The orange horizontal lines are the location of the eyes, nose and chin (the bottom point of the face). The eye-nose separation and nose-chin separation are denoted by h and d respectively. They are taken

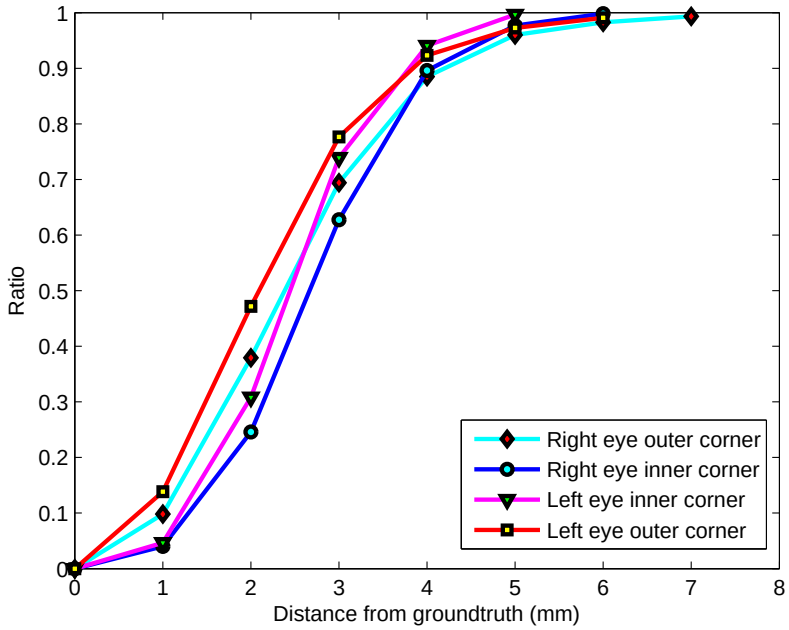


Figure 5.5: Detection error of the eye corners.

as the *'length unit'* to measure the face along the vertical direction, and render positions to draw horizontal baselines. Similarly, in the vertical direction, the location of the four eye corners and the eye centers are selected to draw vertical baselines which intersect the horizontal baselines. The heuristic points are then selected from the intersections of these baselines.

According to the research done by Kotsia *et al.* (2008), eyebrows and mouth area convey the most important information of facial expressions. Thus, the majority of heuristic points are selected around the eyebrows (points 8-11, 27) and mouth area (points 18-25, 28 and 29). This is a flexible scheme to generate heuristic points, in which the locations of some heuristic points can adjust according to different expressions. For example, d would be longer on a surprised face because of the opening mouth, so points 18-25, 28 and 29 will consequently be lowered to cover mouth area. Once the x, y-coordinates of the heuristic points are obtained from range images, the x, y, z-coordinates can be easily determined by finding the corresponding vertex on the uniformly sampled 3D point cloud.

Figure 5.6: T-area for registration.

这里能不能理解为用一个patch的信息来代替一个点的信息，从而能弥补points标记不是特别准确的不足？

5.2 Feature Extraction

Although our heuristic point generation based on the distance ratios of the fiducial points can adjust according to the changes in facial shape, the heuristic points are not as accurate as manually labelled ones. Thus we cannot assume perfect alignment of patches surrounding these heuristic points. To overcome this, we select a subset of features within a patch that are useful in expression recognition despite errors in landmark location. To facilitate this sub-patch feature selection, we choose to extract depth features sampled by a discrete grid on the patch. Our features can be essentially viewed as a discrete approximation of the rings used by Maalej *et al.* (2011) but also offer feature selection to choose arbitrary sub-patches – a process that could not be accomplished with the concentric geodesic ring features used by Maalej *et al.* (2011).

为什么需要choose任意一个sub-patch？

5.2.1 Patch-based Depth Feature Extraction

As shown in Figure 5.8, once a heuristic point has been located, we use a sphere with radius r centered at this point to crop a cluster of points. Then, a cubic patch is fitted to the cropped points using the code from D'Errico (2006). The fitted patch is then sampled on

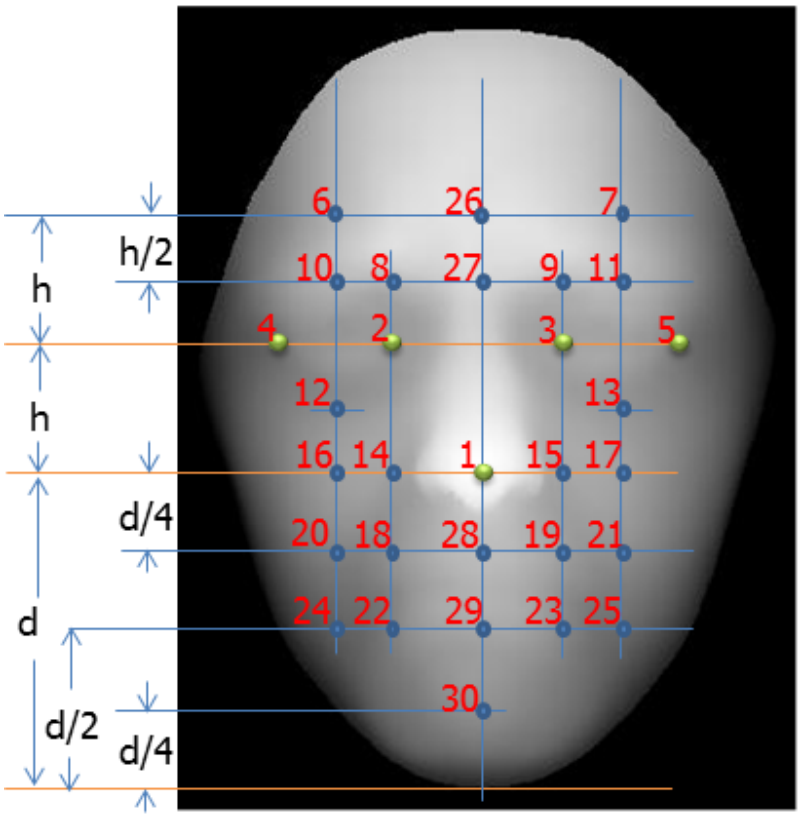


Figure 5.7: Schema of generating heuristic points.

a uniform 74×74 grid, but only the central 64×64 samples covering the points in r region are kept as the patch-based depth feature, in order to avoid the artifacts at boundaries. All sampled patches end up with equal resolution which is necessary for the classification. Figure 5.9 shows the same patch on the mouth corner of three different subjects under the six expressions.

降维究竟是为了什么？是为了减小数据维数，还是为了抽取特征？

The 64×64 depth feature matrix of each patch is then reshaped into a 4096-dimension row vector. Thus, each 3D face is represented by a 30×4096 matrix, as there are 30 patches. A dimension of 4096 is quite large for a feature vector that only describes a local patch, and there are overlaps between adjacent patches. Fortunately, it is possible to compress these vectors by projecting them into a linear subspace defined by 2DPCA. The goal is to discard the redundant information in preparation for feature selection. Assuming that there are N 3D face samples in the training set, the ith training sample is denoted by an m×n (our case 30×4096) matrix $A_i(i = 1, 2, ..., N)$, and the average of all training samples

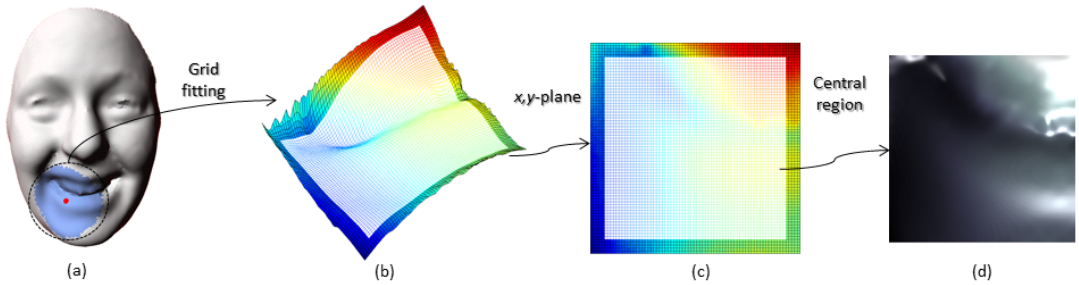


Figure 5.8: Patch-based depth feature extraction on 3D face surface.

is denoted by $\overline{A}$. Then, the scatter matrix C can be calculated by

$$C = \frac{1}{N}\sum_{i=1}^{N}(A_i - \overline{A})^T(A_i - \overline{A}). \tag{5.1}$$

According to Yang *et al.* (2004), the criterion of 2DPCA can be expressed by

$$J(X) = X^T C X, \tag{5.2}$$

where X is a unitary column vector. The optimal projection vector that maximizes the criterion is the eigenvectors of C corresponding to the largest eigenvalue. Normally, we select a set of the projection vectors, $X_1, ..., X_d$, subject to the orthonormal constraints. This can be achieved by applying the Singular Value Decomposition (SVD) on the scatter matrix C as

$$C = USV^T, \tag{5.3}$$

V代表什么？

where U is a 4096×4096 matrix of the eigenvectors and S is a diagonal matrix of eigenvalues, both sorted in descending order. The first d columns of U are the optimal projection vectors. To determine d, the ratio of the first d eigenvalues over the total eigenvalues is calculated by

$$\eta = \frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{4096}\lambda_i}, \tag{5.4}$$

where λ_i is the ith eigenvalue. In our experiment, the ratio η always reaches 0.99 swiftly at only $d = 50$. Thus, the first $d = 50$ eigenvectors are kept as the optimal projection matrix U_d, and used to compress the samples as

$$F = (A - \overline{A})U_d, \tag{5.5}$$

where F is a 30×50 matrix.

F有什么作用？它是不是特征矩阵？

To optimize the parameter r and grid size, we tested three different radiuses (25mm, 30mm and 35mm), and fitted into 20×20, 32×32 and 64×64 grids. The recognition rates of these

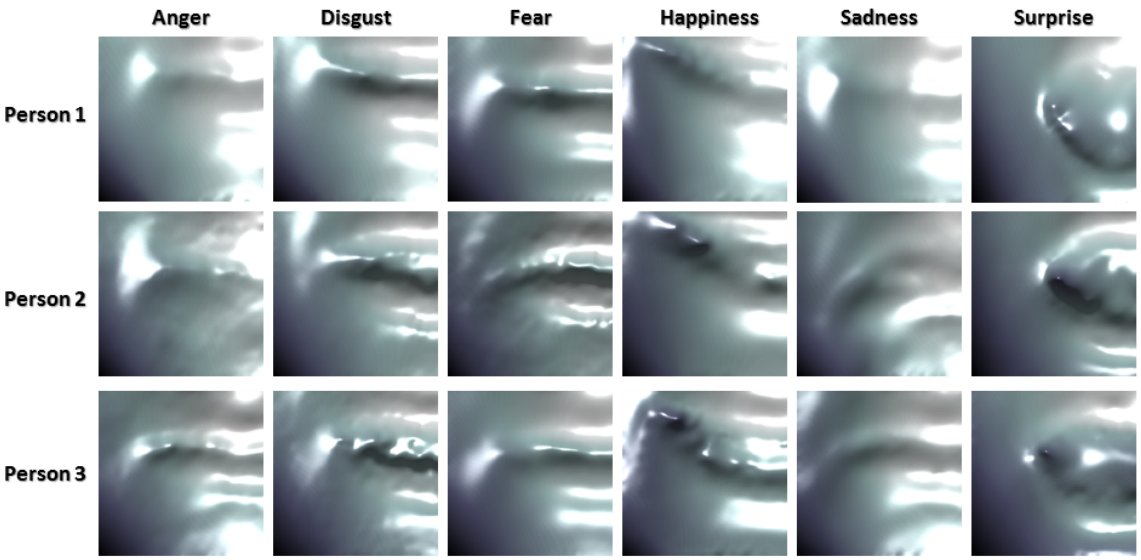


Figure 5.9: Comparison of 3D facial patch (mouth corner) under different expressions. The images are for the same three persons in Fig. 5.1

Table 5.2: Recognition rates of different parameters (patch radius r and fitting grid size) in feature selection.

Rates(%)	20×20	32×32	64×64
Radius=25mm	84.3	84.4	85.0
Radius=30mm	84.0	84.7	**85.4**
Radius=35mm	85.0	84.7	84.4

settings are given in Table 5.2. It can be seen that the patch-based features around the 30 heuristic points are not very sensitive to the radius of the cropping sphere and the fitting grid size. r=30mm and 64×64 grid are hence used for all our remaining experiments.

5.2.2 Feature Selection

Identifying the most characterizing features of the observed data is crucial to minimize the classification error. The idea of feature selection is that a simple combination of individually good features does not necessarily lead to good classification performance. That is to say, "the m best features are not the best m features" (Peng *et al.*, 2005). We adopt the framework of the *minimal-redundancy-maximal-relevance* (mRMR) (Peng *et al.*, 2005) to select the best features for recognition. This involves a two-stage selection algorithm.

是先确定m，还是先选择候选特征？应该是先选择候选特征再确定m的值

First, the mRMR criterion is used to select mutually exclusive features $S = \{x_1, ..., x_m\}$ that jointly have the largest characterizing power on each of the six prototypic expressions class c:

characterizing power是怎样计算的？

$$\begin{cases} max\Phi(D,R) = D(S,c) - R(S), \\ D(S,c) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c), \\ R(S) = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i,x_j) \end{cases} \tag{5.6}$$

where $I(x_i;c)$ is the mutual information value between individual feature and class, $I(x_i,x_j)$ is the mutual information value between two features.

When candidate features are selected, the next task is to determine the optimal number of features m. A wrapper that tests features with an SVM classifier is utilized to decide the size of the feature set, with the direct goal of minimizing the recognition error of the specific classifier on the training set.

为什么是800个特征？增量搜索是什么意思？

We performed 10-fold cross validation to select discriminant features. Each time, we selected a feature set $S_i, i = 1, ..., 10$ enclosing 800 features by incremental search (Peng *et al.*, 2005), in which the features are arranged in descending order of characterizing power. The common features $\bar{S} = \bigcap_{i=1}^{10} S_i$ are taken as the feature candidates. With the purpose of determining the optimal number of features m, we give the common features of the first $50k$ features in $S_i, i = 1, ..., 10$ to the classifier, where k is the iteration number. The common features that yields the best recognition performance is considered as the optimal feature set. As shown in Figure 5.10, there are 169 common features in the first 300 feature candidates (k=6), and these 169 features are adopted as *"the best m features"* since they achieved the best recognition rates.

300个候选特征，169个公共特征

common features与feature candidates的关系是怎样的，是as还是in？

5.3 Classification

Support vector machines (SVMs) have proven to be powerful for facial expression classification. SVM achieves the best performance according to a comprehensive study (Shan *et al.*, 2009). Therefore, we adopt SVMs as the classifiers for facial expression recognition. SVMs attempt to find the hyperplane that maximizes the margin between the positive and negative observations for a specified class. Given a training set of labelled examples $\{(x_i, y_i), i = 1, ..., k\}$ where $y_i \in \{-1, 1\}$, a testing example x is labelled by the following function:

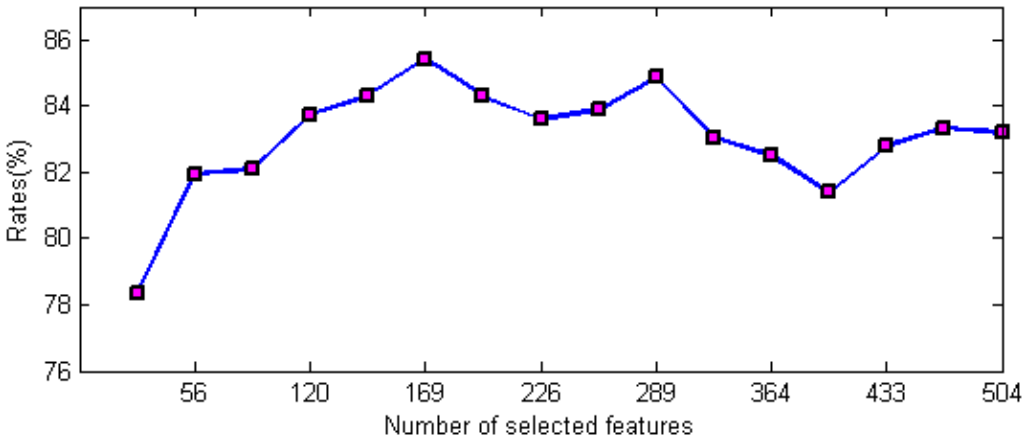


Figure 5.10: Recognition rates of different size of selected features.

Table 5.3: Confusion matrix of recognition on BU-3DFE database.

	AN	DI	FE	SA	HA	SU
AN	**80.9**±3.7	3.8	4.0	10.4	0.8	0.1
DI	8.0	**81.5**±2.7	5.3	1.6	2.7	0.9
FE	4.1	7.1	**70.8**±3.1	4.0	9.9	4.3
SA	13.0	1.7	5.3	**79.6**±3.1	0.4	0.0
HA	0.2	0.9	7.3	0.0	**91.1**±2.0	0.6
SU	0.3	1.5	3.0	0.5	0.7	**94.0**±1.7

$$f(x) = sgn(\sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b) \tag{5.7}$$

where α_i are Lagrange multipliers of a dual optimization problem that determine the classification hyperplane, $K(\cdot, \cdot)$ is a kernel function, and b is the threshold parameter of the hyperplane.

SVMs make binary decisions. However, there are six classes in facial expression recognition. Therefore, we use LIBSVM (Chang and Lin, 2011b) for the training and testing of SVMs, which achieves multi-classes classification according to the one-against-rest technique. With regard to the parameter selection, we carry out coarse-to-fine grid search in a 10-fold cross-validation on the training dataset.

5.4 Experimental Results

BU-3DFE database is one of the very few publicly available databases of annotated 3D facial expressions. It consists of 100 subjects (56 females and 44 males) from different ethnic ancestries and ages. Each subject has 25 facial scans, including one neutral face and 24 faces of 6 prototypic expressions with 4 levels of intensity. The 3D locations of 83 facial landmarks are provided for each 3D face. These manually labeled landmarks are widely used by most existing analysis algorithms.

The experiment is performed on 54-vs-6 setup, which is a commonly used protocol by most methods (Maalej *et al.*, 2011; Tekguc *et al.*, 2009; Tang and Huang, 2008; Wang *et al.*, 2006; Sha *et al.*, 2011). The samples of 60 subjects (30 females and 30 males) with two high-intensities for each expression (03 and 04), which are randomly selected from the 100 subjects in BU-3DFE. In order to conduct person-independent facial expression recognition, we randomly split these 60 subjects into 10 folds, take 9 folds (54 subject, 648 samples) as training data, and the remaining fold (6 subjects, 72 samples) as the testing data.

Following the process of other methods (Maalej *et al.*, 2011; Sha *et al.*, 2011) that used the BU-3DFE database, we select 60 subjects to form a 54-vs-6 setup. However, one issue is that precisely which 60 subjects are selected is never clearly specified by previous methods. This is an issue for performance comparison since different random samples of 60 subjects can give very different results and the selection of 60 "*easy*" faces can give very high accuracy. To ensure unbiased experimental results, we perform 20 random selections and conduct 10-fold cross validation on each of the 20 sets. Thus our total experiments are $20 \times 10 = 200$. The recognition results for each of the 20 times are shown in the box plot of Figure 5.11. Note the significant variations in expression recognition between different sets of 60 subjects.

The recognition rates and stand derivations across all 20 random selections of 10-fold experiments are averaged and reported in Table 5.3. The proposed method achieved a 83% average recognition rate for the six prototypic expressions. The major confusions are 13.0% (sadness is misclassified as anger), 10.4% (anger is misclassified as sadness) and 9.9% (fear is misclassified as happiness).

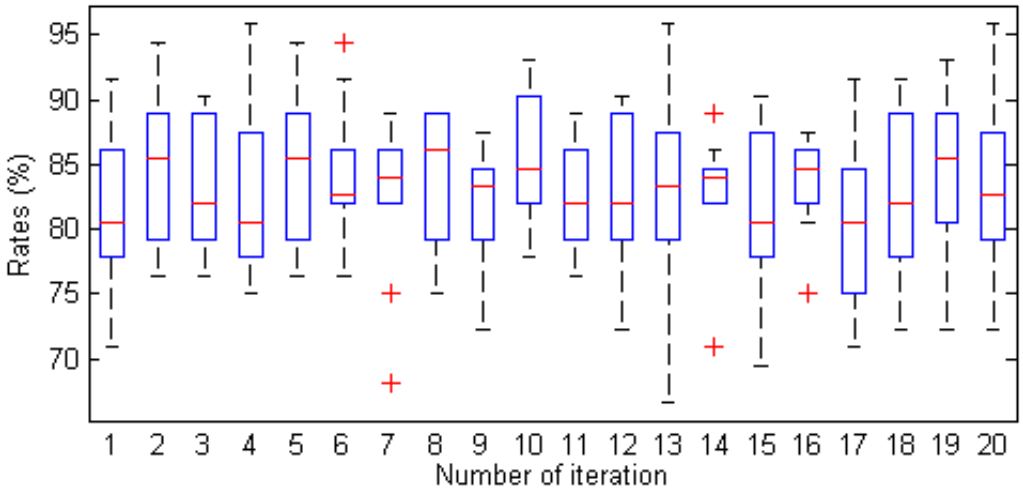


Figure 5.11: Boxplot of 20 times repeated 10-fold cross validation results of the proposed method.

5.4.1 Comparisons

Table 5.4 compares our method to existing manual and automatic 3D facial expression recognition techniques. The first row in the table reports the results of the performance of human experts on the same BU-3DFE database using the same two expression intensity levels i.e. 03 and 04. These experiments were performed by two psychologists who are the experts in human facial expression research (Yin *et al.*, 2006a). It is surprising to see that not even humans have perfect accuracy on this database and even more surprisingly, the method proposed by Maalej *et al.* (2011) performs better than humans. Our method has the best performance among automatic methods and compares well with other manual methods except Maalej *et al.* (2011). It is worth noting that our results are averaged over 20 random picks of 60 subjects multiplied by 10-fold experiments for each pick, whereas the results reported by others are based on a single random pick of 60 subjects.

5.4.2 Analysis and Discussion

Although landmark detection on 3D face models remains an open problem, it is inevitable in designing a fully automatic facial expression analysis system. The experimental results reveal that our method outperforms existing automatic techniques (Lemaire *et al.*, 2011, 2013), with better accuracy for every single expression (except sadness). In addition, clearly the errors in our heuristic points make the recognition task much more difficult

Table 5.4: Comparison between the proposed method and other 3D facial expression recognition approaches. The type of "manual" means the landmarks used in the corresponding method are manually labeled, while "auto" means points are automatically detected or not necessary.

Method	Type	AN	DI	FE	SA	HA	SU	Avg
Yin *et al.* (2006a)	Human expert	*94.9*	*95.4*	*96.4*	*96.2*	*99.4*	*99.0*	*96.8*
Maalej *et al.* (2011)	Manual	97.9	99.2	99.7	99.3	98.6	98.2	98.8
Tekguc *et al.* (2009)	Manual	86.0	87.3	85.3	82.9	93.4	94.7	88.2
Tang and Huang (2008)	Manual	86.7	84.2	74.2	82.5	95.8	99.2	87.1
Wang *et al.* (2006)	Manual	80.0	80.4	75.0	80.4	95.0	90.8	83.6
Sha *et al.* (2011)	Manual	78.7	83.9	69.8	84.8	88.5	95.4	83.5
The proposed	Auto	**80.9**	**81.5**	**70.8**	**79.6**	**91.1**	**94.0**	**83.0**
Lemaire *et al.* (2013)	Auto	74.1	74.9	64.6	74.5	89.8	90.9	78.1
Lemaire *et al.* (2011)	Auto	69.4	78.2	42.8	82.9	88.8	92.5	75.8

than methods based on manual landmarks. However, as shown in Table 5.4, the depth feature extracted around 30 heuristic points still achieved comparable results to many manual techniques (Maalej *et al.*, 2011; Tekguc *et al.*, 2009; Tang and Huang, 2008; Wang *et al.*, 2006; Sha *et al.*, 2011) which use the manually labeled 83 landmarks.

The local depth feature utilized in proposed method facilitates an effective feature selection. That is why we can achieve comparable performance with those methods using manual landmarks. The discrete depth features are projected to a subspace by 2DPCA for dimension reduction. By discarding the redundant dimensions, the resulting feature vectors conserve most of the essential information with large variance. However, the variances in the resulting feature vector are not purely caused by facial expressions. It also contains the changes caused by the facial differences in subjects and the inaccuracy of the heuristic points. We performed facial expressions on the *'contaminated features'*, and only achieved an average recognition rate of 75.4%. This shows that feature selection is vital to our performance, increasing the accuracy significantly to 83%.

5.5 Chapter Summary

This paper presented a fully automatic 3D static facial expression recognition method using local patch-based depth features. We extracted depth feature around 30 heuristic points, generated from 5 fiducial points, to represent facial expressions. A multi-class SVM was trained to classify the expressions based on the extracted feature after mRMR feature selection. The proposed method outperformed existing fully automatic methods by a significant margin.

Chapter 6

Automatic 4D Facial Expression Recognition

Ever since the public release of the BU-4DFE database (Yin *et al.*, 2008) which contains 4D expression data (i.e., dynamic 3D facial expression sequences), the temporal component is introduced in 4D facial expression recognition, and lots of works have been done on facial expression recognition using 3D dynamic sequence (Sun *et al.*, 2008; Sandbach *et al.*, 2012a; Fang *et al.*, 2011; Reale *et al.*, 2013). Temporal information has been shown to be able to significantly improve the accuracy of expression recognition (Sandbach *et al.*, 2011), which motivates the use of video sequences. Sun *et al.* (2008) proposed a facial expression classification based on frame-by-frame features, and it only achieves an average recognition rate of 65.1%. Such low performance is due to the frame-by-frame feature extraction, which may not be able to capture the temporal information sufficiently.

不仅要记录脸部的形变，还要记录形变的时间，是不是这个意思？

In fact, facial expression is inherently a spatio-temporal process, which means that an effective feature extraction should be able to extract not only the deformation of facial features, but also the relative timing of facial actions as well as their temporal evolution. Therefore, it is essential to measure the dynamics of facial expressions. For such purpose, Le *et al.* (2011) used facial level curves to extract spatio-temporal features by comparing the curves across frames using Chamfer distances. This method achieves an average recognition rate of 92.2% when tested on three expressions from the BU-4DFE database: sadness, happy and surprise. Another 4D spatio-temporal "Nebula" feature is proposed by Reale *et al.* (2013) to improve expression and facial movement analysis performance, in which the spatio-temporal volume is voxelized and represented by histogram of curvatures. This method achieves an average recognition rate of 76.9%, with noticeable high recognition rates on happy and surprise, but much lower recognition rates on those easily-confused expressions such as anger and sadness (Pantic and Rothkrantz, 2000; Fasel and Luettin, 2003; Xue *et al.*, 2014).

频域怎样理解？

In this chapter, we propose a method to extract spatio-temporal features of facial expression dynamics by analyzing 4D data in frequency domain. Inspired by the success of discrete cosine transform in video compression, 3D-DCT is applied on the local depth

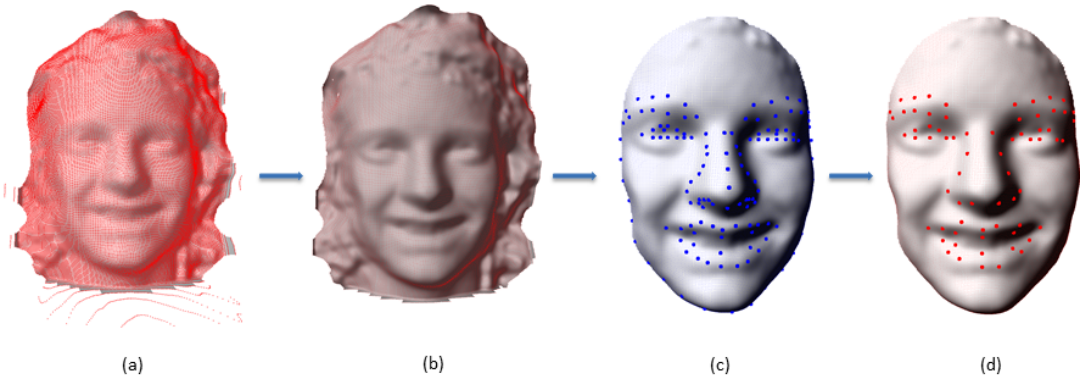


Figure 6.1: Pre-processing of BU-4DFE face model. (a) Raw face model of BU-4DFE database, the red dots are the vertices. (b) The denoised face model. (c) The cropped facial area, with 130 detected landmarks. (d) The 68 selected fiducial points for feature extraction.

这130个点是不是来源于“五棵树”算法在2D图片上探测的130个点？

patch-sequence generated from the original sequences based on the automatic detected facial landmarks. The compact low-frequency 3D-DCT coefficients are selected as the feature vector, which can represent expression dynamics without loss of the subtle information conveyed by those easily-confused expressions. The extracted features are classified by the nearest-neighbor classifier after feature selection and dimension reduction by linear discriminant analysis (LDA) (Belhumeur *et al.*, 1997). The experimental results show that the proposed method achieves an average recognition rate of 78.8%, by improving the recognition rate of anger to 85.0% and sadness to 78.0% significantly.

6.1 Data Pre-processing

The raw 3D face models (frames of the expression sequences) in BU-4DFE database are quite noisy, as illustrated in figure 6.1(a). The face models contain the very obvious outlier vertices at the bottom part, which has a significant impact on landmark detection and feature extraction. In the pre-processing step, we first apply noise filtering to remove the outlier vertices and then implement landmark point detection.

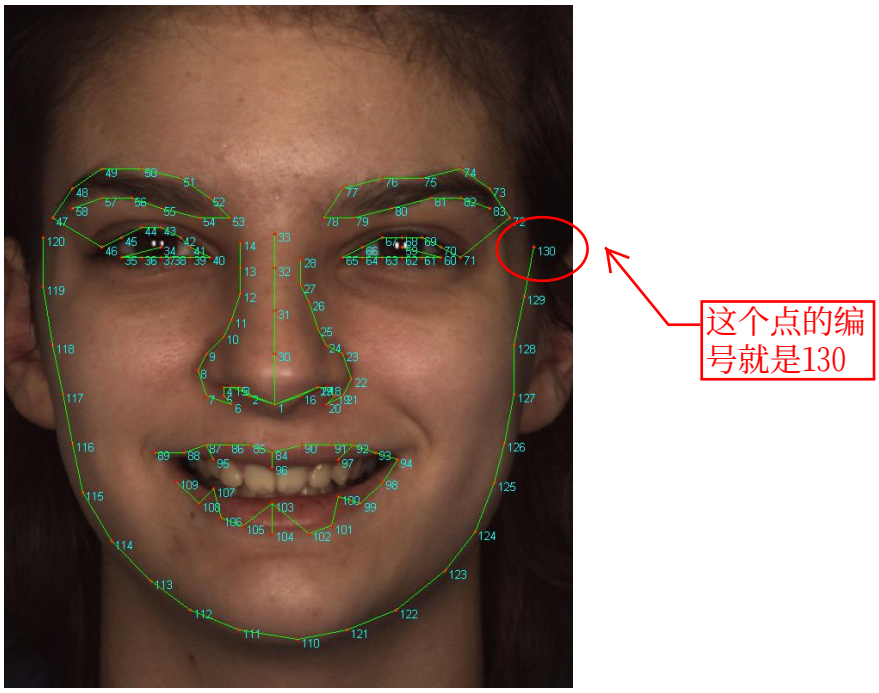


Figure 6.2: Tree structure of the landmark detection model. It contains 130 landmark points, and 5 trees covering nose, left eye, right eye, mouth and face contour.

6.1.1 Noise Filtering

这里是说什么点很钝？

Figure 6.1(a) shows that the vertices of the facial part on the BU-4DFE face models are very dense, while the outliers are quite sparse. Thus, we design a filter based on the length of edges connecting model vertices, since the edges connecting outliers will be much longer than those connecting the normal vertices on facial part. For a given face model, the designed filter calculates the mean length m and standard deviation std of all the edges. Any vertex corresponding to edges longer than $m + 5 \cdot std$ will be filtered out.

The denoised facial model is shown in Figure 6.1(b). It still includes hair and neck. However, only facial area is of interest for expression analysis. In order to crop the facial area out from the denoised model, the vertex which has the largest z-axis value is considered as the nose tip. A sphere with radius of 70mm is centered at this vertex to crop the facial area out (see Figure 6.1(c)).

从图中看，截取的区域不是正圆，实际操作中是如何计算的？

6.1.2 Landmark Point Detection

In order to extract local features from the denoised 3D face model automatically, we need to identify the landmark points around important facial components. Currently, robust landmark detection on 3D faces is still an open problem due to the significant topology changes caused by expressions, such as mouth opening in surprise. However, every 3D face model (i.e., a frame of the expression dynamic sequence) in BU-4DFE database is released with a 2D image texture, and the correspondence between the 3D model vertices and 2D image pixels is recorded in a model file. Thus, 2D facial landmarks can be detected on the texture image and then the 3D positions of the landmarks can be located on the face model through the correspondence between pixels and vertices.

这句话怎样理解？

The facial landmark localization method proposed by Zhu and Ramanan (2012) is utilized to detect the 2D landmarks in the texture. As the tree-structure model is trained by covering the whole face in one tree, it is not accurate enough to characterize some extreme expressions contained in the BU-4DFE database. In order to improve the detection accuracy over different expressions, we retrained 5 tree-structured models to detect 130 landmarks, which covers the nose, right eye, left eye, mouth and face contour, as shown in Figure 6.2. The 130 detected landmarks are then back-projected onto the corresponding 3D model (see in Figure 6.1(c)). In order to compare with the existing results achieved using ground truth, 68 of the 130 detected points are picked out for feature extraction.

在2D图像上，利用五棵树模型得到130个点，再将这些点对应到3D模型上

为什么要从130个点中选出68个点？选择的标准是什么？

6.2 Feature Extraction

Each facial expression sequence included in BU-4DFE database normally contains about 100 frames, each of which is a 3D face model. After performing landmark detection on each frame, a sliding window with the width of 16 frames and sliding stride of 4 frames is applied on each of the expression sequences to generate a group of subsequences. Obviously, every subsequence generated contains 16 consecutive frames. Unlike the methods which extract expression features frame-by-frame (Sun *et al.*, 2008; Sun and Yin, 2008), this paper proposes to extract features from frame sequences that can represent spatio-temporal facial expression dynamics. This is accomplished by applying 3D-DCT on the local depth patch-sequences.

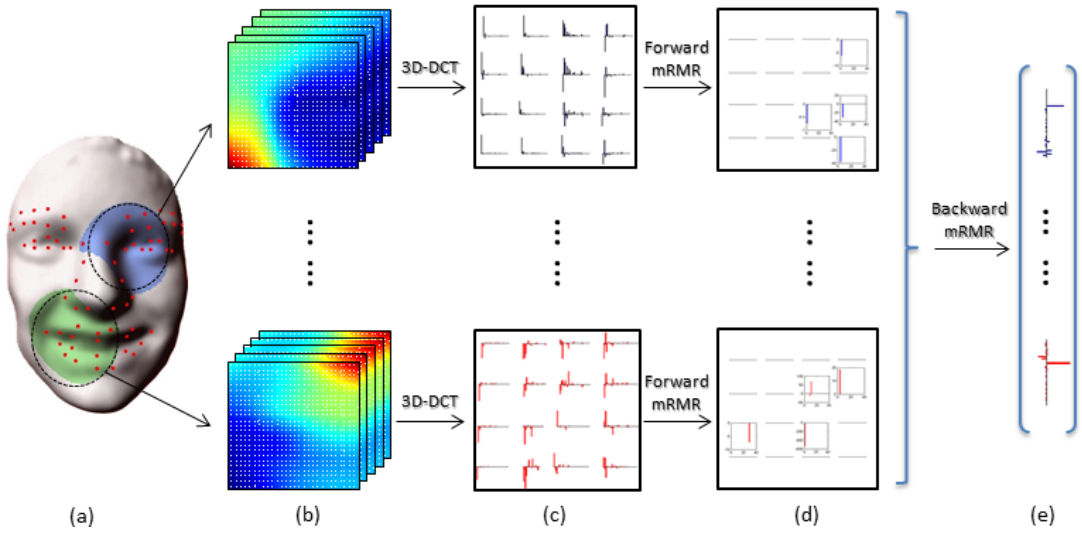


Figure 6.3: Schema of feature extraction and selection. (a) One cropped 3D face frame with 68 landmarks. Two cubic patches are fitted to the point cloud around left inner eye corner (blue patch) and right mouth corner (green patch). (b) Local depth features are sampled from the fitted patch, and one patch-sequence is formed by putting the sampled depth feature around same fiducial points(left inner eye corner or mouth corner) from consecutive frames together. (c) 3D-DCT coefficients of the patch-sequence. (d) The forward mRMR feature selection is applied on 3D-DCT coefficients of each patch-sequence, and the "best m coefficients" are shown. (e) The selected features of every patch-sequence are putting together, and the backward feature selection is applied to determine the optimal feature set for whole face.

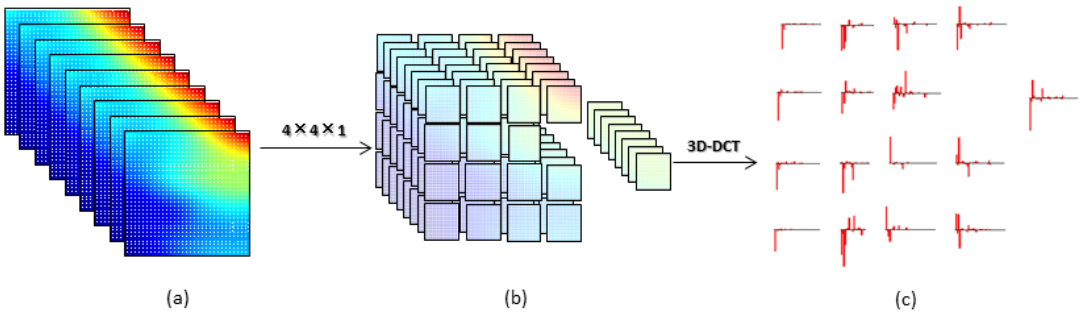


Figure 6.4: Demonstration of 3D-DCT on one depth patch-sequence. (a) One local depth patch-sequence. (b) The patch-sequence is divided equally into $4 \times 4 \times 1$ cells. (c) The 3D-DCT coefficients, one bar chart represents the selected 29 low-frequency coefficients from one cell in (b).

Table 6.1: The selected frequency coefficients of 3D-DCT.

Frequency	(u, v, w)			
DC	$(0,0,0)$			
AC	$(0,0,1)$	$(0,0,2)$	$(2,0,1)$	$(2,2,0)$
	$(0,1,0)$	$(0,2,0)$	$(2,1,0)$	$(1,2,2)$
	$(1,0,0)$	$(2,0,0)$	$(1,1,2)$	$(2,1,2)$
	$(0,1,1)$	$(0,1,2)$	$(1,2,1)$	$(2,2,1)$
	$(1,0,1)$	$(0,2,1)$	$(2,1,1)$	$(0,0,3)$
	$(1,1,0)$	$(1,0,2)$	$(0,2,2)$	$(0,3,0)$
	$(1,1,1)$	$(1,2,0)$	$(2,0,2)$	$(3,0,0)$

6.2.1 Local Depth Patch-sequence

As illustrated in Figure 6.3(a), once a fiducial point is located, we use a sphere with radius r centered at this point to crop a cluster of points. Next, a cubic patch is fitted to the cropped points using the grid-fitting code from D'Errico (2006). The fitted patch is then sampled on a uniform 74×74 grid, but only the central 64×64 samples covering the points in r region are kept as the patch-based depth feature, in order to avoid the artifacts at boundaries. All sampled patches end up with equal resolution which is necessary for the following feature extraction. For a given subsequences with 16 consecutive frames, the extracted 64×64 patches which around the same fiducial point of every frame yield a depth patch-sequence with the dimension of 64×64×16, as shown in Figure 6.3(b). Thus, each expression subsequence is represented by 68 local depth patch-sequences, since there are 68 fiducial points on every frame.

6.2.2 3D Discrete Cosine Transform

Inspired by the success of discrete cosine transform in image and video compression (Servais and De Jager, 1997; Chan and Siu, 1997), we propose an alternative expression dynamics representation based on the 3D discrete cosine transform (3D-DCT) which has a set of fixed projection bases (i.e., cosine basis functions). Using these fixed projection bases, it is only necessary to compute the corresponding projection coefficients (3D-DCT coefficients) to represent the local depth patch-sequences. In fact, 3D-DCT leads to a object representation with sparse high-frequency transform coefficients if a signal is self-correlated in both spatio and temporal dimensions, which is desirable for depth patch-

Algorithm 1:	**Forward Feature Selection**
Input:	Frequency coefficients of 3D-DCT of depth patch-sequence.
Initialization:	LDA classification error: $E = 1$, Based on mutual information, Candidate feature set: $C_0 = \{x_i \| 1 \leq i \leq n\}$, Optimal feature set: $S_0^* = \Phi$, $cnt = 1$;
While: $cnt \leq n$	
	For $k = 1 : sizeof(C_{k-1})$ 1. $S_k^* = \{S_{k-1}^*, x_k\}$, 2. do LDA classification based on S_k^*, record error $e(k)$, 3. $i = argmin\ e(k)$, **End** **If** $min\ e(k) \leq E$ 1. $E = min\ e(k)$, 2. $S_k^* = \{S_{k-1}^*, x_i\}$, 3. $C_k = \{C_{k-1}^* - x_i\}$, **Else** Break, **End** $cnt = cnt + 1$
End	
Ouput:	Optimal feature set: S_k^*

sequence representation. That is to say, by discarding these high-frequency coefficients, we simultaneously obtain a compact 3D-DCT based expression sequence representation and the reconstruction error introduced by removing a subset of high-frequency coefficients is typically small.

The 3D-DCT is based on a set of cosine basis functions which are determined by the dimensions of the 3D signal and thus independent of the input video data. The goal of the discrete cosine transform is to express a discrete signal, such as the depth patch-sequence, as a linear combination of mutually uncorrelated cosine basis functions, each of which encodes frequency-specific information of the discrete signal. In general, 3D-DCT can be applied to 3D signal $((f_{3D}(x,y,z))_{N_1\times N_2\times N_3})$ as following:

$$\begin{aligned} C_{3D}(u,v,w) = \alpha_1(u)\alpha_2(v)\alpha_3(w) \sum_{x=0}^{N_1-1}\sum_{y=0}^{N_2-1}\sum_{z=0}^{N_3-1} f_{3D}(x,y,z) \\ \cdot \left\{ cos\left[\frac{\pi(2x+1)u}{2N_1}\right] cos\left[\frac{\pi(2y+1)v}{2N_2}\right] cos\left[\frac{\pi(2z+1)w}{2N_3}\right] \right\} \end{aligned} \tag{6.1}$$

where $u \in \{0,1,\cdots,N_1-1\}$, $v \in \{0,1,\cdots,N_2-1\}$, $w \in \{0,1,\cdots,N_3-1\}$, and $\alpha_k(u)$ is defined as

$$\alpha_k(u) = \begin{cases} \sqrt{\frac{1}{N_k}}, & if\ u=0; \\ \sqrt{\frac{2}{N_k}} & otherwise. \end{cases} \tag{6.2}$$

For each patch-sequence with the dimension of $64 \times 64 \times 16$, we divide this volume into 16 cells ($4 \times 4 \times 1$, as shown in Figure 6.4(b)). Each cell has a dimension of $16 \times 16 \times 16$, on which 3D-DCT is applied. The DC coefficient and 28 low-frequency AC coefficients are kept in a 29-dimensional vector to represent the corresponding cell. Table 6.1 records the index of the selected coefficients of 3D-DCT. As shown in Figure 6.4(c), each small bar chart (3D-DCT coefficients) represents one cell, and all these 16 bar charts together represent the corresponding local depth patch-sequence.

6.3 Feature Selection and Classification

In pattern recognition problems, identifying the most characterizing features of the observed data is crucial to minimize the classification error. The idea of feature selection is that a simple combination of individually good features does not necessarily lead to good classification performance. That is to say, "the m best features are not the best

m features" (Peng *et al.*, 2005). We adopt the framework of the *minimal-redundancy-maximal-relevance* (mRMR) (Peng *et al.*, 2005; Ding and Peng, 2005) to select the best features for recognition as following:

First, the mRMR criterion is used to select mutually exclusive features $S = \{x_1, ..., x_m\}$ that jointly have the largest characterizing power on each of the six prototypic expressions class c:

$$\begin{cases} max\Phi(D, R) = D(S, c) - R(S), \\ D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c), \\ R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \end{cases} \tag{6.3}$$

where $I(x_i; c)$ is the mutual information value between individual feature and class, and $I(x_i, x_j)$ is the mutual information value between two features.

When the candidate features are selected, the next task is to determine the optimal number of features m. A wrapper that tests features with a nearest-neighbor (NN) classifier is utilized to determine the size of the feature set, with the direct goal of minimizing the recognition error of the specific classifier on the training set.

In the proposed method, we perform 2-tier feature selection from the bottom up, corresponding to patch-level and face-level. In the first tier, mRMR feature selection (Peng *et al.*, 2005) is applied on each patch-sequence to select the "best m features" of each patch. This is done by a forward feature selection algorithm (Algorithm 1). The forward feature selection will not stop until the recognition rate drops when adding one feature from the candidate feature set. The selected "best m features" of each patch are put together and fitted to second tier feature selection.

In the second tier, the selected features of all patches by forward algorithm are put together, and mRMR feature selection is applied on these features again to determine the "best m features" of the whole face. Since these features are pre-selected, we use backward feature selection algorithm (Algorithm 2) to condense the input features and refine the optimal feature set. Although Chen *et al.* (2013) pointed out that keeping some of the redundant features may be useful for classification, feature reduction is still needed because some of the features yield confusion. In Figure 6.5, the recognition rate and number of features of each iteration of the backward feature selection algorithm are plotted. It can be seen that the recognition rate keeps increasing when the features with low characterizing power are removed.

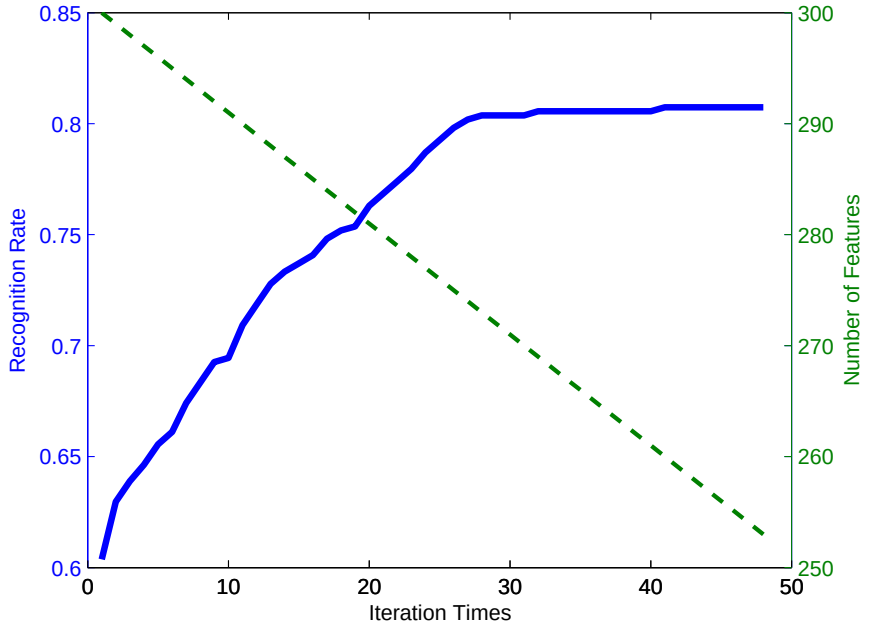


Figure 6.5: The recognition rate of backward feature selection.

Finally, the features contained in the optimal feature set by backward feature selection is considered as the "best m features", for the purpose of expression classification. The dynamic 3D facial expression sequences are finally represented by these optimal features, and classified by nearest-neighbor classifier after dimension reduction by linear discriminant analysis (LDA) (Belhumeur *et al.*, 1997).

6.4 Experimental Results

6.4.1 Experiment Setup

The BU-4DFE database in the first dataset consists of faces in 3D videos. It involves 101 subjects (58 females and 43 males) of various ethnicities. For each subject the six prototypic expressions (Angry, Disgust, Fear, Happy, Sad and Surprise) (Ekman and Friesen, 1978) were recorded gradually from a neutral face, onset, apex, offset and back to neutral, using the dynamic facial acquisition system Di3D and producing roughly 60,600 3D face modes(frames), with corresponding texture images. Each prototypic 3D expression video sequence lasts about four seconds. The temporal resolution of the 3D videos is 25 fps and

Algorithm 2: **Backward Feature Selection**

Input: The union of the optimal feature sets from forward feature selection.

Initialization: LDA classification error: $E = 1$,
Optimal feature set:
$S_0^* = \{x_i | 1 \leq i \leq n\}$,
$cnt = 1$;

While: $cnt \leq n$

For $k = 1 : sizeof(S_{k-1}^*)$
1. $S_k^* = \{S_{k-1}^* - x_k\}$,
2. do LDA classification based on S_k^*, record error $e(k)$,
3. $i = argmin\ e(k)$,

End

If $min\ e(k) \leq E$
1. $E = min\ e(k)$,
2. $S_k^* = \{S_{k-1}^* - x_i\}$,

Else
Break,

End

$cnt = cnt + 1$

End

Ouput: Optimal feature set: S_k^*

Table 6.2: Confusion matrix of 6 prototypic expressions recognition with 3D-DCT feature on the BU-4DFE database.

	AN	DI	FE	SA	HA	SU
AN	**85.0**	10.4	1.3	2.3	0.2	0.7
DI	9.3	**74.6**	3.8	1.8	3.6	7.0
FE	4.3	6.4	**62.0**	4.8	12.6	9.9
SA	17.0	2.1	1.3	**78.0**	0.4	1.1
HA	2.4	3.7	6.1	0.0	**86.2**	1.6
SU	0.4	3.2	5.8	0.7	2.8	**87.1**

Table 6.3: Comparison of 6 prototypic expressions recognition performance on the BU-4DFE database.

	AN	DI	FE	SA	HA	SU	AvgRates
Sun and Yin (2008)	92.4	87.6	85.4	97.8	80.3	99.3	90.4
Xue *et al.* (2014)	71.8	56.3	51.7	72.3	66.8	63.5	63.8
Sandbach *et al.* (2012a)	48.2	66.1	46.2	57.1	88.2	82.6	64.7
Fang *et al.* (2011)	68.3	79.7	67.9	81.3	71.6	96.1	75.8
Reale *et al.* (2013)	76.3	74.0	60.0	70.9	90.8	89.8	76.9
The proposed	**85.0**	**74.6**	**62.0**	**78.0**	**86.2**	**87.1**	**78.8**

each 3D model consists of approximately 35,000 vertices. Each frame is released with 83 manually labeled facial landmark points. In order to fairly compare with the existing results which using the manually labeled ground truth landmarks, we select 68 fiducial points (see red points in Figure 6.1(d)) from the detected 130 landmarks for feature extraction, which locate on the similar positions with manually labeled landmarks.

The proposed method is tested on a subset of the BU-4DFE database. We select 60 subjects (30 females and 30 males) randomly from the BU-4DFE database. In order to guarantee person-independent facial expression recognition, we follow the commonly used 54-vs-6 setup (Sun and Yin, 2008; Wang *et al.*, 2006; Xue *et al.*, 2014), which means that the 3D expression sequences from 54 subjects are chosen as the training set, and the remaining 6 subjects' sequences are taken as the testing set. To ensure unbiased experimental results, we conduct 10-fold cross validation in all our experiments and the average recognition rates are reported.

6.4.2 Expression Recognition Results

The first experiment is to recognize the six prototypic facial expressions sequences in the BU-4DFE database. We generate 15 subsequences (16-frame length) from each original expression sequence, and extract 3D-DCT features to represent these subsequences. The 2-tier feature selection is applied on the extracted features, followed by dimension reduction using linear discriminant analysis, and finally classified by nearest-neighbor classifier. Table 6.2 records the recognition confusion matrix of the proposed method. 17.0% of the sadness expression are misclassified as anger, but the worst recognition happens on the fear expression.

Table 6.4: Confusion matrix of recognition AN-HA-SU expressions on the BU-4DFE database based on proposed 3D-DCT features.

	AN	HA	SU
AN	**100**	0	0
HA	0.1	**99.9**	0
SU	0.1	0	**99.9**

The proposed method is compared with some existing approaches in Table 6.3. The method proposed by Sun and Yin (2008) achieves an average recognition rate of 90.4% on BU-4DFE database. However, it is not automatic because the generic model adaptation is controlled by 83 pre-defined ground truth landmarks. We implement the method proposed by Xue *et al.* (2014) and extract local depth feature frame-by-frame on BU-4DFE database, the average recognition rate is only 63.8%. A close recognition rate 65.1% is achieved by the work of Sun *et al.* (2008), in which the classification is performed on a frame-by-frame basis rather than by constructing spatio-temporal features. This is to show that feature extraction frame-by-frame is insufficient to represent expression dynamics accurately. Instead, Reale *et al.* (2013) proposed a 4D spatio-temporal "Nebula" feature to improve facial expression analysis performance, and achieves an average rate of 76.9%, with noticeably recognition rates of 90.8% and 89.8% on happy and surprise respectively. But the recognition rates of the easily-confused expressions, such as anger, fear and sadness, are relatively lower. This is because the histogram of the curvature features is not sensitive enough to reflect the subtle changes conveyed by the easily-confused expressions. In contrast, the proposed method achieves a higher average recognition rates 78.8% over all the 6 prototypic expressions. Furthermore, the recognition of rates of the easily-confused expression are improved significantly, especially on anger expression 85.0% and sadness expression 78.0%. Despite the 2-tier feature selection, the performance improvement attributes to the 3D-DCT features extracted from depth patch-sequences, which can preserve most of the changes caused by facial expression in compact low-frequency coefficients.

The methods proposed in (Fang *et al.*, 2011; Le *et al.*, 2011; Sandbach *et al.*, 2011) propose to recognize only 3 expressions, either anger, happy and surprise (AN-HA-SU) or sad, happy and surprise (SA-HA-SU). We also conduct similar experiments following this setup. Table 6.4 and table 6.5 record the confusion matrices of the recognition of AN-HA-SU and SA-HA-SU expressions. Table 6.6 compares the performance of the recognition rates over expression group AN-HA-SU and SA-HA-SU, which reveal that the proposed method outperforms other existing approaches in 3-class expression recognition.

Table 6.5: Confusion matrix of recognition SA-HA-SU expressions on the BU-4DFE database based on proposed 3D-DCT features.

	SA	HA	SU
SA	**100**	0	0
HA	0.2	**99.8**	0
SU	1.7	0.9	**97.4**

Table 6.6: Comparison of 3-class expressions recognition base on the BU-4DFE database.

	AN	HA	SU	SA	HA	SU
Fang *et al.* (2011)	97.3	96.3	96.5	98.9	97.3	91.0
Le *et al.* (2011)	-	-	-	91.7	95.0	90.0
Sandbach *et al.* (2011)	76.3	89.1	83.7	-	-	-
The proposed	**100**	**99.9**	**99.9**	**100**	**99.8**	**97.4**

6.4.3 Discussion

In 6-class recognition, misclassification happens among the easily-confused expressions, such as anger, disgust, fear and sadness. The reason is illustrated in Figure 6.6, which shows the expression samples in the subspace after LDA projection. Figure 6.6(a) shows that in the projected subspace the 'cloud' of fear samples (red triangles) has intersections with the 'cloud' of all the rest 5 expressions, especially with 'cloud' of happy samples, which is the reason why 12.6% of the fear expression samples are misclassified as happy. Similarly, the 'cloud' of anger samples (green plus) lies between the 'cloud' of disgust and sadness, with great overlaps. This explains that 10.4% anger expression samples are misclassified as disgust, and 17.0% of sadness expression samples are misclassified as anger.

While in 3-class recognition, there are only 2-dimension left after LDA projection. Either in SA-HA-SU classification or in AN-HA-SU classification, only one 'hard' expressions left, and almost has no confusion with happy and surprise. It can be seen in Figure 6.6(b) and Figure 6.6(c). As the result, the proposed method achieves an average recognition rate of 99.9% on AN-HA-SU classification and 99.1% on SA-HA-SU classification.

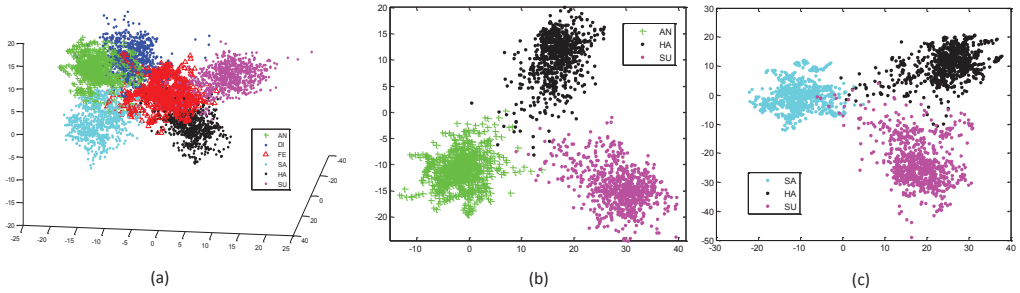


Figure 6.6: The expression samples projected into the subspace from LDA. (a) In 6-class recognition, the first 3 dimensions of the samples in LDA subspace are plotted. (b) In AN-HA-SU recognition, the subspace from LDA only has two dimensions, and the 3 expressions have little overlap. (c) In SA-HA-SU recognition, the subspace from LDA has two dimensions, the 3 expressions overlap slightly.

6.5 Chapter Summary

This paper proposed an automatic 4D facial expression recognition method based on dynamic 3D facial expression sequences. Three dimension discrete cosine transform is utilized to extract compact spatio-temporal features from local patch-sequence to represent facial expression dynamics. In order to get more characterizing features for classification, two rounds mRMR feature selection are applied on the extract 3D-DCT coefficients. The experimental results show that the proposed feature extraction method can preserve the subtle information conveyed by easily-confused expressions, and outperforms the existing methods, especially on the easily-confused expressions.

Chapter 7

Discriminative Expression Feature Selection in 4D data

4D包含哪些信息？

It has been noted that facial expressions are conveyed by different facial parts/components (Pardàs and Bonafonte, 2002; Bourel *et al.*, 2001). Pardàs and Bonafonte (2002) show that eyebrows and mouth are the components that carry the maximum amount of information relevant to expressions being conveyed, and mouth conveys more information than the eyebrows. Another work by Bourel *et al.* (2001) reveal that sadness is mainly conveyed by the mouth area. Similarly, Kotsia *et al.* (2008) show that the occlusion of the mouth reduces the recognition rate by more than 50%. This result is highly consistent with the results of Pardàs and Bonafonte (2002). These researches inspire us to ask: what parts/components of human face carry the information that can best distinguish the six basic expressions? In other words, what are the most expressive parts of human face that convey the most discriminative information for facial expression? Answers to these questions can be crucial for performance improvement of automatic expression recognition.

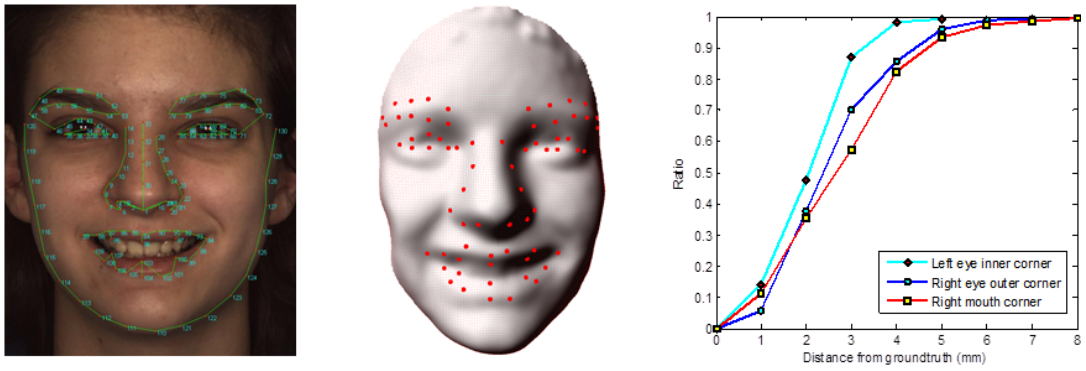
One possible approach to answer such questions is to recognize facial expressions under partial facial image occlusion, such as the work mentioned above (Pardàs and Bonafonte, 2002; Bourel *et al.*, 2001; Kotsia *et al.*, 2008). The impact on recognition rate reflects how discriminative the occluded facial area is, such as the area of mouth in Bourel *et al.* (2001). However, the occlusion in these work (Pardàs and Bonafonte, 2002; Bourel *et al.*, 2001; Kotsia *et al.*, 2008) is manually imposed on the face in a fixed manner by applying a predefined binary mask on facial images empirically, which cannot be adjusted according to the different input images containing different expressions. Therefore, it is necessary to develop an algorithm that can search the most discriminative facial parts automatically for different image data.

这一章的创新点主要是能够自动进行occlusion

This chapter proposes an automatic method to determine the most discriminative facial parts for 4D facial expression recognition. The local depth patch-sequence is generated from automatically detected facial landmarks, and Histogram of Oriented 3D-Gradients (HOG3D) feature is then extracted to represent 3D facial deformation over time. Two-stage (forward and backward) mRMR feature selection is conducted to determine the most

sequence是否意味着处理的对象是一段视频？

3D梯度直方图特征是什么？



(a) (b) (c)

这个图表达不够严谨，上一章表述这个问题要好一些

Figure 7.1: Landmark detection on BU-4DFE face models. (a) Tree structure of the landmark detection model. It contains 130 landmark points, and 5 trees covering nose, left eye, right eye, mouth and face contour. (b) The 68 selected fiducial points for feature extraction on 3D models. (c) The accumulation ratio of the error distance from the detected landmarks to the corresponding groundtruth.

抽取的时空特征和frame-by-frame特征有什么不同？

discriminative facial parts. The propose method consists of: (1) an automatic local depth patch-sequence based method to extract spatio-temporal features (real 4D features) to represent 4D facial expressions rather than to extract features frame-by-frame; (2) a data-driven method for determining the most expressive facial parts for 6 basic expressions; (3) a hierarchical classifier which uses features of different facial parts in each tier. The proposed method outperforms the state-of-art approaches by a significance margin.

7.1 Feature Extraction

7.1.1 Landmark Point Detection

Every 3D face model (i.e., a frame of the dynamic expression sequence) in the BU-4DFE database (Yin *et al.*, 2008) is released with a 2D image texture, and the correspondence between the 3D model vertices and 2D image pixels is recorded in a model file. Thus, 2D facial landmarks can be detected on the texture image and then the 3D positions of the landmarks can be located on the face model through the correspondence between the pixels and vertices.

The facial landmark localization method proposed by Zhu and Ramanan (2012) is utilized to detect the 2D landmarks in the texture image. As the tree-structure model is trained

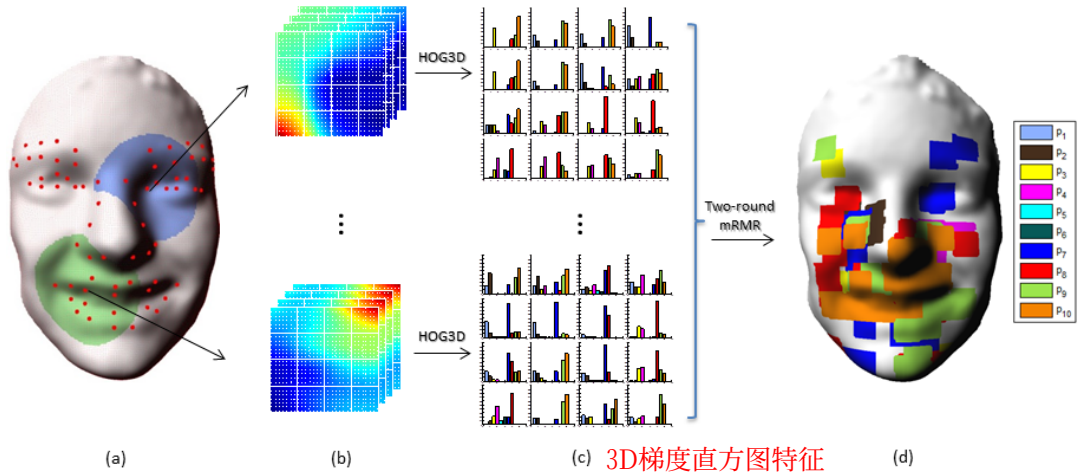


Figure 7.2: Schema of feature extraction and selection. (a) One cropped 3D face frame with 68 landmarks. Two cubic patches are fitted to the point cloud around left inner eye corner (blue patch) and right mouth corner (green patch). (b) Local depth features are sampled from the fitted patch, and one patch-sequence is formed by putting the sampled depth feature around same fiducial points(left inner eye corner or mouth corner) from consecutive frames together. (c) HOG3D features extracted from the patch-sequence. (d) After two-stage feature selection, the resulting expressive facial parts/components are plotted on a 3D face model. The color of the patch stands for its characterizing gradient's direction.

to cover the whole face in one tree, it is not accurate enough to characterize some extreme expressions contained in the BU-4DFE database. As same as in section 6.1.2, in order to improve the accuracy of landmark detections over different expressions, we retrained 5 tree-structured models to detect 130 landmarks, which covers the nose, right eye, left eye, mouth and face contour, as shown in Figure 7.1(a). The detected landmarks are then back-projected onto the corresponding 3D model (see in Figure 7.1(b)).

The distances from the detected landmarks to the corresponding groundtruth provided by BU-4DFE data are calculated to check the detection accuracy. Figure 7.1(c) illustrates the detection error of three points (left eye inner corner, right eye outer corner, and right mouth corner). It can be seen that over 80% of the errors are less than 4 mm.

7.1.2 Histogram of Oriented 3D-Gradients

Each facial expression sequence included in the BU-4DFE database normally contains about 100 frames, each frame is a 3D face model. After performing landmark detection on each frame, a sliding window with the width of 16 frames and sliding stride of 4 frames

这个16是怎么来的？

is applied on each of the expression sequences to generate a group of subsequences. Obviously, every subsequence generated contains 16 consecutive frames. Unlike the methods which extract expression feature frame-by-frame (Sun *et al.*, 2008; Sun and Yin, 2008), the proposed method extracts real 4D features from subsequences in an attempt to represent spatio-temporal facial expression dynamics. This is accomplished by extracting HOG3D features (Klaser and Marszalek, 2008) on the local depth patch-sequences. 3D梯度直方图特征

为什么要强调这句话？patch的分辨率本来就应该是相同的呀

As illustrated in Figure 7.2(a), once a fiducial point is located, a sphere with radius r is centered at this point to crop a cluster of points. Next, a cubic patch is fitted to the cropped points using the grid-fitting code from D'Errico (2006). The fitted patch is then sampled on a uniform 74×74 grid, but only the central 64×64 samples covering the points in r region are kept as the patch-based depth feature, in order to avoid the artifacts at boundaries. All sampled patches end up with equal resolution which is necessary for the subsequent feature extraction. For a given subsequences with 16 consecutive frames, the extracted 64×64 patches which are around the same fiducial point of every frame yield a depth patch-sequence with the dimension of 64×64×16, as shown in Figure 7.2(b). Each expression subsequence is represented by 68 local depth patch-sequences, since there are 68 fiducial points on every frame.

这里能不能理解为是68*64*64*16？
为什么这篇文章不将64*64合并为4096?

In order to generate the histogram of oriented 3D gradients, the 64×64×16 depth patch-sequence is divided equally into 4×4×1 cuboid cells, each of which has dimension of 16×16×16. For every cuboid cell $c(x, y, t)$, 3D gradients along x, y, t-directions are denoted by partial derivatives $[c_x', c_y', c_t']^T = [\frac{\partial}{\partial x}c, \frac{\partial}{\partial y}c, \frac{\partial}{\partial t}c]^T$, and the mean gradient is denoted by $\bar{\mathbf{g}}_\mathbf{c} = [\bar{c_x}', \bar{c_y}', \bar{c_t}']^T$. In order to calculate the histogram of 3D-Gradients, a *regular n-sided polyhedron* is centered at the origin of a three-dimensional Euclidean Coordinate. The mean gradient $\bar{\mathbf{g}}_\mathbf{c}$ is then projected on axes which go through the origin and the center points $\mathbf{p_i} = [x, y, t]$ of all n facets. Let P be the projection matrix of the n center points

n如何确定？

$$P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n]. \tag{7.1}$$

The projection **h** of the mean gradient vector $\bar{\mathbf{g}}_\mathbf{c}$ is calculated as follows:

h是不是就是直方图的柱子的高度？

$$\mathbf{h} = \frac{P \cdot \bar{\mathbf{g}}_\mathbf{c}}{\|\bar{\mathbf{g}}_\mathbf{c}\|_2}. \tag{7.2}$$

In this work, we choose $n = 20$ to make a regular icosahedron which contains 20 regular triangle facets. The center points of a icosahedron is

为什么要用20面体？

eta是什么意思？

$$\begin{aligned} &(\pm 1, \pm 1, \pm 1), \quad (0, \pm 1/\eta, \pm\eta), \\ &(\pm 1/\eta, \pm\eta, 0), \quad (\pm\eta, 0, \pm 1/\eta), \end{aligned} \tag{7.3}$$

depth patch-sequence --> cuboid cell --> 求偏导数 --> 求偏导数的均值 --> 构建投影矩阵 --> 投影计算 --> 计算的h值就是柱子的高度 --> 进行阈值处理

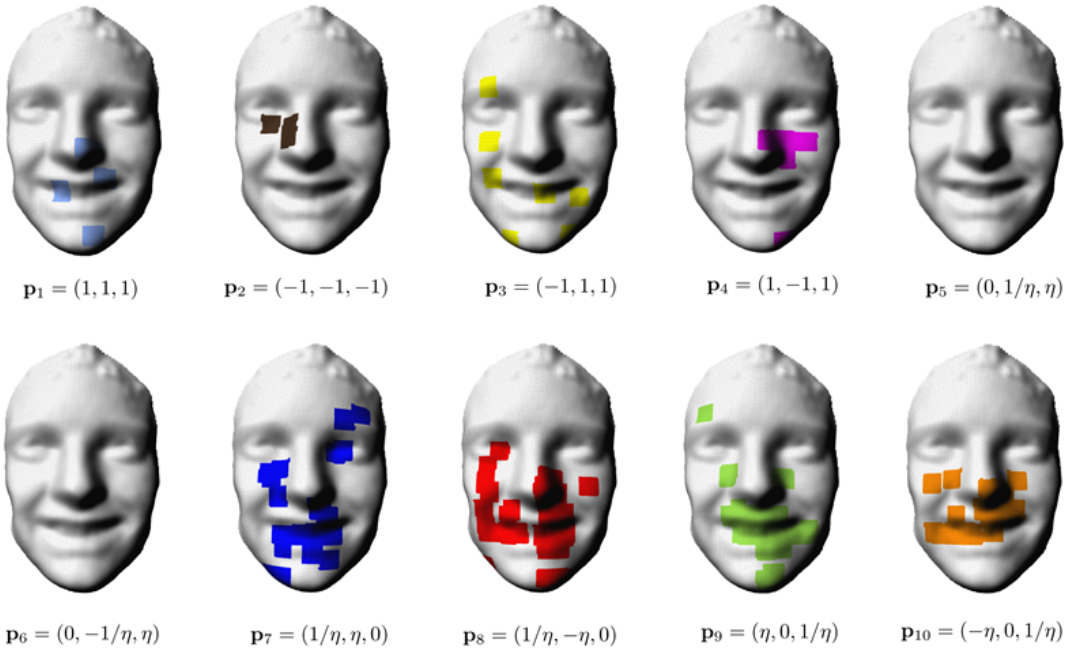


Figure 7.3: The oriented discriminative parts of face on six basic expression models. The colour stands for the selected orientation of the 3D-gradient feature in the corresponding region.

10是怎么来的？20 / 2 = 10，因为相对的两个面其实是同一根柱子

where $\eta = \frac{1+\sqrt{5}}{2}$ is the golden ratio. Each pair of opposite facets corresponds to one histogram bin since they are along the same axis. 二十面体的面跟直方图之间的关系？

For each cuboid cell, a 10-bin histogram is extracted, and it needs to be thresholded by $t = \mathbf{p}_i^T \cdot \mathbf{p}_j$ since the projection directions are not mutually orthogonal. Consequently, the HOG3D features of one depth patch-sequence are obtained by concatenating the histograms of corresponding 16 cells.

HOG3D特征是针对一个depth patch-sequence抽取的，最后的特征应该是一个柱状图的矩阵，即矩阵的每一个元素都是柱状图

这句话是什么意思？

7.2 Expressive Facial Parts Determination

7.2.1 Two-stage Feature Selection

As mentioned above, some parts of the face carry the maximum amount of information related to the facial expression being displayed, such as mouth and eyebrows. However, the amount of information carried in them are not sufficient for good expression recognition. A straightforward question is what parts of human face jointly carry the most discriminative

information for recognition of six basic expressions. That is to say, features from these parts/areas can jointly have the largest characterizing power on each of the six basic expressions, and consequently lead to the best recognition performance.

The *minimal-redundancy-maximal-relevance* (mRMR) feature selection framework (Peng *et al.*, 2005; Ding and Peng, 2005) is adopted here to identify the most discriminative facial parts for the recognition of the six basic expressions. The idea of feature selection is that a combination of individually good features does not necessarily lead to good classification performance. That is to say, "the m best features are not the best m features" (Peng *et al.*, 2005). In our case, the whole face is covered by 68 patches around corresponding landmarks. Assuming that a HOG3D feature set $S = \{h_1, ..., h_m\}$ can best distinguish the six basic expression, this set can be identified from bottom up by a two-stage feature selection process.

Firstly, the mRMR criterion is used to select mutually exclusive candidate features $Q = \{h_1, ..., h_n\}, n > m$ that jointly have the largest characterizing power on each of the six prototypic expressions class c. When the candidate features are selected, the next task is to determine the optimal number of features m. A wrapper that tests features with a classifier is utilized to determine the size of the optimal feature set, with the direct goal of minimizing the recognition error of the specific classifier on the training set. To ensure the generalized performance of the selected features, 10-fold cross-validation is conducted in each stage of the selection process.

In the first stage, mRMR feature selection is applied on each patch-sequence to determine the most discriminative sub-patch. This is done by a forward feature selection algorithm. In each iteration, the algorithm adds one feature from the patch according to the descending order of candidates features' characterizing power. This forward feature selection will not stop until the recognition rate starts to drop when adding one more feature from the candidate feature set. In the ith-fold cross-validation, the classification error is set to 1 initially, and Q_i is the candidate feature set. The wrapper first searches for a subset C_i^1 with one feature by choosing the feature h_1^* that renders the greatest error reduction. Then the wrapper selects the feature h_2^* from the set $Q_i - C_i^1$ so that the feature set $C_i^2 = \left\{C_i^1, h_2^*\right\}$ leads to the largest error reduction. This forward feature selection will continue until the classification error e begins to increase when adding a new feature h_{k+1}^* with $e_{k+1} > e_k$. That is to say, the optimal feature number of this patch $m_i = k$ and the optimal feature set $P_i = C_i^k$. Eventually, the features selected from all the 10 fold are put together to form the optimal feature set of one patch $P^* = P_1 \cup P_2 \cup \cdots \cup P_i \cup \cdots \cup P_{10}$. The parts/areas where P^* is selected are considered as the expressive parts of the current patch.

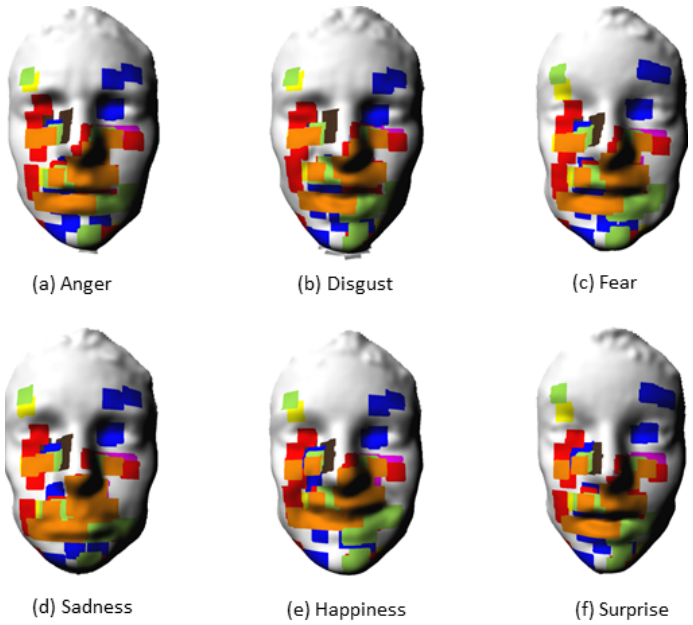


Figure 7.4: Visualization of the selected most discriminative regions of face models with six basic expressions.

In the second stage, the features of expressive parts selected by the forward algorithm are put together, and mRMR feature selection is applied on these features again to determine the expressive parts of the whole face. Since these features are from pre-selected candidate areas, we use a backward feature selection algorithm to condense the input features. This can be done by eliminating one candidate features in each iteration according to the ascending order of characterizing power. This process continues until the recognition rate starts decreasing. Similar as the first stage, a 10-fold cross-validation is conducted, and the union of the optimal features of each fold are preserved as "the m best features" for facial expression recognition.

Finally, the facial parts where the preserved features come from are considered as the most discriminative areas for expression recognition. The selected expressive facial areas and its 3D gradients orientations (represented by color) are shown in Figures 7.3. It shows the most discriminative/expressive parts of human face selected based on a data-driven algorithm with no pre-defined rules. We can see that the majority of the selected area covers mouth and eyebrows (see in Figure 7.3, directions $\mathbf{p}_3, \mathbf{p}_9, \mathbf{p}_{10}$), which perfectly matches the results of the occlusion related works (Pardàs and Bonafonte, 2002; Bourel *et al.*, 2001; Kotsia *et al.*, 2008). Moreover, the deformations of mouth area in the directions $\mathbf{p}_7$ to $\mathbf{p}_{10}$ are always considered as discriminative to human facial expressions by the proposed data-driven algorithm, which explains why mouth area carries more expression informa-

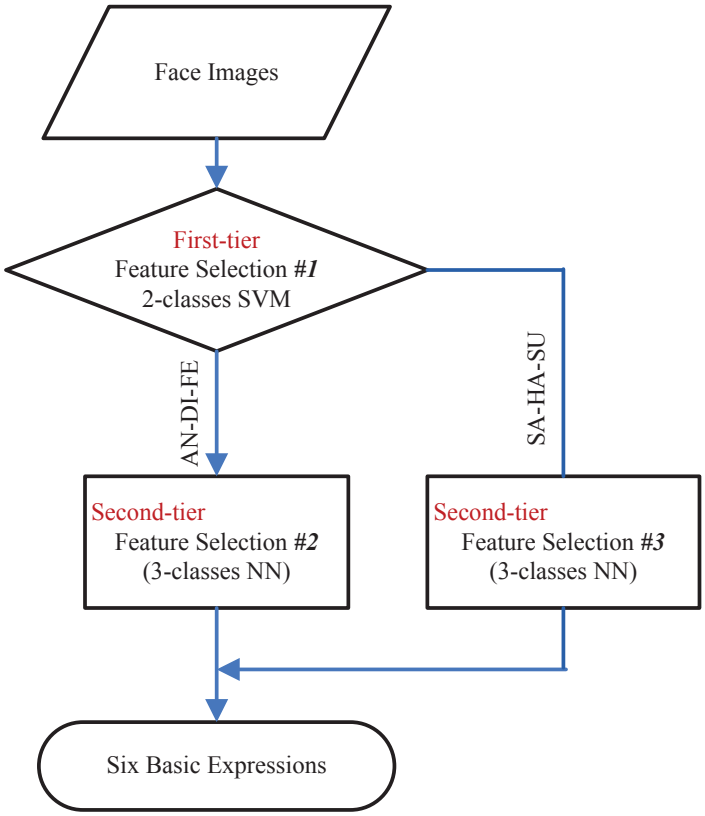


Figure 7.5: The flowchart of the hierarchical classification.

tion than eyebrows (Pardàs and Bonafonte, 2002). Figure 7.4 illustrates the distribution and the gradient directions of the expressive facial parts on 3D face model with different expressions. Take the blue area covering left eyebrow as an example, it means the gradient information along the direction $\mathbf{p}_7$ is very discriminative for expression recognition. The importance of the cheek area is identified by the proposed data-driven algorithm (see Figure 7.4), which has not been mentioned in previous works.

7.2.2 Hierarchical Classification

The six basic expressions are not mutually exclusive. Some of them are easily-confused, even for a human observer, such as anger and sad (Pantic and Rothkrantz, 2000), fear and disgust (Fasel and Luettin, 2003). It is difficult to classify all the 6 classes in one time, because the nonlinear overlaps caused by the subtlety and variance of expressions. Thus, hierarchical classification is adopted to divide and conquer this problem, as shown in Figure 7.5 . In this work, each tier of the classification is wrapper with an independent feature selection to learn the most discriminative features. All the learning processes are

Table 7.1: Confusion matrix of recognition on selected onset expression sequences from the BU-4DFE database.

	AN	DI	FE	SA	HA	SU
AN	**92.46**	2.93	1.11	3.50	0.00	0.00
DI	0.71	**92.13**	0.59	0.00	5.66	0.91
FE	0.91	1.11	**96.98**	0.00	1.00	0.00
SA	1.70	0.00	0.00	**98.30**	0.00	0.00
HA	0.00	0.67	0.50	0.50	**98.33**	0.00
SU	0.00	0.38	0.00	0.00	0.83	**98.78**

totally data-driven with a direct goal of maximize the recognition accuracy, which means no pre-defined rules are used.

7.3 Experimental Results

7.3.1 Experiment Setup

The BU-4DFE database is the first database consists of faces in 3D videos. It involves 101 subjects (58 females and 43 males) of various ethnicities. For each subject the six prototypic expressions (Angry, Disgust, Fear, Happy, Sad and Surprise) (Ekman and Friesen, 1978) were recorded gradually from a neutral face, onset, apex, offset and back to neutral, using the dynamic facial acquisition system Di3D and producing roughly 60,600 3D face modes(frames), with corresponding texture images. Each prototypic 3D expression video sequence lasts about four seconds. The temporal resolution of the 3D videos is 25 fps and each 3D model consists of approximately 35,000 vertices. Each frame is released with 83 manually labeled facial landmark points.

The experimental setups of existing methods on BU-4DFE database vary a lot. Some works (Canavan *et al.*, 2012) select a subset of 60 subject and follow the 54-vs-6 setup, while some other approaches (Sandbach *et al.*, 2012a; Jeni *et al.*, 2012) select certain frames from the original sequence. For example, in Reale *et al.* (2013), only 481 onset frames with the expressions from neutral to apex are selected, while in Fang *et al.* (2012, 2011), the frames with ambiguous expressions are manually removed, and 507 sequences are selected. In order to provide better comparison with the existing works, we conduct

experiments using two different setups, details of which are described in the following 2 subsections. To ensure unbiased experimental results, we conduct 10-fold cross validation in all our experiments and the average recognition rates are reported.

7.3.2 Results for Onset Sequences

In the first experimental setup, to facilitate a fair comparison, we conduct experiments using a similar setup as methods with selected expression sequences. Similar as Jeni *et al.* (2012) and Reale *et al.* (2013), we select 651 onset frames sequences from the BU-4DFE database, and generate training and testing samples from exclusive subjects to guarantee person-independent expression recognition. Since some expression sequences do not start from the neutral frame, the number of sample from each expression is not always the same.

The HOG3D features are extracted from the onset sequences, and two-stage feature selections are conducted on the patch level and face level to find out the expressive area. The features from the expressive parts are then compressed by linear discriminant analysis (LDA) (Belhumeur *et al.*, 1997), and the recognition is accomplished by a Nearest-Neighbor classifier. The confusion matrix of the classification is recorded in Table 7.1, and the proposed method achieves an average recognition rate of 96.64%. Table 7.2 compares the proposed method with other existing methods. It shows that the proposed method outperforms other existing results on selected sequences from BU-4DFE database. In addition, our experimental results are obtained from 651 sequences, which is much more than other methods.

7.3.3 Results for Sequences from 60 Subjects

In the survey by Danelakis *et al.* (2014), all the methods listed in Table 7.2 are labeled with automatic method, since most of them can extract features automatically. However, the methods are tested on a pre-selected expression sequences, or even selected frames, which means that those methods are not truly automatic. In fact, the expression sequences in BU-4DFE database contain plenty of abnormalities, such as the corrupted 3D faces, the missing onset stage of expressions and inconsistencies of expressions etc. This is a significant challenge to the facial expression algorithms. Thus, we set up an experiment under a more realistic scenario to test the performance of the proposed method.

In this setup, there is no manual selection involved, such as selecting better expression

Table 7.2: Comparison of 6 prototypic expressions recognition performance on the BU-4DFE database. '-' means the corresponding data is not available.

	AN	DI	FE	SA	HA	SU	AvgRates
Sun *et al.* (2010)	-	-	-	-	-	-	94.37
Canavan *et al.* (2012)	83.60	83.20	81.30	78.00	92.10	89.50	84.80
Jeni *et al.* (2012)	80.00	59.00	42.00	64.00	78.00	85.00	78.18
Sandbach *et al.* (2012a)	51.92	62.71	46.15	68.97	75.28	82.56	64.60
Fang *et al.* (2011)	68.31	79.69	67.89	71.64	81.31	86.10	75.82
Fang *et al.* (2012)	92.42	91.67	81.06	88.64	98.48	93.94	91.00
Reale *et al.* (2013)	76.30	74.00	60.00	70.90	90.80	89.80	76.90
The proposed	**92.46**	**92.13**	**96.98**	**98.30**	**98.33**	**98.78**	**96.64**

Table 7.3: Confusion matrix of 6 prototypic expressions recognition on the BU-4DFE database.

	AN	DI	FE	SA	HA	SU
AN	**81.11**	5.33	2.89	8.00	1.67	1.00
DI	9.89	**70.56**	8.89	3.22	3.33	4.11
FE	6.56	9.78	**65.56**	2.11	10.44	5.56
SA	14.00	0.44	1.89	**82.33**	0.56	0.78
HA	1.11	3.22	5.22	0.33	**86.22**	3.89
SU	0.56	2.33	4.11	1.22	3.44	**88.33**

Table 7.4: Confusion matrix of hierarchical classification on the BU-4DFE database.

	AN	DI	FE	SA	HA	SU
AN	**82.11**	10.89	6.00	0.56	0.00	0.44
DI	10.67	**71.56**	13.11	1.22	2.11	1.33
FE	9.67	12.00	**69.22**	1.89	3.89	3.33
SA	9.22	1.78	1.56	**87.33**	0.11	0.00
HA	0.78	3.56	3.44	0.00	**92.22**	0.00
SU	0.00	2.11	3.67	0.00	0.00	**94.22**

sequences, consistent frames etc. A subset of 60 (30 females and 30 males) subjects is randomly selected from original BU-4DFE database. In order to guarantee person-independent facial expression recognition, the commonly used 54-vs-6 setup (Sun and Yin, 2008; Wang *et al.*, 2006) is adopted, which means that the 3D expression sequences from 54 subjects are chosen as the training set, and the remaining 6 subjects' sequences are taken as the testing set. First, all the six basic expression are classified by a nearest-neighbor classifier after LDA projection. Table 7.3 shows the confusion matrix of the classification, leading to the average recognition rate of 79.0% and this result is still slightly better than the results which achieved on selected data (Jeni *et al.*, 2012; Sandbach *et al.*, 2012a; Reale *et al.*, 2013). It can be seen from Table 7.3 that the major confusions happen between sadness and anger (14%), fear and happiness (10.44%). Thus, the hierarchical classifier is used to improve the performance further.

We implement the hierarchical classification as follows. First, a SVM is embedded in the 1st-tier to separate six basic expression into two groups: AN-DI-FE and SA-HA-SU. Next, in 2rd-tier, two Nearest-Neighbor classifiers are trained to classified the samples into single expression category. The confusion matrix is recorded in Table 7.4. The average recognition rate of 82.80% is achieved, which is comparable or even better than the results obtained on selected samples such as Jeni *et al.* (2012), Sandbach *et al.* (2012a), and Reale *et al.* (2013).

7.3.4 Discussion

From the comparison in Table 7.2, it can be seen that the proposed method achieves the best performance of 96.64% on selected onset sequences. Even when tested on a much more difficult sample set, i.e. the subset of randomly selected 60 subjects, an average recognition of 82.80% is achieved by hierarchical classification. This is still comparable to

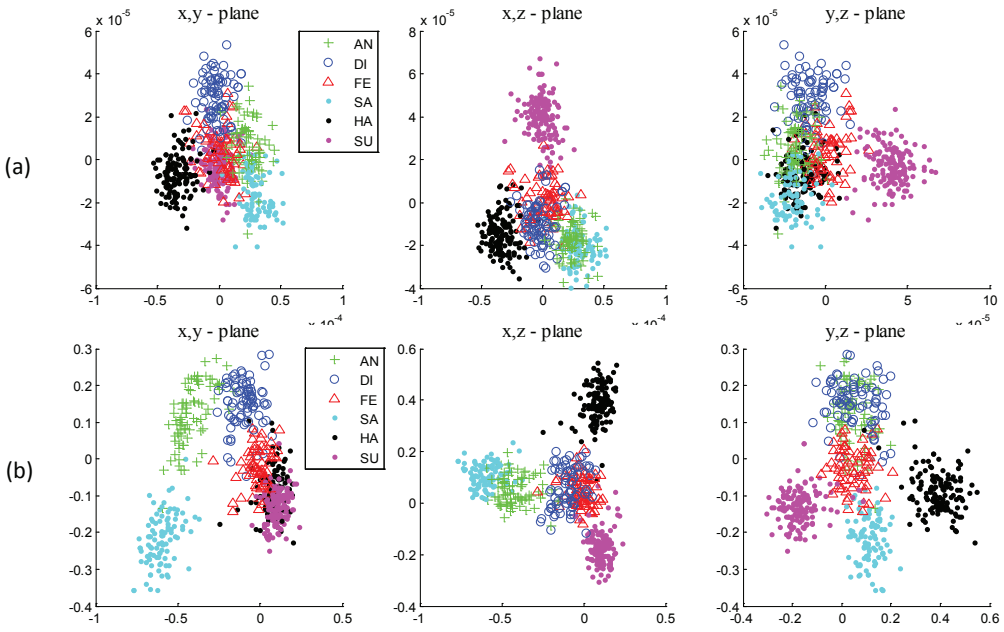


Figure 7.6: The expression samples projected into the subspace from LDA. (a) The top row illustrates the projection of training samples without feature selection. (b) The bottom row demonstrates the projection of training samples after feature selection.

or even better than the results listed in Table 7.2. The reason is as follows:

First, the proposed method extracts the spatio-temporal HOG3D features to represent 4D facial expressions. It is a real 4D feature that can encode the intra-frame deformations and inter-frame dynamics at the same time. This is better than extracting expression information frame-by-frame, in which the resulted features can not sufficiently describe the 3D mesh changes over time.

Second, the most discriminative face parts/components are identified by the proposed two-stage feature selection, and only the features of the most expressive areas are used to distinguish expressions. These features could achieve the best classification results since they jointly have the most characterizing power for the six basic expressions. As shown in Figure 7.6, the first 3 dimensions of all the samples in the LDA subspace are projected to $xy-$,$xz-$,$yz-$planes. It is obvious that the features from the expressive areas (in the bottom row of figure 7.6) are better for classification than the original features, because the six expression classes are well separated in the second row (represented by features from expressive parts), but inseparable in the first row, especially the expressions fear, anger and disgust.

Finally, the 2-tired hierarchical classifier could eliminate confusions in expression recognition. It also facilitates feature selection in each tier, which can use 'learned features' to improve the recognition accuracy. It can be seen by comparing the confusion matrices in Table 7.3 and Table 7.4. In hierarchical classification, the recognition rates of all the expressions are improved, especially the recognition of fear (by 3.66%), sad (by 5.00%), happiness (by 6.00%) and surprise (by 5.89%). In Table 7.3, the major confusion happens in anger and sad (14.00% of sadness are misclassified as anger), and fear and happiness (10.44% of fear are misclassified as happiness). In contrast, hierarchical classification reduces the misclassified rates to 9.22% (anger and sadness) 3.89% (fear and happiness) respectively.

7.4 Chapter Summary

This chapter proposed an automatic 4D facial expression recognition method based on dynamic 3D facial expression sequences. The most characterizing face parts/components for expression representation are identified by two-stage feature selection. The HOG3D features extracted from the most discriminative parts are fed to a 2-tiered hierarchical classifier to accomplish facial expression recognition. The experimental results show that the proposed method outperforms the existing methods by a significant margin. The landmark detection in this work still uses the information of 2D texture images, and a 3D landmark detector that can work directly on 3D information would be investigated in future.

Chapter 8

Conclusion

This thesis has proposed methods to improve the performance of automatic person-independent facial expression recognition. This problem has been explored from 2D image/video to 3D image/video. The facial landmark detection method is extended and applied in order to achieve fully automatic facial expression recognition. This means there is no manual intervention in the expression feature extraction and selection process. The common problems of facial expression recognition, like landmark detection on 3D faces, easily-confused facial expressions classification, spatio-temporal representation for dynamic analysis, has been addressed. The weaknesses of the existing approaches have been analyzed and alternative approaches have been proposed accordingly.

In Chapter 2, a review of related work in the fields of automatic facial expression recognition is presented. The framework of a facial expression recognition system is first briefly described. This is follow by the introduction of the property of 2D and 3D facial expression images and popular benchmark databases. Secondly, a review of the 2D and 3D facial expression feature extraction methods is presented. Finally, previous work on feature selection are then discussed, focusing on dimension reduction and improving the classification performance.

In Chapter 3, the colour spaces for facial expression recognition are investigated in detail. The Uncorrelated Colour Space and Discriminant Colour Space are derived with the purpose of expression recognition. Then, 3 experiments are conducted on on Oulu-CASIA NIR&VIS facial expression database and CurtinFaces database. The performances of the UCS, DCS are compared with RGB colour space and gray images. In addition, the experiment conducted by crossing images source: traning on the Kinect expression captures and testing on the images from Panasonic camera, and vice versa. The Kinect captures are in low resolution but images from Panasonic camera are in high resolution. This is a special case of person-independent facial expression recognition, and a trial on how to recognize the low-resolution expressions based on high-resolution images.

In Chapter 4, a 2-tierd hierarchical classifier focusing on the recognition of easily-confused expressions is implemented. The LBP features are extracted from local regions of 2D

images, and concatenated to form feature vector to represent the expression conveyed by the corresponding image. In order to improve the recognition performance of the easily-confused expressions, a hierarchical classification is applied. Two different SVMs are trained in each tier of the hierarchical classifier. In first tier, the easily-confused expressions such as anger and sadness are merged into on class, and a 5-class SVMs are trained for classification (the merged class and the rest 4 prototypic expressions: Disgust, fear, happiness, surprise.). In second tier, the samples that classified as merged class are then separated by a 2-class SVMs into anger and sadness. It is worth noting that different features are used in each tier of the classification. Comparison with the existing methods, the proposed method can eliminate the confusion between anger and sad significantly.

In Chapter 5, a fully automatic 3D facial expression recognition method is proposed. This work is designed for recognizing facial expressions based on static 3D face model. In order to achieve a fully automatic recognition, 5 points (four eye corners and nose tip) are first detected to serve as the fiducial points for face alignment and feature extraction. This is done by applying Haar-cascade clssier on the range images rendered from the raw 3D point cloud to detect the fiducial points. Then, the 5 fiducial points are used to define a T-area for alignment and generate 25 heuristic points to all over the whole face. According to the T-area, the 3D faces are aligned by ICP algorithm and local depth features are uniformly sampled around the heuristic points to represent facial expressions. After feature selection, the selected features are fed to SVMs classifier to accomplish expression recognition. The performance achieved by the proposed method is the best among existing automatic methods, and also comparable to those approaches which require human interfere.

In Chapter 6, the problem of 4D facial expression recognition is addressed. Unlike the majority of the existing methods, this chapter tries to extract real 4D feature to represent spatio-temporal 3D facial expression dynamics. A tree-structure model is used to detect 130 landmarks for the faces in BU-4DFE database, especially for those faces with extreme expressions. The 3D-DCT features are then extracted around the selected 68 detected landmarks to represent 3D facial expressions dynamics. This is followed by a two-round mRMR (*minimal-redundancy-maximal-relevance*) feature selection to reduce the feature dimension and improve the recognition performance. The proposed method is tested by conducting 6-class recognition and 3-class recognition. In both cases, the proposed method outperform other existing methods on same database.

In Chapter 7, a method to identify the most discriminative facial parts/components for 3D dynamic expressions is presented. Human facial expression are conveyed by different facial regions, and previous work on this issue are all occlusion based. The facial regions, such

as eyebrow and mouth, are blocked by a predefined binary mask, which can not adjust to different data. In this chapter, the identification is conduct on 4D expression data, with the HOG3D features extract from local depth patch-sequences. A hierarchical classification embedded with 2-stage feature selection is utilized to pick the most discriminative facial parts out with the direct goal of maximizing recognition rates. The data-driven selection result shows mouth, cheeks, and eyebrow carry most of expression related information. Different with the 2D expression data, the result shows that cheek area carries important expression information in 3D facial data.

8.1 Summary of Contributions

The contributions of this thesis include the following:

- A quantitative performance comparison of four colour spaces, such as UCS, DCS, RGB and Gray image, for facial expression recognition. This thesis tries to figure what is the best colour space for facial expression recognition.
  - ≻ Unlike in face recognition, DCS can not achieve consistent better performance than the rest colour space for facial expression recognition.
  - ≻ The crossing image source experiment is a specific trial for expression recognition, especially when training on high-resolution images and testing on low-resolution images.
- The use of hierarchial classification to improve the recognition performance of easily-confused expressions. The significance of this includes:
  - ≻ For those easily-confused expressions, hierarchical classification can divide an conquer the problem by merging the easily-confused expressions in the first tier, and then separating them in the second tier.
  - ≻ The 2-tier structure allows the classifier to use different features in each tiers. The features could be either extracted by different descriptors from original data, or selected with different task in each tiers.
- Proposing a 2-stage feature selection to identify the most discriminant facial region for expression representation and recognition. This has several advantages:
  - ≻ The 2-stage feature selection is applied bottom up. In first stage, local features are selected based on current interest region, such as the depth patches in chapter

6. The result features of first stage are given to the second stage to identify *"the best m features"* for the whole face.

- ≻ The feature selection is wrapped with classification. With different recognition task in each tier, the selection process could select the best features for current tier.
- ≻ The proposed 2-stage feature selection are used to identify the most discriminative facial regions for expression recognition.

8.2 Future Work

In general, there are certain limits to the circumstances that the proposed approaches can be applied to. As mentioned in chapter 1, the current state-of-art facial expression recognition system focused on achieving different aspects of the ideal properties, such as fully automatic, person-independent, real-time etc. However, it is need to integrate all of these ideas together to refine facial expression recognition system. For the proposed method, the efficiency and recognition performance would ideally be improved.

8.2.1 Real-time Recognition

Although increasing number of work which achieved automatic facial expression recognition, few of them could be real-time. The computational cost lies either in fiducial points detection or the consequent feature extraction. The fiducial points are very important for expression feature extraction, which is almost inevitable to design a high-performance system. Lots of algorithm, such as ASM, AAM, Haar-cascade, have been proposed for landmark detection on 2D human face images. However, the computing cost are still quite high, and the accuracy are expected to be improved. Moreover, 3D face images have gained its popularity in facial expression analysis, but landmark detection on 3D face model is still an open problem, especially real-time detection and tracking on 3D data. Until now, very few work focuses on solving this problem. The ultimate goal of 3D facial expression recognition systems will be real-time analysis, requiring real-time alignment and tracking. Approaches with low computational cost for this two operations need be investigated. For expression feature extraction, effective descriptors is required, especially for 3D dynamic expression analysis. If the extract features are in high-dimensional space, the effective dimension reduction methods need to be applied to balance the computational cost and feature discrimination.

8.2.2 Easily-confused Expressions

It has been noticed for a long time that the six basic expressions, namely anger, disgust, fear, sadness, happiness, and surprise, are not mutually exclusive. From the results of survey papers by Bettadapura (2012), similar confusion happened between anger and sadness (Aleksic and Katsaggelos, 2006; Sebe *et al.*, 2007), anger with fear (Kotsia and Pitas, 2007; Kotsia *et al.*, 2008),. In contrast, surprise and happiness are easiest to recognize (Michel and El Kaliouby, 2003; Pardàs and Bonafonte, 2002). The confusions caused by the easily-confused expressions will affect the overall performance significantly. In order to separate the easily-confused anger and sadness, chapter 3 uses specific features and hierarchical classification to reduce the confusions. However, except happiness and surprise, the recognition rate of the rest 4 expressions are still not good enough. A possible way to solve this problem is using expression-specific classifiers. For each expression, a unique feature could be extracted to train a expression-specific classifier, and the the final decision could be achieved by weighted voting.

8.2.3 Spontaneous Expression Recognition

Most of existing research works on the basis of recognizing deliberately displayed/posed expressions, such as the six prototypic expressions. Recently, research effort has begun to shift to recognizing more complex and spontaneous expressions, such as lack of attention, boredom, frustration, pain etc. The major challenge that the researches face is the non-availability of spontaneous expression data. If the subjects become aware of the recording, their expressions lose authenticity immediately (Sebe *et al.*, 2007). The BP4D-Spontaneous Zhang *et al.* (2014) data is the pioneer work which attempt to capture spontaneous facial expression data for research, in which the elicited expressions of subjects are recorded. For recording spontaneous affective behavior, a trade-off between the acquisition of natural emotional expressions and data quality is applied. If the recording environment is too constrained, genuine emotion and social signaling become difficult to elicit. However, If the recording environment is unconstrained, substantial error may be introduced into the recordings. Though well-validated emotion techniques which meet the challenge mentioned by Coan and Allen (2007) are used, it is hard to grantee the expression authenticity. Furthermore, real life facial expression analysis is much more difficult. The factors, such as head motion, low resolution, low expression intensity, will complicate facial expression analysis.

8.2.4 Temporal Information

The majority of the existing work on dynamic facial expression analysis address the expression recognition as a time-series problem. These works try to extract features from discrete expression frames, and sequential models like HMMs are then used trained to finish expression recognition. However, temporal information has been shown to be able to improve the performance of recognition Sun *et al.* (2008), the timing of facial actions may be as important as their configurations. In fact, the differences of spontaneous and deliberate facial expressions may be reflected by the temporal parameters (Cohn *et al.*, 2002), such as the intensity and duration of expressions. This is consistent with neuropsychological models (Rinn, 1984). Thus, spatio-temporal descriptors, which can encode the deformation of facial features, the relative timing of facial actions as well as their temporal evolution, is need to be developed for dynamic expression analysis.

Bibliography

Aleksic, P. and Katsaggelos, A. (2006). Automatic facial expression recognition using facial animation parameters and multistream hmms. *Information Forensics and Security, IEEE Transactions on*, **1**(1), 3–11. 34, 97

Bartlett, M., Littlewort, G., Fasel, I., and Movellan, J. (2003). Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, pages 53–53. IEEE. 34, 40

Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573. Ieee. 15

Bartlett, M. S., Braathen, B., Littlewort-Ford, G., Hershey, J., Fasel, I., Marks, T., Smith, E., Sejnowski, T. J., and Movellan, J. R. (2001). Automatic analysis of spontaneous facial behavior: A final project report. *Institute for Neural Computation MPLab TR2001*, **8**. 14

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **19**(7), 711–720. 65, 73, 88

Berretti, S., Del Bimbo, A., and Pala, P. (2012). Real-time expression recognition from dynamic sequences of 3d facial scans. In *Proceedings of the 5th Eurographics conference on 3D Object Retrieval*, pages 85–92. Eurographics Association. 19

Besl, P. J. and McKay, N. D. (1992). Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics. 52

Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*. 2, 3, 4, 97

Bourel, F., Chibelushi, C. C., and Low, A. A. (2001). Recognition of facial expressions in the presence of occlusion. In *BMVC*, pages 1–10. Citeseer. 6, 7, 35, 39, 42, 79, 85

Bourgain, J. (1985). On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, **52**(1-2), 46–52. 18

Canavan, S., Sun, Y., Zhang, X., and Yin, L. (2012). A dynamic curvature based approach for facial activity analysis in 3d space. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 14–19. IEEE. 87, 89

Chan, Y.-L. and Siu, W.-C. (1997). Variable temporal-length 3-d discrete cosine transform coding. *Image Processing, IEEE Transactions on*, **6**(5), 758–763. 69

Chang, C.-C. and Lin, C.-J. (2011a). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 40

Chang, C.-C. and Lin, C.-J. (2011b). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3), 27. 59

Chang, Y., Vieira, M., Turk, M., and Velho, L. (2005). Automatic 3d facial expression analysis in videos. In *Analysis and Modelling of Faces and Gestures*, pages 293–307. Springer. 18

Chen, D., Cao, X., Wen, F., and Sun, J. (2013). Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE. 72

Cheon, Y. and Kim, D. (2009). Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, **42**(7), 1340–1350. 13

Coan, J. and Allen, J. (2007). Oxford handbook on emotion elicitation and assessment. 97

Cohen, I., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, **91**(1), 160–187. 34

Cohn, J. F., Schmidt, K., Gross, R., and Ekman, P. (2002). Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 491. IEEE Computer Society. 98

Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **23**(6), 681–685. 13, 19

Dahmane, M. and Meunier, J. (2011). Emotion recognition using dynamic grid-based hog features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 884–888. IEEE. 15

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE. 14, 15

Danelakis, A., Theoharis, T., and Pratikakis, I. (2014). A survey on facial expression recognition in 3d video sequences. *Multimedia Tools and Applications*, pages 1–39. 2, 3, 88

Darwin, C. (1872). The expression of the emotions in man and animals. *London, UK: John Marry*. 1

Daugman, J. G. (1988). Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **36**(7), 1169–1179. 14

D'Errico, J. R. (2006). Understanding gridfit. *Information available at: http://www. mathworks. com/matlabcentral/fileexchange/loadFile. do*. 54, 69, 82

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, **3**(02), 185–205. 72, 84

Donato, G., Bartlett, M., Hager, J., Ekman, P., and Sejnowski, T. (1999). Classifying facial actions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **21**(10), 974–989. 14

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, **17**(2), 124. 1, 34

Ekman, P. and Friesen, W. V. (1978). *Manual for the facial action coding system*. Consulting Psychologists Press. 9, 73, 87

Essa, I. A. and Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **19**(7), 757–763. 14

Fang, T., Zhao, X., Shah, S. K., and Kakadiaris, I. A. (2011). 4d facial expression recognition. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1594–1601. IEEE. 64, 75, 76, 77, 87, 89

Fang, T., Zhao, X., Ocegueda, O., Shah, S. K., and Kakadiaris, I. A. (2012). 3d/4d facial expression analysis: an advanced annotated face model approach. *Image and Vision Computing*, **30**(10), 738–749. 87, 89

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, **36**(1), 259–275. 2, 3, 4, 22, 34, 64, 86

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7**(2), 179–188. 26

Gu, H. and Ji, Q. (2004). Facial event classification with task oriented dynamic bayesian network. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–870. IEEE. 34

Jeni, L. A., Lőrincz, A., Nagy, T., Palotai, Z., Sebők, J., Szabó, Z., and Takács, D. (2012). 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, **30**(10), 785–795. 19, 87, 88, 89, 90

Jones, C. and Abbott, A. L. (2006). Color face recognition by hypercomplex gabor analysis. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 6–pp. IEEE. 22

Klaser, A. and Marszalek, M. (2008). A spatio-temporal descriptor based on 3d-gradients. 82

这篇论文讲3D梯度直方图特征

Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Image Processing, IEEE Transactions on*, **16**(1), 172–187. 34, 97

Kotsia, I., Buciu, I., and Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, **26**(7), 1052–1067. 6, 7, 34, 35, 39, 53, 79, 85, 97

Lajevardi, S. M. and Wu, H. R. (2012). Facial expression recognition in perceptual color space. *Image Processing, IEEE Transactions on*, **21**(8), 3721–3733. 4, 22

Lanitis, A., Taylor, C. J., and Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **19**(7), 743–756. 19

Le, V., Tang, H., and Huang, T. S. (2011). Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 414–421. IEEE. 64, 76, 77

Lemaire, P., Ben Amor, B., Ardabilian, M., Chen, L., and Daoudi, M. (2011). Fully automatic 3d facial expression recognition using a region-based approach. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 53–58. ACM. 17, 61, 62

Lemaire, P., Chen, L., Ardabilian, M., Daoudi, M., *et al.* (2013). Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients. In *Workshop 3D Face Biometrics,*. 48, 61, 62

Littlewort, G., Fasel, I., Bartlett, M. S., and Movellan, J. R. (2002). Fully automatic coding of basic expressions from video. *University of California, San Diego, San Diego, CA*, **92093**. 14

Liu, C. (2008). Learning the uncorrelated, independent, and discriminating color spaces for face recognition. *Information Forensics and Security, IEEE Transactions on*, **3**(2), 213–222. 22

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE. 11, 15, 35, 37, 39, 42, 46

Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE. 14

Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(12), 1357–1362. 14

Maalej, A., Amor, B. B., Daoudi, M., Srivastava, A., and Berretti, S. (2010). Local 3d shape analysis for facial expression recognition. In *20th International Conference on Pattern Recognition (ICPR 2010)*, pages 4129–4132. 17

Maalej, A., Amor, B. B., Daoudi, M., Srivastava, A., and Berretti, S. (2011). Shape analysis of local facial patches for 3d facial expression recognition. *Pattern Recognition*, **44**(8), 1581–1589. 17, 48, 49, 54, 60, 61, 62

Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, **60**(2), 135–164. 13, 37

Mian, A. (2011). Robust realtime feature detection in raw 3d face images. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 220–226. IEEE. 50, 51

Michel, P. and El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM. 34, 35, 97

Mpiperis, I., Malassiotis, S., Petridis, V., and Strintzis, M. G. (2008a). 3d facial expression recognition using swarm intelligence. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2133–2136. IEEE. 18

Mpiperis, I., Malassiotis, S., and Strintzis, M. G. (2008b). Bilinear elastically deformable models with application to 3d face and facial expression recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE. 18

Mpiperis, I., Malassiotis, S., and Strintzis, M. G. (2009). Bilinear decomposition of 3-d face images: An application to facial expression recognition. In *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS'09. 10th Workshop on*, pages 1–4. IEEE. 18

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, **29**(1), 51–59. 14, 36

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **24**(7), 971–987. 14, 36

Orrite, C., Gañán, A., and Rogez, G. (2009). Hog-based decision tree for facial expression classification. In *Pattern Recognition and Image Analysis*, pages 176–183. Springer. 15

Pantic, M. and Rothkrantz, L. (2004). Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **34**(3), 1449–1461. 37

Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(12), 1424–1445. 1, 3, 4, 22, 34, 48, 64, 86

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE. 11

Pardàs, M. and Bonafonte, A. (2002). Facial animation parameters extraction and expression recognition using hidden markov models. *Signal Processing: Image Communication*, **17**(9), 675–688. 6, 7, 35, 39, 42, 79, 85, 86, 97

Patras, I. and Pantic, M. (2004). Particle filtering with factorized likelihoods for tracking facial features. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 97–102. IEEE. 13

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **27**(8), 1226–1238. 49, 57, 58, 72, 84

Person, K. (1901). On lines and planes of closest fit to systems of points in space. *philosophical magazine*, **2**(6), 559–572. 24

Rajapakse, M., Tan, J., and Rajapakse, J. (2004). Color channel encoding with nmf for face recognition. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 3, pages 2007–2010. IEEE. 22

Ramanathan, S., Kassim, A., Venkatesh, Y., and Wah, W. S. (2006). Human facial expression recognition using a 3d morphable model. In *Image Processing, 2006 IEEE International Conference on*, pages 661–664. IEEE. 17

Reale, M., Zhang, X., and Yin, L. (2013). Nebula feature: a space-time feature for posed and spontaneous 4d facial behavior analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE. 19, 21, 64, 75, 76, 87, 88, 89, 90

Rinn, W. E. (1984). The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, **95**(1), 52. 98

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **20**(1), 23–38. 13

Rudovic, O., Pavlovic, V., and Pantic, M. (2012). Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2634–2641. IEEE. 46

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast mr images. *Medical Imaging, IEEE Transactions on*, **18**(8), 712–721. 19

Samal, A. and Iyengar, P. A. (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition*, **25**(1), 65–77. 1

Sandbach, G., Zafeiriou, S., Pantic, M., and Rueckert, D. (2011). A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 406–413. IEEE. 19, 64, 76, 77

Sandbach, G., Zafeiriou, S., Pantic, M., and Rueckert, D. (2012a). Recognition of 3d facial expression dynamics. *Image and Vision Computing*, **30**(10), 762–773. 64, 75, 87, 89, 90

Sandbach, G., Zafeiriou, S., Pantic, M., and Yin, L. (2012b). Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, **30**(10), 683–697. 2, 3, 22, 48

Sebe, N., Lew, M., Sun, Y., Cohen, I., Gevers, T., and Huang, T. (2007). Authentic facial expression analysis. *Image and Vision Computing*, **25**(12), 1856–1863. 34, 97

Servais, M. and De Jager, G. (1997). Video compression using the three dimensional discrete cosine transform (3d-dct). In *Communications and Signal Processing, 1997. COMSIG'97., Proceedings of the 1997 South African Symposium on*, pages 27–32. IEEE. 69

Sha, T., Song, M., Bu, J., Chen, C., and Tao, D. (2011). Feature level analysis for 3d facial expression recognition. *Neurocomputing*, **74**(12), 2135–2141. 16, 60, 62

Shan, C., Gong, S., and McOwan, P. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, **27**(6), 803–816. 14, 34, 36, 40, 58

Soyel, H. and Demirel, H. (2007). Facial expression recognition using 3d facial feature distances. In *Image Analysis and Recognition*, pages 831–838. Springer. 16

Srivastava, R. and Roy, S. (2009). 3d facial expression recognition using residues. In *TENCON 2009-2009 IEEE Region 10 Conference*, pages 1–5. IEEE. 16

Sun, Y. and Yin, L. (2008). Facial expression recognition based on 3d dynamic range model sequences. In *Computer Vision–ECCV 2008*, pages 58–71. Springer. 19, 67, 75, 76, 82, 90

Sun, Y., Reale, M., and Yin, L. (2008). Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE. 19, 64, 67, 76, 82, 98

Sun, Y., Chen, X., Rosato, M., and Yin, L. (2010). Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, **40**(3), 461–474. 89

Suwa, M., Sugie, N., and Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. In *International joint conference on pattern recognition*, pages 408–410. 1

Tang, H. and Huang, T. S. (2008). 3d facial expression recognition based on properties of line segments connecting facial feature points. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE. 16, 60, 62

Tao, H. and Huang, T. S. (1999). Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1. IEEE. 14

Tekguc, U., Soyel, H., and Demirel, H. (2009). Feature selection for person-independent 3d facial expression recognition using nsga-ii. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pages 35–38. IEEE. 60, 62

Tian, Y., Kanade, T., and Cohn, J. (2001). Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **23**(2), 97–115. 12, 37

Tian, Y., Kanade, T., and Cohn, J. (2002). Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 229–234. IEEE. 14, 15, 34

Tsalakanidou, F. and Malassiotis, S. (2009). Robust facial action recognition from real-time 3d streams. In *Computer Vision and Pattern Recognition Workshops*, pages 4–11. 19

Valstar, M., Patras, I., and Pantic, M. (2005). Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 76–76. IEEE. 13, 15, 34, 35, 40

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, **57**(2), 137–154. 50

Wang, J., Yin, L., Wei, X., and Sun, Y. (2006). 3d facial expression recognition based on primitive surface feature distribution. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1399–1406. IEEE. 17, 48, 60, 62, 75, 90

Xiao, J., Baker, S., Matthews, I., and Kanade, T. (2004). Real-time combined 2d+ 3d active appearance models. In *CVPR (2)*, pages 535–542. 13

Xue, M., Mian, A., Liu, W., and Li, L. (2014). Fully automatic 3d facial expression recognition using local depth features. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1096–1103. IEEE. 64, 75, 76

Yang, J. and Liu, C. (2008). A discriminant color space method for face representation and verification on a large-scale database. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE. 22, 26

Yang, J., Zhang, D., Frangi, A. F., and Yang, J.-y. (2004). Two-dimensional pca: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **26**(1), 131–137. 25, 56

Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006a). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE. 11, 48, 49, 61, 62

Yin, L., Wei, X., Longo, P., and Bhuvanesh, A. (2006b). Analyzing facial expressions using intensity-variant 3d data for human computer interaction. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1248–1251. IEEE. 19

Yin, L., Chen, X., Sun, Y., Worm, T., and Reale, M. (2008). A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE. 11, 64, 80

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, **32**(10), 692–706. 97

Zhang, Z., Lyons, M., Schuster, M., and Akamatsu, S. (1998). Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 454–459. IEEE. 14, 15, 34

Zhao, G., Huang, X., Taini, M., Li, S. Z., and PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, **29**(9), 607–619. 11, 28

Zhao, X., Huang, D., Dellandréa, E., and Chen, L. (2010). Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3724–3727. IEEE. 17

Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., and Metaxas, D. (2012). Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2562–2569. IEEE. 46

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE. 67, 80