Inferring Facial and Body Language

Shan, Caifeng

For additional information about this publication click this link.
http://www.eecs.qmul.ac.uk/tech_reports/RR-08-01.pdf

Inferring Facial and Body Language

Caifeng Shan

RR-08-01

February 2008

Queen Mary
University of London
Department of Computer Science

INFERRING FACIAL AND BODY LANGUAGE

Caifeng Shan

Submitted to the University of London in partial fulfillment of the requirements for the degree of Doctor of Philosophy

QUEEN MARY, UNIVERSITY OF LONDON

2007

Declarations

I, Caifeng Shan, declare that this thesis titled "Inferring Facial and Body Language" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institutions, this has been clearly stated

- Where I have consulted the published work of others, this is always clearly attributed

- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work

- I have acknowledged all main sources of help

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Signed:
_____

Date:
_____

Abstract

Machine analysis of human facial and body language is a challenging topic in computer vision, impacting on important applications such as human-computer interaction and visual surveillance. In this thesis, we present research building towards computational frameworks capable of automatically understanding facial expression and behavioural body language.

The thesis work commences with a thorough examination in issues surrounding facial representation based on Local Binary Patterns (LBP). Extensive experiments with different machine learning techniques demonstrate that LBP features are efficient and effective for person-independent facial expression recognition, even in low-resolution settings. We then present and evaluate a conditional mutual information based algorithm to efficiently learn the most discriminative LBP features, and show the best recognition performance is obtained by using SVM classifiers with the selected LBP features. However, the recognition is performed on static images without exploiting temporal behaviors of facial expression.

Subsequently we present a method to capture and represent temporal dynamics of facial expression by discovering the underlying low-dimensional manifold. Locality Preserving Projections (LPP) is exploited to learn the expression manifold in the LBP based appearance feature space. By deriving a universal discriminant expression subspace using a supervised LPP, we can effectively align manifolds of different subjects on a generalised expression manifold. Different linear subspace methods are comprehensively evaluated in expression subspace learning. We formulate and evaluate a Bayesian framework for dynamic facial expression recognition employing the derived manifold representation. However, the manifold representation only addresses temporal correlations of the whole face image, does not consider spatial-temporal correlations among different facial regions.

We then employ Canonical Correlation Analysis (CCA) to capture correlations among face parts. To overcome the inherent limitations of classical CCA for image data, we introduce and formalise a novel Matrix-based CCA (MCCA), which can better measure correlations in 2D image data. We show this technique can provide superior performance in regression and recognition tasks, whilst requiring significantly fewer canonical factors. All the above work focuses on facial expressions. However, the face is usually perceived not as an isolated object but as an integrated part of the whole body, and the visual channel combining facial and bodily expressions is most informative.

Finally we investigate two understudied problems in body language analysis, gait-based gender discrimination and affective body gesture recognition. To effectively combine face and body cues, CCA is adopted to establish the relationship between the two modalities, and derive a semantic joint feature space for the feature-level fusion. Experiments on large data sets demonstrate that our multimodal systems achieve the superior performance in gender discrimination and affective state analysis.

Acknowledgments

I would like to express my deep gratitude to my two supervisors, Prof. Shaogang Gong and Prof. Peter W. McOwan, for their enthusiastic supervision and continued support during the last three years. It was through Sean and Peter that I learned independent exploration and thought. Sean deserves special thanks for keeping sharing his knowledge and advice, and I learned a lot from the productive and inspirational discussions with him.

I also thank other members of the Vision Group who supply friendly working environment and simulating discussions: Pengwei Hao, Fabrizio Smeraldi, Lourdes de Agapito Vicente, Tao Xiang, Andrew Graves, Xavier Llado, Alessio Del Bue, Jianguo Zhang, John Qiu, Yogesh Raja, Lukasz Zalewski, Hayley Hung, Alex Leung, Chris Jia, David Russell, Yong Wang, Jun Li, Milan Verma, Bushra Akhtar, and Samuel Pachoud.

My thanks also go to technical staff, Tim Kay, Matt Bernstein, Derek Coppen, David Hawes, Keith Clarke, and Tom King, for solving the innumerable hardware/software problems that cropped up during the course of the work, and the administrative staff, Julie Ringham, Gill Carter, Joan Hunter, Rupal Vaja, Sue White, Colin Powell, and Carly Wheeler, for their friendly and generous help during the three years.

I would like to acknowledge the financial support I have had over the years: full research studentship of Queen Mary, the International Travel Grant of the Royal Academy of Engineering, and the Royal Society International Joint Project.

Most of all, I am indebted to my family without whose endless sacrifice and support there would never be possible for this thesis to happen. Specifically I am grateful to my wife Na Li, and I cannot thank her enough for her love and encouragement.

Publication List

1. C. Shan, S. Gong, and P. W. McOwan, "Matrix-based Canonical Correlation Analysis for Facial Expression Analysis", Submitted to *IEEE Transactions on Circuits and Systems for Video Technology*.

2. C. Shan, S. Gong, and P. W. McOwan, "A Generalized Facial Expression Manifold for Dynamic Expression Analysis", Submitted to *IEEE Transactions on Systems, Man and Cybernetics, Part B*.

3. C. Shan, S. Gong, and P. W. McOwan, "Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study", *Image and Vision Computing* (Minor Revision).

4. C. Shan, S. Gong, and P. W. McOwan, "Fusing Gait and Face Cues for Human Gender Recognition", Accepted by *Neurocomputing*.

5. C. Shan, S. Gong, and P. W. McOwan, "Beyond Facial Expressions: Learning Human Emotion from Body Gestures", In *British Machine Vision Conference (BMVC'07)*, pages 242-251, Warwick, UK, September, 2007.

6. C. Shan, S. Gong, and P. W. McOwan, "Capturing Correlations Among Facial Parts for Facial Expression Analysis", In *British Machine Vision Conference (BMVC'07)*, pages 469-478, Warwick, UK, September, 2007.

7. C. Shan, S. Gong, and P. W. McOwan, "Learning Gender from Human Gaits and Faces", In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS'07)*, London, UK, September, 2007.

8. C. Shan, S. Gong, and P. W. McOwan, "Dynamic Facial Expression Recognition Using A Bayesian Temporal Manifold Model", In *British Machine Vision Conference (BMVC'06)*, pages 297-306, Edinburgh, UK, September, 2006.

9. C. Shan, S. Gong, and P. W. McOwan, "A Comprehensive Empirical Study on Linear Subspace Methods for Facial Expression Analysis", In *IEEE Workshop on Vision for Human-Computer Interaction, in conjunction with CVPR'06*, pages 153-158, New York, USA, June, 2006.

10. C. Shan, S. Gong, and P. W. McOwan, "Appearance Manifold of Facial Expression", In *IEEE Workshop on Human-Computer Interaction, in conjunction with ICCV'05*, pages 221-230, Beijing, China, October, 2005.

11. C. Shan, S. Gong, and P. W. McOwan, "Conditional Mutual Information Based Boosting for Facial Expression Recognition", In *British Machine Vision Conference (BMVC'05)*, pages 399-408, Oxford, UK, September, 2005.

12. C. Shan, S. Gong, and P. W. McOwan, "Recognizing Facial Expressions at Low Resolution", In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS'05)*, pages 330-335, Como, Italy, September, 2005.

13. C. Shan, S. Gong, and P. W. McOwan, "Robust Facial Expression Recognition Using Local Binary Patterns", In *IEEE International Conference on Image Processing (ICIP'05)*, pages 370-373, Genoa, Italy, September, 2005.

Contents

List of Figures

List of Tables

1 Introduction

Human communication consists of two main aspects: verbal and nonverbal. "Verbal" means "of or concerned with words", and verbal communication, which often refers to spoken language, conveys messages that are made up of words. On the contrary, nonverbal communication is the process of communication through sending and receiving wordless messages, which can be communicated through facial expression, eye contact, gaze, body movement and posture, gesture, tone of voice, speaking style, touch, and so on. Although spoken language is indispensable for sharing ideas and feelings, nonverbal cues play a vital role in social interpersonal communication. For example, we express emotion, mood, attitude, and attention through nonlinguistic messages. Mehrabian [101] claimed that *words* (Verbal), *tone of voice* (Vocal) and *body language* (Visual) are three basic elements in face-to-face communication, and up to 93% of communication (of feelings and attitudes) is nonverbal: body language accounts for 55%, tone of voice accounts for 38%, and words only account for 7%. Here body language is a broad term representing visual nonverbal behaviour including facial expression, body movement and posture, gesture, and so on. These visual cues are the major and fundamental means for nonverbal communication. A psychological study [3] indicates that the visual channel carrying facial expressions and body gestures is most important in human judgement of behavioural cues. Human judges seem to be most accurate in their judgement when they are able to observe the face and the body. In this thesis, in contrast to facial expression, bodily expression (including body movement, posture, gesture, and gait) is considered as body language. Two examples of facial and body language are shown in Figure 1.1.

Facial and body language are the main ways for humans to communicate their emotions

1

Figure 1.1: Facial and body language in G8 Summit (from BBC News) (*Left*) "Blair and Putin look one another straight in the eye. Blair's smile is slightly nervous, Putin relaxed and confident, his left arm grasps Blair's upper arm. Putin looks like the man in charge." (*Right*) "Bush is covering his mouth with his hand, as if he has something to hide. Putin is smirking, cheeks slightly raised as if genuine amusement. Is this joke less than polite? It is interesting that Putin has to lean over far more than Bush to hear his aside."

and intentions, which are the most complex messages communicated by human nonverbal behaviour [117]. There is considerable history associated with the link between emotions and facial and body language. Charles Darwin [38] was the first to describe in detail the specific facial and bodily expressions associated with emotions in animals and humans. The human face is the preeminent means of expressing and interpreting somebody's affective states based on the shown facial expressions. Darwin argued that all mammals show emotions reliably in their faces. Paul Ekman's influential studies [49] on facial expression determined that expressions of anger, disgust, fear, joy, sadness and surprise are universal. The human body configuration and movement also reveal and enhance emotions. For example, an angry face is more menacing when accompanied by a fist. When we see a bodily expression of emotion, we immediately know what specific action is associated with a particular emotion, leaving little need for interpretation of the signal, as is the case for facial expressions [42]. Furthermore, a recent psychological study [100] suggests the recognition of facial expression is strongly influenced by the concurrently presented emotional body language.

Human facial and body language also reveal other information including identity, age, gender, attractiveness, and personality. Human face has been widely used to gather this

information [115]. One can also tell a lot about people from their body language. For example, gait, the style of walking, has been widely studied for human identification [109]. Psychophysical studies [85] have suggested that people may recognize the gender of walkers by their gaits. Careful observation of gait can provide insight into the walker's overall state of health, and gait changes also occur as part of the natural process of human aging [124].

1.1 Automatic Facial and Body Language Understanding

In the information society we entered, computers (and computing devices) have become pervasive in our daily life. Moreover, it is widely believed that computers will be embedded everywhere in human environments in the future, receding into the background of our lives, which is often referred to as "ubiquitous computing" [169]. This vision of future brings with it great challenges for Human-Computer Interaction (HCI) designs. The conventional HCI devices like keyboard, mouse and visual displays, which assume that the human will be unambiguous and fully attentive while controlling information and command flow, cannot provide natural and efficient interactions with the computing devices diffused throughout future smart environments [117]. As Pantic *et al.* stated in [117], "we must approach HCI in a different way, moving away from computer-centered designs toward human-centered designs, made for humans based on models of human behaviour". Human-center computing targets computer systems that can unobtrusively perceive and understand human behaviour in unstructured environments and respond appropriately [29]. Next generation HCI designs will be built on " human behaviour computing", to realize human-like interactive functions such as understanding and emulating human affective and social signaling. Meanwhile, computer vision-based facial and body language understanding is a major and fundamental step to achieve this aim.

In addition to more intelligent HCI, there are numerous potential applications of automatic facial and body language understanding, for example:

- Computer animation: computer-generated personified virtual characters with believable facial expression and body language, which can be used for entertainment, education

and customer service [106].

- Visual surveillance and security: human identification or categorization (based on gender, age, and so on) using behavioural signals (face, gait, etc.); automatic assessment of boredom, inattention, and stress in situations where firm attention is essential, for example, driver monitoring [59].

- Medical diagnosis: diagnosing early psychological disorders, or identifying specific mental processes from facial expression [115]; inferring the walker's state of health from his/her gait.

- Emotion-related research (behavioural science, neurology, psychiatry, etc): improving the processing of emotion data by providing more efficient, reproduceable and accurate measurements of emotional expressions.

- Law enforcement: providing reliable cues in establishing credibility and concealing deceit [9].

- Content-based video/image retrieval: automatically describing and annotating images and videos based facial and body behaviour occurring in them.

- Education: automated tutoring systems that can recognize the emotional and cognitive states of pupils.

Driven by its important applications explained above and the theoretical interests of cognitive, psychological and medical scientists, automatic facial and body language understanding has received much attention recently [117, 115, 73]. However, although human cognitive process appears to detect and interpret facial and body behavioural signals with little or no effort, design and development of an automated system that accomplishes this task is rather difficult.

One main research topic is the modeling and understanding of affective facial and bodily expressions [122]. This is because machine analysis of affective behaviours is a key component to realize human-like HCI designs. The recently initiated research area of *affective*

computing [125] focuses on sensing, detecting and interpreting human affective states and devising appropriate means for handling the affective information to enable computers to express and recognize affect [115]. A great deal of attention has been focused on how emotions are communicated through facial expression [120]. Although much progress has been made, recognizing facial expression with a high accuracy remains difficult due to its subtlety, complexity, and variability. Moreover, little attention has been placed on the modeling of affective body language. Since both face and body contribute towards conveying the emotional state of an individual, how to integrate the two modalities also needs to be investigated.

Another research area is human identification and categorization based on facial and body behaviour. There have been extensive studies on human identification based on face [74] or gait [109]. Facial cues have also been exploited for gender classification [104] and estimation of the age of individuals [128]. Although humans can learn significant information about people by their gaits, such as their gender and approximate age, or whether they are tired, in pain or drunk, few studies have been conducted on developing automated systems for human categorization based on gaits.

1.2 Approach

The goal of our research is to address vision-based automatic facial and body language understanding. In this thesis, we present research building towards computational frameworks capable of automatically understanding facial expression and behavioural body language. In particular, we study the following problems.

1.2.1 Feature Selection and Representation

A vital step for successful vision-based facial and body language understanding is deriving an effective feature representation from original images, which includes identifying an appropriate feature representation scheme and selecting the most discriminative features. The thesis work commences with a thorough examination in issues surrounding facial representation based on statistical local features. The method of Local Binary Patterns (LBP) [112] is em-

pirically investigated for person-independent facial expression recognition. Different machine learning methods, including template matching, Support Vector Machine (SVM) [163], and linear programming [62], are systematically examined using several public databases. Extensive experiments illustrate that LBP features are effective and efficient for facial expression recognition.

Most of the existing work attempts to recognize facial expressions from data collected in a highly controlled environment with very high resolution frontal faces [152]. However, in real-world environments, input face images are often in lower resolution. Tian *et al.* [153, 152] recently made attempts to recognize facial expressions with lower resolution. In this work, we also investigate LBP features for low-resolution facial expression recognition. In addition to the evaluation on different image resolutions, we performed experiments on real-world compressed low-resolution video sequences. It is observed that LBP features perform stably and robustly over a useful range of low resolutions of face images, yielding promising performance in compressed video sequences captured in real-world environments.

We then adopt AdaBoost [55] to extract the most discriminative LBP features from face images for better expression recognition. However, AdaBoost requires very expensive training time. We present an efficient learning procedure based on Conditional Mutual Information (CMI) [54]. The CMI based algorithm learns a sequence of weak classifiers by maximizing their mutual information about a candidate class, conditional to the response of any weak classifier already selected, thus avoiding the selection of ineffective weak classifiers. Extensive experiments show that the CMI based method enables much faster training, and the best recognition performance can be obtained by using SVM classifiers with the selected LBP features. However, the recognition is performed on static images without exploiting temporal behaviours of facial expression.

1.2.2 Manifold Analysis of Facial Expression

The temporal dynamics of human behaviour is a critical factor for successful interpretation of the observed behaviour [117]. The differences between facial (or bodily) expressions are often conveyed more powerfully by dynamic transitions between different stages of expressions

rather than any single state represented by a still image, and this is especially true for spontaneous facial expressions without any deliberate exaggerated posing [2]. In this work, we present a method to capture and represent the expression dynamics by discovering the underlying low-dimensional manifold.

To address the limitations in the existing work [25, 26, 72], we exploit Locality Preserving Projections (LPP) [66] to learn the expression manifold in the LBP based dense appearance feature space. Compared to Locally Linear Embedding (LLE) [135] and Isomap embedding [151] adopted previously in [25, 72, 89], LPP provides explicit mapping from the input space to the reduced space, so is better suited to facial expression recognition. The LBP based appearance features offer advantages over the sparse 2D feature points [25, 26, 72] in describing detailed facial deformations that are important to facial expression modeling. One challenging problem in expression manifold learning is to obtain a generalized representation for facial expressions from different subjects. By deriving a universal discriminant expression subspace using a supervised LPP (SLPP), we effectively align manifolds of different subjects on a generalized expression manifold. More crucially, we evaluate our approach with a large number of subjects, and its generalization ability and robustness is evidently verified.

Although different linear subspace techniques have been developed, it is still unknown which technique is most suitable for discriminant expression subspace learning. Therefore, we comprehensively evaluate linear subspace methods, including traditional Principal Component Analysis (PCA) [158] and Linear Discriminant Analysis (LDA) [15], and recently proposed graph-based methods LPP, SLPP, Orthogonal Neighborhood Preserving Projections (ONPP) [84], and Locality Sensitive Discriminant Analysis (LSDA) [24]. Extensive comparative experiments on several databases demonstrate that SLPP is superior in expression subspace learning on databases used.

We further formulate a Bayesian framework to examine both the temporal and appearance characteristics for dynamic facial expression recognition employing the derived manifold representation. Our method provides superior performance to both static frame-based methods and state-of-the-art dynamic models [176] in person-dependent recognition experiments. As facial expressions vary in intensity, it is helpful to estimate the expression intensity for quan-

titative assessment of facial expression. We show that the expression intensity can be easily estimated on the expression manifold using the Fuzzy K-Means method [18]. However, our manifold based representation only addresses temporal dynamics or correlations of the whole face image, does not consider the spatial-temporal correlations among different facial regions.

1.2.3 Correlation Analysis of Facial Parts

As facial muscles are contracted in unison to display facial expressions, different facial parts have strong correlations. Capturing and analyzing correlations among facial parts are important for modeling facial expressions precisely. Most of the existing work on facial expression analysis did not explicitly model these correlations. In this work, we employ Canonical Correlation Analysis (CCA) [70], a statistical technique that is well suited for relating two sets of signals, to model correlations among facial parts.

When applying CCA to image data, the original two-dimensional images have to be reshaped into one-dimensional vectors, as the traditional CCA is based on the vector-space model. However, this matrix-to-vector operation leads to some problems. For example, the intrinsic 2D structure of image matrices is removed, so the spatial information stored therein is discarded. To address these problems, we introduce a novel Matrix-based Canonical Correlation Analysis (MCCA) for better correlation analysis of 2D image or matrix data in general. MCCA takes a 2D matrix based data representation rather than the 1D vector based representation in classical CCA. MCCA seeks canonical factors in two dimensions to maximize the correlations between two sets of matrices. Unlike classical CCA, there is no closed-form solution for the optimization problem in MCCA. Instead, we propose an iterative solution with a convergence proof. We evaluate the proposed MCCA for capturing correlations of facial parts. Experimental results demonstrate that MCCA can better measure correlations in 2D image data, providing superior performance in regression and recognition tasks, whilst requiring much fewer canonical factors.

All the above work focuses on facial expressions. However, the face is usually perceived not as an isolated object but as an integrated part of the whole body. As indicated in the psychological study [3], the visual channel combining facial and bodily expressions is most

informative.

1.2.4 Multimodal Facial and Body Language Analysis

We investigate two understudied problems in body language analysis, gait-based gender discrimination and affective body gesture recognition, and further integrate face and body cues for improved performance. Gender classification is an important visual task for human beings, as many social interactions critically depend on the correct gender perception. A large number of studies have investigated gender classification by human faces [104]. However, in unconstrained real-world situations, for example, when people walking at a distance, face information is either unreliable or unavailable. In contrast, human gait can be detected at a distance or at low resolution, and psychophysical studies [85, 8, 99, 126] suggest that people can recognize the gender of walkers by their gaits. Therefore, human gait may provide important alternative cues for gender discrimination. However, there have been few studies on gait-based gender recognition in computer vision. Compared to facial gender classification, this problem is relatively understudied. In this work, we investigate gender discrimination by human gaits in image sequences using machine learning methods. With the Gait Energy Image (GEI) [64] based gait representation, our experiments illustrate that a linear decision surface can be derived to discriminate gender with high confidence (93-94%).

In affective computing, emotion analysis from facial expressions has been widely studied. However, little attention has been placed on affective body posture and gesture analysis, although bodily expression plays a vital role in conveying emotional states, and the perception of facial expressions is strongly influenced by the concurrently presented body language [3, 100]. This is probably due to the high variability of emotional body posture and gesture that can be displayed. Existing studies on vision-based gesture recognition have been primarily carried out on non-affective gestures such as sign languages [171]. In this work, we investigate affective body gesture analysis in video sequences. Particularly, we exploit spatial-temporal features based on space-time interest point detection [43] for representing body gestures in videos.

Each modality, face or body, in isolation has its inherent weakness and limitations. In real

life, human face and body are indeed perceived as an integrated whole. So fusing face and body cues provides a potential way to accomplish improved gender discrimination or affect analysis. A recent psychological study [100] suggest the integration of facial expressions and body gestures is a mandatory automatic process occurring early in human processing stream. Therefore, face and body cues cannot be considered mutually independently and combined at decision or module level but, on the contrary, should be processed in a joint feature space [73]. We exploit CCA to fuse the two modalities at the feature level. Our motivation is that, as face and body cues are two sets of measurements for gender or affective states, conceptually the two modalities are correlated, and their relationship can be established using CCA. CCA derives a semantic "gender" or "affect" space, in which face and body features are compatible and can be effectively fused. Experiments on large dataset [178, 61] demonstrate that our multimodal recognition system achieves better recognition performance in gender discrimination and affect analysis than that based on the single modality.

1.3 Contributions

The main contributions of this thesis can be summarized as follow:

1. We empirically investigate facial representation based on LBP features for person-independent facial expression recognition [143], including the low-resolution settings [142]. Different machine learning methods are systematically examined on several public databases. We also present and evaluate a Conditional Mutual Information based algorithm for efficient LBP feature selection [141].

2. We exploit Locality Preserving Projections to learn the facial expression manifold in the dense appearance feature space. By deriving a universal discriminant expression subspace using a supervised LPP, we align manifolds of different subjects on a generalized expression manifold [140]. Different linear subspace learning techniques are comprehensively evaluated in expression subspace learning [144]. We formulate a Bayesian framework for dynamic facial expression recognition employing the derived manifold representation [145].

3. We employ Canonical Correlation Analysis to model the correlations among facial parts. To overcome the inherent limitations of the traditional CCA for image data, we propose a novel Matrix-based Canonical Correlation Analysis for better correlation analysis of 2D image or matrix data in general [147].

4. We investigate two understudied problems, gait-based gender discrimination and affective body gesture recognition, and further integrate face and body cues for improved performance [148, 146]. CCA is adopted to establish the relationship between the two modalities, and derive a semantic joint feature space for the feature-level fusion.

1.4 Thesis Outline

The thesis is structured as follows. We present in Chapter 2 a critical review of previous work on machine analysis of facial and body language. In Chapter 3, we investigate Local Binary Pattern based facial representation. Chapter 4 introduces manifold learning of facial expression, and empirical evaluation on different linear subspace techniques. A Bayesian framework is presented for dynamic facial expression recognition employing the derived manifold representation. In Chapter 5, we discuss capturing correlations among facial parts using Canonical Correlation Analysis, and propose a Matrix-based CCA for better correlation analysis of 2D image. Chapter 6 presents studies on gait-based gender discrimination and affective body gesture recognition, and multimodal analysis by combining face and body cues. Chapter 7 concludes the thesis and discusses future work.

2 Literature Review

Because of its important applications and the theoretical interests from cognitive and psychological scientist, machine analysis of facial and body language has attracted much attention in the last decade [120, 117, 73]. Although much progress has been made, it is still rather difficult to develop a computer vision system capable of automatically understanding facial and body language. In this chapter, we review previous work on facial expression analysis (Section 2.1) and body language analysis (Section 2.2), to give a foundation and context for the work presented in this thesis.

2.1 Facial Expression Analysis

Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications [155]. Facial expression analysis has received interest from behavioural scientists since the work of Darwin in 1872 [38]. Suwa *et al.* [150] made the first attempt to automatically analyze facial expressions from image sequences in 1978. There have been major advances in the computer vision literature for facial expression analysis over the last two decades. See survey papers [118, 52, 120, 155, 115] for detailed review.

Facial expressions can be described at different levels [155]. Two mainstream description methods are facial affect (emotion) and facial muscle action (action unit) [115]. Psychologists suggest that some basic emotions are universally displayed and recognized from facial expressions [38, 49], and the most commonly used facial expression descriptors are the six basic emotions [29] (anger, disgust, fear, joy, surprise, and sadness; see Figure 2.1 for examples). This is also reflected by the research on automatic facial expression analysis. Most facial expression analysis systems developed so far target facial affect analysis, and attempt

Figure 2.1: Prototypic emotional facial expressions: Anger, Disgust, Fear, Joy, Sadness, and Surprise (from left to right). From the Cohn-Kanade database [78].

to analyze a set of prototypic emotional facial expressions [118, 120]. There have also been some tentative efforts to detect cognitive and psychological states like interest [77], pain [96], and fatigue [59]. To describe subtle facial changes, Facial Action Coding System (FACS) [48] has been widely used for manually labeling of facial actions in behavioural science. FACS associates facial changes with actions of the muscles that produce them. It defines 44 different action units (AUs) (see Figure 2.2 for some examples of AUs). It is possible to map AUs onto the basic emotions using a finite number of rules [48]. Automatic AU detection has been widely studied recently [44, 154, 121, 181, 95]. The major problem of AU-related research is the need of highly trained experts to manually perform FACS coding frame by frame. Approximately 300 hours of training are required to achieve minimal competency of FACS, and each minute of video tape takes around two hours to code comprehensively [20]. Another possible descriptor is the bipolar dimensions of *Valence* and *Arousal* [133]. Valence describes the pleasantness, with positive (pleasant) on one end (e.g. happiness), and negative (unpleasant) on the other (e.g. disgust). The other dimension is arousal or activation, for example, sadness has low arousal, whereas surprise has a high arousal level. Different emotional labels can be plotted at various positions on a two-dimensional plane scanned by these two axes.

The general approach to automatic facial expression analysis consists of three steps [155]: face acquisition, facial feature extraction and representation, and facial expression recognition.

1. Face acquisition is a pre-processing stage to detect/locate the face region in input images or sequences. Numerous techniques have been proposed for face detection [74],

Figure 2.2: Examples of facial action units and their combination defined in FACS [115].

due to its practical importance in many computer vision applications. The real-time face detection scheme proposed by Viola and Jones [165] is arguably the most commonly employed face detector, which consists of a cascade of classifiers trained by AdaBoost employing Harr-wavelet features. The detected face region is usually aligned based on the eye position that can be detected in the face region [156]. To handle large head motion in video sequences, head tracking and pose estimation can also be adopted.

2. After locating the face, the next step is to extract facial features from original face images for facial representation. There are mainly two approaches to this task: geometric feature-based methods and appearance-based methods [74].

3. The last stage is to classify different expressions based on the extracted facial features. Depending on whether the temporal information is used, the recognition approaches are generally divided as image-based or sequence-based.

In the following sections, we survey the existing work in facial feature representation (Section 2.1.1) and facial expression recognition (Section 2.1.2).

2.1.1 Facial Feature Representation

Facial feature representation is to derive a set of features from original face images to effectively represent faces. If inadequate features are used, even the best classifier could fail
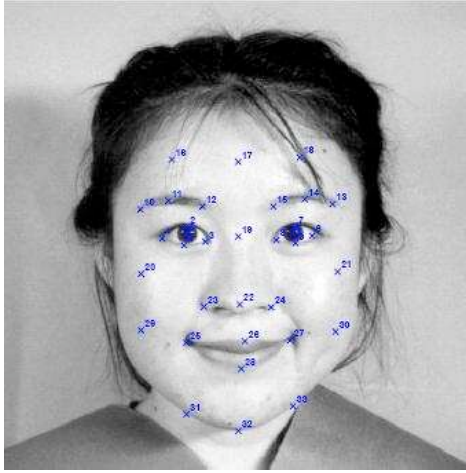
Figure 2.3: Geometric features [182]: 34 fiducial points for representing the facial geometry.

to achieve accurate recognition. Two types of features can be extracted: geometric features and appearance features [155]. Geometric features present the shape and locations of facial components (including mouth, eyes, brows, and nose), which are extracted to form a feature vector that represents the face geometry. Appearance features present the appearance changes (skin texture) of the face, including wrinkles, bulges and furrows. Image filters, such as Gabor wavelets [39], can be applied to either the whole-face or specific facial regions to extract appearance features.

Geometric Features

Fiducial facial feature points have been widely used in facial representation. Zhang *et al.* [182] used the geometric positions of 34 fiducial points (as shown in Figure 2.3) as facial features. A shape model defined by 58 facial landmarks was adopted in [25, 26]. Pantic and her colleagues [121, 116, 159] also utilized a set of facial characteristic points to describe facial actions. They developed a robust facial point detector [166], and some detection results are shown in Figure 2.4.

In [153], Tian considered two types of geometric facial features, location features and shape features. Specifically, the author extracted six location features (eye centers, eyebrow inner endpoints, and corners of the mouth), which are transformed into 5 parameters (as shown in Figure 2.5), and the mouth shape features. To extract the latter, an edge detector is first

Figure 2.4: The facial point detection results [115].

Figure 2.5: Geometric features [153]: (*Left*) location features; (*Right*) normalized face and zones of the edge map of the normalized face.

applied to the normalized face to get the edge map, which is divided into 3×3 zones as shown in Figure 2.5; the mouth shape features are then computed from zonal shape histograms of the edges in the mouth region.

In image sequences, facial movements can be qualified by measuring the geometrical displacement of facial feature points between the current frame and the initial frame. This method has shown validity in previous work [119, 154, 77]. Tian *et al.* [154] developed multi-state facial component models to track and extract the geometric facial features, including lip, eyes, brows and cheeks (as shown in Figure 2.6). Cohen *et al.* [28] adopted a model-based face tracker to track head motion and local deformation of facial features such as the eyebrows, eyelids, and mouth. The tracked motions of various facial features at each frame
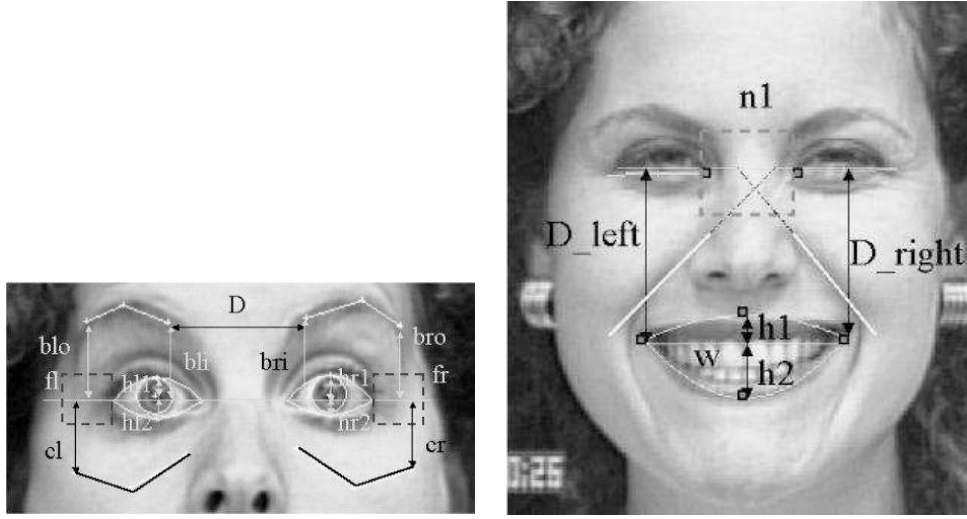
Figure 2.6: Geometric and appearance features [154]. (*Left*) Upper face features: 12 parameters describe the motion and shape of the eyes, brows, and cheeks; 2 parameters describe the state of crow's-feet wrinkles, and 1 parameter describes the distance between the brows. (*Right*) Lower face features: 6 parameters describe lip shape, state and motion, and 3 describe the furrows in the nasolabial and nasal root regions.

are referred as Motion-Units (MUs) (as shown in Figure 2.7). The MUs represent not only the activation of a facial region, but also the direction and intensity of the motion. Infrared eye (pupil) detection and tracking [80, 181, 59] have been adopted to enhance facial motion measurement. For example, Kapoor *et al.* [80] used the infrared eye tracking to localize and normalize eye and eyebrow regions, which are analyzed using PCA to recover shape parameters.

As facial movements can produce optical flow in the image, optical flow analysis has been widely used to model muscles activities or estimate the displacements of feature points [172, 51, 69, 176]. For example, Essa and Pentland [51] utilized optic flow to estimate facial activity in a detailed anatomical and physical model of the face. Motion estimates from optic flow were refined by the physical model in a recursive estimation and control framework and the estimated forces were used to classify facial expressions. Although it is effective to obtain facial motion information by computing dense flow between successive image frames, flow estimation has its disadvantages such as easily disturbed by lighting variation and non-rigid
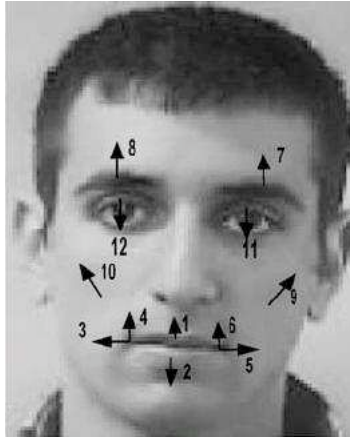
Figure 2.7: The Motion-Units introduced in [28].

motion, and sensitive to the inaccuracy of image registration and motion discontinuities [181].

The geometric feature-based facial representations commonly require accurate and reliable facial feature detection and tracking, which is difficult to accommodate in many situations. More crucially, geometric features usually cannot encode changes in skin texture such as wrinkles and furrows that are critical for facial expression modeling. It is especially difficult to describe subtle spontaneous facial expressions using sparse geometric features. In addition, Tian's experiments in [153, 152] demonstrate that geometric features are not available or reliable in low-resolution facial images captured in real environments.

Appearance Features

In contrast to geometric features, appearance features encode changes in skin texture such as wrinkles, bulges, and furrows. Appearance features include Gabor wavelets [39, 97], Harr-like wavelets [165, 168], the learned statistical image filters such as PCA [158], LDA [15, 97], Independent Component Analysis (ICA) [44, 11], and Local Feature Analysis (LFA) [44], those based on Active Appearance Model (AAM) [86, 32], temporal templates [161], and features based on edge-oriented histograms. Appearance-based methods suffer less from issues of initialization and tracking errors.

Most of the existing appearance-based methods have adopted Gabor-wavelet features [182, 97, 62, 95, 9, 156]. Gabor filters are obtained by modulating a 2D sine wave with a Gaussian
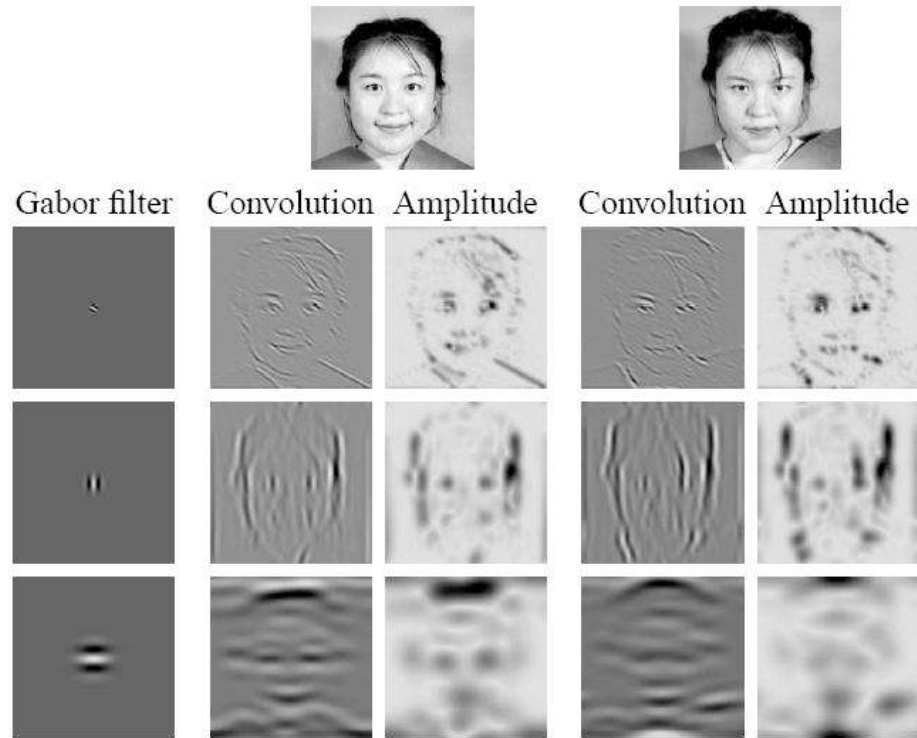
Figure 2.8: Gabor-wavelet representation [182]: two examples with three Gabor kernels.

envelope. Representations based on the outputs of Gabor filters at multiple spatial scales, orientations, and locations have proven successful for facial image analysis. For example, in Lyons *et al.* 's work [97], each face is represented using a set of Gabor filters at the facial feature points sampled from a sparse grid covering the face. Zhang *et al.* [182] compared geometric features (the geometric positions of 34 fiducial points, as shown in Figure 2.3) and a set of multi-scale and multi-orientation Gabor-filters coefficients at these points (as shown in Figure 2.8). Experimental results show that Gabor-wavelet coefficients are much more powerful than geometric positions. This is possibly because Gabor features can better describe facial deformation in details. Tian [152] compared geometric features and Gabor-wavelet features with different image resolutions, and her experiments show that Gabor-wavelet features work better for low-resolution face images.

Donato *et al.* [44] explored different approaches to facial representation for facial action recognition, which includes holistic spatial analysis, such as PCA, ICA, LFA and LDA, and methods based on the outputs of local filters, such as Gabor wavelet representation and local

principal components. Best performances were obtained using the Gabor-wavelet representation and the ICA representation. These experimental results provide converging evidence for the importance of using local filters, high frequencies, and statistical independence for classifying facial actions.

Active Appearance Model (AAM) [32] describes both shape and appearance in the PCA space. Utilizing AAM, Lanitis *et al.* [86] characterize face deformation due to facial expressions by a set of appearance parameters, which could be used to determine facial expressions. Zeng *et al.* [179] used a 3D face tracker to extract facial texture images. A holistic, monochrome, spatial-ratio face template was used in [5].

Feature selection methods have been exploited to select the most effective appearance features. Guo and Dyer [62] introduced a linear programming technique, Feature Selection via Linear Programming (FSLP), that jointly performs feature selection and classifier training so that a subset of features is optimally selected together with the classifier. Comparison experiments on the JAFFE database showed that the performance of FSLP is comparable with SVM while requiring much fewer features. Wang *et al.* [168] boosted Harr feature based Look-Up-Table weak classifiers using Adaboost for facial expression recognition. Bartlett *et al.* [12, 10] selected a subset of Gabor filters using AdaBoost. More recently Whitehill and Omline [170] compared Gabor filters, Harr-like filters, and the edge-oriented histogram for AU recognition using SVMs and Adaboost as the classifiers. They found that AdaBoost performs better with Harr-like filters, while SVMs perform better with Gabor filters. This may be attributed to the fact that the pool of Harr features was much larger. AdaBoost performs feature selection and does well with redundancy, while SVMs were calculated on the full set of filters and don't do well with redundancy.

The appearance-based representations contain more information than those representations based on the relative positions of a finite set of facial features [9], so usually providing superior performance [182, 44, 152]. For example, the Gabor-wavelet representation [95] outperforms the performance upper-bound computed based on manual feature tracking. However, some recent studies indicate that this claim does not always hold [162, 115]. For example, Valstar *et al.* [162, 159] presented a AU detection method using the facial representation based on

tracked facial points, which detects a similar number of AUs with similar or higher recognition rates than other methods. It seems that using both geometric and appearance features might be the best choice [115]. Tian *et al.* [154] considered both permanent facial features (brows, eyes, mouth) and transient facial features (deepening of facial furrows) for AU analysis (see Figure 2.6).

Although Gabor-wavelet based representations have been widely adopted, it is computationally expensive to convolve face images with a set of Gabor filters to extract multi-scale and multi-orientation coefficients. It is inefficient in both time and memory due to the high redundancy of Gabor-wavelet features. For example, in [10], the Gabor-wavelet representation derived from each 48×48 face image has the high dimensionality of $O(10^5)$. Local Binary Patterns (LBP) features recently have been introduced for facial images analysis [1, 63]. The most important properties of LBP features are their tolerance against illumination changes and their computational simplicity. In this thesis (Chapter 3), we investigate in more details facial representation based on LBP features for facial expression analysis.

Despite the high dimensionality of face image space, face images lie intrinsically on much lower dimensional subspaces. Subspace analysis method such as PCA, LDA and ICA have been widely exploited in facial expression analysis [44]. Recently a host of linear subspace methods [66, 84, 24] have been developed. However, it is still unknown whether these graph-based methods are effective for facial expression analysis. In this thesis (Chapter 4), we comprehensively investigate and compare a number of linear subspace methods for facial expression analysis.

2.1.2 Facial Expression Recognition

Many classification techniques have been applied to recognize facial expressions, including Neural Networks (NN) [182, 154, 162], Support Vector Machines (SVM) [80, 95, 5], Bayesian Networks (BN) [28, 27], k-Nearest Neighbor (kNN) [97, 44], rule-based classifiers [119, 121, 116], Hidden Markov Models (HMM) [113, 28, 176], Dynamic Bayesian Networks (DBN) [77, 181, 156], and so on. In this section, we review image-based (or static) expression recognition and sequence-based (dynamic) expression recognition respectively. The image-

based methods use features extracted from a single image to recognize the expression of that image, while the sequence-based approaches aim to capture the temporal pattern in a sequence to recognize the expression for one or more images.

Image-Based Expression Recognition

Lyons [97] adopted a nearest neighbour classifier to recognize facial images using discriminant features computed by applying PCA and LDA to the Gabor-wavelet features. With geometric features extracted using a dual-view (front and profile) face model, Pantic and Rothkrantz [119] performed facial expression recognition by comparing the AU-coded description of an observed expression against rule descriptors of six basic emotions. Recently they further adopted the rule-based reasoning to recognize action units and their combination [121]. Yacoob and Davis [172] used local motions of facial features to construct a mid-level description of facial motions, which were classified into one of six facial expressions using a set of heuristic rules.

Tian *et al.* [154] used a three-layer Neural Network with one hidden layer to recognize action units by a standard back-propagation method. The input to the NN are the parametric descriptions of nontransient and transient facial features derived from their multistate face and facial component models. Most of the existing facial expression recognition approaches attempt to recognize facial expressions from data collected in a highly controlled environment with very high resolution frontal faces (e.g. 200×200 pixels) [152]. Tian *et al.* recently made attempts to recognize facial expressions with lower resolution (e.g. 50×70 pixels) [153, 152]. In the real-time system described in [153], to handle the full range of head motion, not face but head was first detected, and the head pose was then estimated. For faces of frontal and near frontal views, the geometric features were computed as inputs to a NN for recognition. Tian [152] further explored the effects of different image resolutions for each step of facial expression analysis.

Cohen *et al.* [28] adopted Bayesian network classifiers to classify each frame as one of the basic emotions based on face tracking results. They compared Naive-Bayes classifiers where the features are assumed to be either Gaussian or Cauchy distributed, and Gaussian Tree-

Augmented Naive (TAN) Bayes classifiers. Naive-Bayes classifiers use a very strict and often unrealistic assumption that the features are independent given the class, while Gaussian TAN classifiers can learn dependencies between the features without adding much complexity, and the resultant tree structure is assured to maximize the likelihood function. It is difficult to collect a large amount of training data. Moreover, data labeling is time-consuming, error prone, and expensive. It is very beneficial to construct methods that use scarcely available labeled data and abundant unlabeled data. To address these problems, Cohen *et al.* [27] further proposed to use unlabeled data together with labeled data using Bayesian networks. However, they also pointed out that adding unlabeled data can be detrimental to the performance.

As a powerful discriminative machine learning technique, SVM has been widely adopted for facial expression recognition. Recently Bartlett *et al.* [10, 95] performed systematic comparison of different techniques including AdaBoost, SVM, and LDA for facial expression recognition, and best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training SVM on the outputs of the selected filters. This strategy is also adopted in [156, 159]. For example, Valstar *et al.* [159] recognized AU temporal segments using a subset of most informative spatio-temporal features selected by AdaBoost.

Most of the existing work have been carried out on expression data that were collected by asking subjects to deliberately pose facial expressions [97, 78, 123]. However, these exaggerated facial expressions occur rarely in real-life situations. Spontaneous facial expressions induced in natural environments are more subtle and fleeting, such as tightening of the lips in anger or lowering the lip corners in sadness [155]. A psychological study [50] has indicated that posed expressions may differ in appearance and timing from spontaneous ones. Recently research attention has started to shift to spontaneous facial expression analysis [137, 30, 149, 179, 9]. Sebe *et al.* [137] collected an authentic facial expression database containing spontaneous emotions, and compared a wide range of classifiers, including Bayesian Network, the Decision Trees, SVM, k-Nearest-Neighbor, Bagging, and Boosting, for spontaneous expression recognition. Surprisingly, the best classification results were obtained with the kNN classifier. It seems that all the models tried were not able to entirely capture the

complex decision boundary that separates different spontaneous expressions. Cohn *et al.* [30] developed an automated system to recognize brow actions in spontaneous facial behaviour captured in interview situations. Their recognition accuracy was relatively worse than that for the posed facial behaviour. Recently Zeng *et al.* [179] treated the problem of emotional expression detection in a realistic conversation setting as an one-class classification problem, and adopted Support Vector Data Description to distinguish emotional expressions from non-emotional ones. Bartlett *et al.* [9] recently presented preliminary results on facial action detection in spontaneous facial expressions by adopting their AU recognition approach [95].

	Subjects	Expressions	Type	Labeled
JAFFE Database [97]	10	7 classes	Posed	Yes
Cohn-Kanade Database [78]	100	wide range	Posed	Yes
MMI Database [123]	53	wide range	Posed/Spontaneous	Yes
Authentic Expression Database [137]	28	4 classes	Spontaneous	Yes
RU-FACS [9]	100	wide range	Spontaneous	Yes
UTDallas-HIT [114]	284	11 classes	Spontaneous	No
MIT-CBCL [149]	12	9 classes	Spontaneous	Yes

Table 2.1: Summary of the existing databases of facial expressions.

We summarize the existing databases of facial expressions in Table 2.1. Although several databases containing spontaneous facial expressions have been reported recently [137, 114, 9, 149], most of them currently are not available to the public, due to ethical and copyright issues. In addition, manual labeling of spontaneous expressions is very time consuming and error prone due to subjectivity. One of the available databases containing spontaneous facial expression was collected at UT Dallas [114], which contains videos of more than 200 subjects. Videos of spontaneous expressions (including happiness, sadness, disgust, puzzlement, laughter, surprise, and boredom) were captured when the subject watching videos that intend to elicit different emotions. The MMI database [123] also contains some (currently 65) videos of spontaneous facial displays [115]. Some other databases [46] were recorded from talk TV shows which contain speech-related spontaneous facial expressions.

Sequence-Based Expression Recognition

While image-based expression recognition is based on the static facial configuration from still images, sequence-based expression recognition models temporal behaviours of facial expressions from image sequences. Psychological experiments [13] suggest that the dynamics of facial expressions are crucial for successful interpretation of facial expressions. This is especially true for natural facial expressions without any deliberate exaggerated posing [2].

Hidden Markov Model is one of the basic probabilistic tools used for time series modeling, and has been widely used for temporal interpretation in speech recognition [127]. HMMs have also been exploited to capture temporal behaviours exhibited by facial expressions. Oliver *et al.* [113] applied HMM for facial expression recognition based on the tracked deformation of mouth shapes in real-time. Each of the mouth-based expressions, e.g. sad and smile, is associated with an HMM trained by using the mouth features, and the facial expression is identified by computing the maximum likelihood of the input sequence with respect to all trained HMMs. Cohen *et al.* [28] proposed a multi-level HMM classifier, which not only performs expression classification on a video segment, bust also automatically segments an arbitrary long video sequence to different expressions segments without resorting to heuristic methods of segmentation. Recently Yeasin *et al.* [176] presented a two-stage approach to classify six basic emotions, and derive the level of interest using psychological evidences. First, a bank of linear classifiers were applied at frame level and the output was coalesced to produce the temporal signature for each observation. Second, temporal signatures computed from the training data set were used to train discrete HMMs to learn the underlying models for each expression.

Dynamic Bayesian Networks are graphical probabilistic models which encode dependencies among sets of random variables evolving in time, with efficient algorithms for inference and learning. HMM actually is a simplified version of DBNs. DBNs are capable of accounting for uncertainty in the facial expression recognition, representing probabilistic relationships among different actions and modeling the dynamics in facial action development. Zhang and Ji [181] explored the multisensory information fusion technique with DBNs for modeling and

understanding temporal behaviours of facial expressions in image sequences. By integrating DBN with a general-purpose facial behaviour description language, Gu and Ji [58] further proposed a task-oriented DBN to represent and classify facial events of interest, which can incorporate prior knowledge of a given application. Hoey and Little [69] used DBNs in unsupervised learning and clustering of facial displays. Kaliouby and Robinson [77] developed a system for inferring complex mental states from videos of facial expressions and head gestures in real-time. Their system was built on a multi-level DBN classifier which models complex mental states as a number of interacting facial and head displays, identified from component-based facial features. More recently, Tong *et al.* [156] proposed to exploit the relationships among different AUs and their temporal dynamics using DBN.

Spontaneous facial expressions differ from posed expressions both in terms of which muscles move and how they move dynamically. For example, spontaneous expressions have fast and smooth onsets, with distinct facial action peaking simultaneously, while posed expressions tend to have slow and jerky onsets, and the actions typically do not peak simultaneously [9]. So facial dynamics is a key parameter in differentiating posed expressions from spontaneous ones [160]. A study in [31] indicates that posed smiles were of larger amplitude and has less consistent relationship between amplitude and duration than spontaneous smiles. Recently Valstar *et al.* [160] experimentally showed that temporal dynamics of spontaneous and posed brow actions are also different from each other. They built a system to automatically discern spontaneous brow actions from deliberately posed ones, based on the parameters of temporal dynamics such as speed, duration, intensity, and the occurrence order. They achieved the classification rate of 90.7% on 189 samples taken from three different databases.

Recently Pantic and Patras [116] introduced facial-action-dynamics recognition using temporal rules on profile-view face image sequences. Particle filtering was exploited to track 15 facial points in input face-profile sequences. Their algorithm performs both automatic segmentation of an input video into facial expressions and recognition of temporal segments (i.e. onset, apex, offset) of 27 AUs occurring alone or in a combination. A recognition rate of 87% was reported on database used.

Bettinger and Cootes [17] described a system prototype to model both the appearance and

behaviour of a person's face. AAM was used to model the appearance of the individual, and an image sequence was represented as a trajectory in the parameter space. They presented a method to break the trajectory into segments, and used a variable length Markov model to learn the relations between groups of segments. Given a long training sequence for an individual containing repeated facial behaviours such as moving head and changing expression, their system can learn a model capable of simulating these simple behaviours. However, how to model facial dynamics for facial expression recognition was not considered in their work. Lee and Elgammal [89] recently introduced a framework to learn decomposable generative models for dynamic appearance of facial expressions where facial motion is constrained to one dimensional closed manifolds.

Variations of face images can be represented as low dimensional nonlinear manifolds embedded in a high dimensional input space [151, 135, 66]. The expression dynamics can be captured in low dimensional manifolds. Chang *et al.* [25] made first attempt to learn the structure of the expression manifold. They compared Locally Linear Embedding (LLE) [135] and Lipschitz embedding [25] for expression manifold learning. In [26], they further proposed a probabilistic video-based facial expression recognition method using manifolds. By exploiting Isomap embedding [151], they also developed an approach for facial expression tracking and recognition [72]. However, there are several noticeable limitations in their work. First, as face images are represented by a shape model defined by sparse 2D feature points, expression manifolds were learned in a facial geometric feature space. Consequently most detailed facial deformations important to capture expressions such as wrinkles and furrows were ignored. There is a need to learn expression manifolds using a much more dense representation. Second, a very small data set was used to develop and verify their models, e.g. two subjects were considered in [25, 72]. To verify a model's generalization potential and robustness, the expression manifold should be evaluated on a large number of subjects. Third, on facial expression recognition, their approach [26] is subject dependent in that each subject was represented by a separate manifold, so the generality and scalability of the approach is still unknown. Moreover, no quantitative evaluation was given in their papers to provide comparison. In this thesis, we aim to address these limitations in Chapter 4.

Human face is usually perceived not as an isolated object but as an integrated part of the whole body. The face and the body both contribute in nonverbal communication. Psychological studies suggested [3, 100] suggest that the visual channel combining facial and bodily expressions is most informative, and the recognition of facial expression is strongly influenced by the concurrently presented emotional body language. Therefore, beyond facial expressions, visual analysis of bodily expressions has generated much interest in computer vision.

2.2 Body Language Analysis

Body language denotes human bodily expression, including body configuration or posture, body movement, gesture, gait, etc. Like facial expression, bodily language is also the main means for humans to communicate their emotion, mood, attitude, and attention. Body language also reveals other information including identity, gender, age, attractiveness, and personality. Vision-based analysis of bodily expression in image sequences, including body posture (or configuration) analysis and body gesture (including gait) analysis, has become one of the most active fields in computer vision. Most of the existing work can be classified as model-based (i.e. use geometric primitives like cones and sphere to model body parts) or appearance-based (i.e. use color, shape or texture information to represent body or body parts). Much progress has been made in visual human motion analysis in the last two decades. See the survey papers [167, 103] for detailed review.

In the following sections, we focus on a survey of previous work on gait analysis for gender recognition and affective body language analysis.

2.2.1 Gait Analysis for Gender Recognition

Human gait, or the style of walking, has been studied in medical science, psychology, and biomechanics for decades [167, 109]. As a unique non-invasive biometric that can be detected and measured at a distance or at low resolution, gait has received much attention in computer vision for human identification recently [134, 110, 109].

The existing approaches to gait analysis can be categorized as model-free analysis and model-based analysis [109]. Model-free approaches are mainly silhouette-based, which first derive the human silhouette by separating the moving walker from the background, and then extract measurements that reflect shape and/or movement for recognition. The simplest approach is to form an average of the silhouette whereas the more complex impose a model on the motion. In model-based approaches, a body model is fitted to the human in every frame to derive kinematic parameters, which describe the movement of the torso and/or the legs. Unlike silhouette-based methods, model-based approaches usually concentrates on body dynamics, omitting body shape.

In addition to identity recognition, gait can also be used for gender discrimination, which was suggested and supported in psychological studies [85, 8, 99]. However, gait-based gender recognition has not been well investigated in computer vision. On the contrary, as human faces provide important visual information for gender perception, a very large number of studies have been focused on gender classification by face.

In the early 1990s various neural network techniques were employed to recognize gender by frontal faces [56, 21]. For example, Golomb *et al.* [56] trained a fully connected two-layer neural network, SEXNET, to identify gender from face images. Recently Moghaddam and Yang [104] investigated nonlinear SVMs for gender classification with low-resolution thumbnail faces, and demonstrated the superior performance of SVMs to other classifiers. Shakhnarovich *et al.* [139] developed a real-time face detection and demographic analysis (female/male and asian/non-asian) system using Adaboost, which delivers slightly better performance than the nonlinear SVMs [104] on unaligned faces from real-world video sequences. Although gender discrimination by face has been widely studied, in unconstrained real-world situations, face information is not always available or reliable, due to the arbitrary walking direction and continuously varying head pose. More crucially, with people walking at a distance, face information can not be measured reliably at low resolution. In these situations, human gait can provide important alternative cues for gender classification, as gaits can be detected and measured at a distance. Therefore, it is necessary to investigate gender discrimination by gait using computer vision techniques.

Gender recognition from the point-light display (as shown in Figure 2.9) of human walking has received much attention in psychological field during the past decades. Kozlowski and Cutting [85] performed the first major experiment with six walkers (three females and three males) of approximately the same height and weight recorded at a sagittal view. Their results showed that human observers can correctly identify the gender of walkers with average recognition rate of 63%, and alterations such as varying the arm swing, changing the walking speed, and occluding portions of the body do not significantly influence recognition performance. Barclay *et al.* [8] carried out further study by examining temporal and spatial factors. They suggested that successful gender recognition requires exposure to approximately two walking cycles, and the rendering speed has a strong influence over recognition. The effect of inversion on the point-lights was also investigated, and it is found that the gender assignments are significantly reversed. They proposed a view-based explanation based on the shoulder-hip ratio, in which men tend to have broader shoulders and smaller hips than women. Cutting *et al.* [36] supported the shoulder-hip concept and proposed a related center-of-moment feature of the torso.

The shoulder-hip ratio and center-of-moment features [8, 36] are mainly based on the structural differences between male and female walkers. There are additional dynamic features of movement that contribute to recognition. Mather and Murdoch [99] found that males have greater shoulder swing and females have more hip swing. Troje [157] compared structural and dynamic features, where dynamic-only stimuli (movement applied to averaged postures) produced better results than with structural information (postures using averaged motions). Although most of the study was conducted using a side-view presentation of walkers to observers, the effect of view angle on gender recognition performance was examined in [99, 157]. It was found that the frontal or oblique views are much more effective than a side view for gender discrimination. More recently, Pollick *et al.* [126] studied human efficiency at gender recognition from point-light walkers.

Most of the previous studies have focused on manual identification of key features that enable perceptual classification between female and male walking styles. Features related to speed, arm swing, shoulder-hip lengths, inversion, and body sway have been examined.
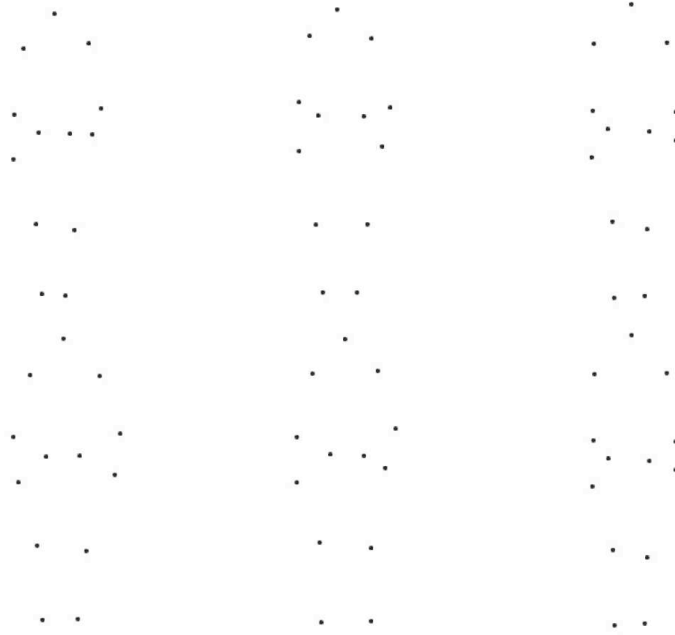
Figure 2.9: Point-light display of human walking at the front view [40]: three female walkers (*top*) and three male walkers (*bottom*).

However, to date there is no conclusive evidence as to which features actually drive the discrimination process. It seems that gender information is not a matter of a single feature, but rather involves multiple combined features. Troje [157] treated the analysis of biological motion as a linear pattern recognition problem, and presented a two-stage PCA framework for recognizing gender. The first PCA decomposed each walker's data into its Eigenspace, and a second PCA was applied to all walker Eigenspaces followed by a linear classifier. He reported the recognition rate of 92.5%. Davis and Gao [40, 41] recently presented an approach for gender recognition of point-light walkers using an expressive three-mode PCA model. Their method first constructs a PCA representation of point-light trajectories for a prototype female and male walker. A large labeled set is then used to automatically learn which trajectories in the prototype PCA representation best express the gender of the walkers. Non-expressive trajectories are removed and the remaining trajectories are weighted to bias the gender estimation method to produce the desired gender labels.

Given the ability of humans to identify gender by gaits, there have been few computer vi-
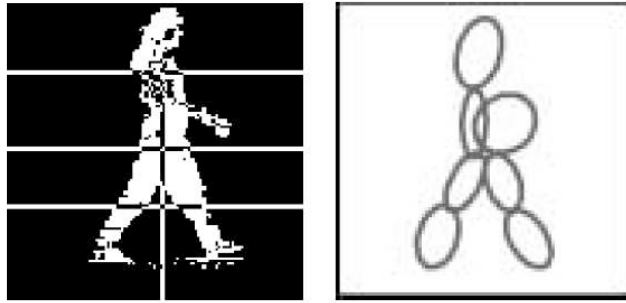
Figure 2.10: The silhouette of a foreground walking person is divided into 7 regions, and ellipses are fitted to each region [90].

sion systems developed for gait-based gender recognition. Compared to facial gender discrimination, this problem is relatively understudied, although recently some tentative attempts appeared [90, 177]. Lee and Grimson [90] extracted appearance features of gaits from image sequences for gender classification. For each scale-normalized binary silhouette, they found the centroid and divided the silhouette into 7 parts roughly corresponding to head/shoulder, arms/torso, thighs, and calves/feet (as shown in Figure 2.10), and then extracted moment-based features from each part to represent gait dynamics. Using SVMs as classifiers, their approach achieved recognition performance of 84.5% on a small data set (10 women and 14 men). More recently Yoo *et al.* [177] studied gender discrimination by gaits using a much larger database (84 males and 16 females). They used a 2D stick figure (with 8 sticks and 6 joint angles) to represent human body structure (as shown in Figure 2.11), which was extracted from body contour by determining body points. Gait features based on motion parameters were calculated from a sequence of stick figures, which were input into SVM classifiers for gender discrimination. Their system produced average recognition performance of 96%. In this thesis, we investigate gait-based gender discrimination on a larger database with more balanced female and male subjects (88 men and 31 women) by adopting a simple silhouette-based gait representation (Chapter 6).

Each modality, gait or face, in isolation has its inherent weakness and limitations for gender discrimination. For the optimal performance, computer vision systems must use as much information as possible from the observation, i.e., combining face and gait cues. Recently
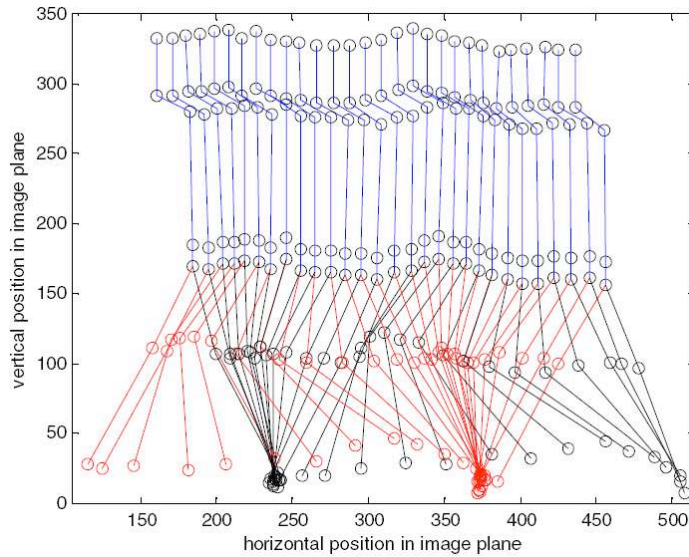
Figure 2.11: The extracted stick figures from an image sequence [177].

several attempts [138, 76, 184] have been made to integrate face and gait cues for the human identification problem. Shakhnarovich and Darrell [138] computed an image based visual hull from a set of monocular views which is then used to render virtual canonical views for frontal face recognition and side-view gait recognition. They studied different approaches to combine the two modalities in the multi-camera indoor environment. Kale *et al.* [76] presented experimental results on fusing face and gait in the single camera case. They used a specially designed database NIST, where subjects walk along an inverted Σ pattern. Recently Zhou *et al.* [183] combined cues of face profile and gait silhouette from the single camera video sequences to recognize human at a distance. In their later work [184], they further combined side face and gait cues for human identification. All these existing studies have been focused on the decision-level fusion of face and gait, while the feature-level fusion is understudied (see [98] for an overview on data fusion). This is mainly because the two modalities may have incompatible feature sets and the relationship between the different feature spaces is unknown. In this thesis, we propose to fuse face and gait cues at the feature level for improved gender discrimination.

Figure 2.12: Examples of affective body gestures (from the FABO database [61]). From *top* to *bottom*: Fear, Joy, Uncertainty, and Surprise.

2.2.2 Affective Body Language Analysis

In affective computing, a great deal of attention has been focused on how emotions are communicated through facial expressions, and a similar although smaller literature exists on the perception of emotion from voice [120]. However, little attention has been placed on visual signals extracted from body parts (see Figure 2.12 for examples of affective body gestures), although there is clear evidence that people express and interpret others' emotional and interpersonal states from bodily expression such as body movement, posture, gesture and gait [22]. Affective computing research should go beyond facial and vocal expressions and might consider issues of perception of bodily expression.

Affective body posture and gesture analysis is still an unresolved area in psychology and nonverbal communication. Coulson [34] presented experiments on attributing six universal emotions to static body postures using computer-generated mannequin figures, and his experimental results suggest that recognition from body posture is comparable to recognition from voice, and some postures are recognized as well as facial expressions. Observers tend to be accurate in decoding some negative emotions like anger and sadness from static body
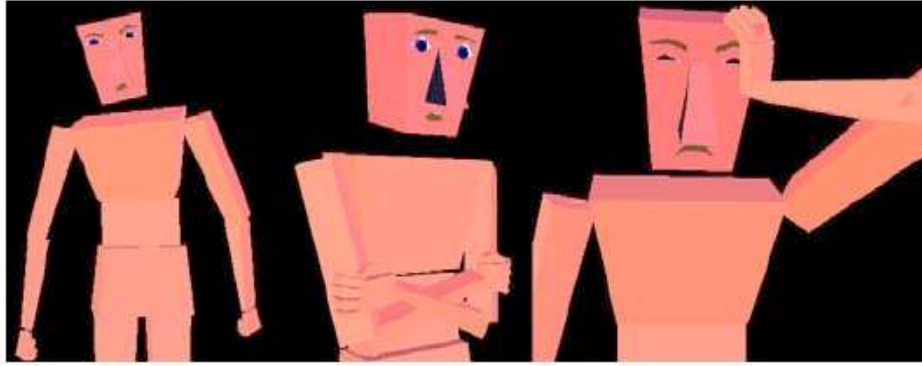
Figure 2.13: Examples of body language displayed by the virtual agent in [106]. From *left* to *right*: anger, defensiveness, and headache.

postures [34] and the gestures like head inclination and face touching often accompany affective states like shame and embarrassment [33]. Body posture involves an 3D presence which offers different percepts depending on the observer's location, and the same posture viewed from different angles does not give rise to the same percept. Statistical techniques were used in [129] to determine a set of posture features in discriminating between emotions. Neagle *et al.* [107] reported a qualitative analysis on affective motion features of virtual ballet dancers, and their results show that human observers are highly accurate in assigning an emotion label to each dance exercise.

Burgoon *et al.* [23] discussed the issue of identifying emotional states from bodily cues for human behaviour understanding. Mota and Picard [105] studied affective postures in an e-learning scenario, where the posture information was collected through a sensor chair. More recently Ravindra *et al.* [130] presented an affective gesture recognition system that recognize child's emotion with intensity through body gesture in the context of a game. Nayak and Turk [106] developed a system that allows virtual agents to express their mental state through body language, endowing them with various personality characteristics, emotions and mental attributes (see Figure 2.13 for some examples). The affective behaviours increase the believability of virtual characters, and therefore the capacity to emotionally connect and interact with humans.

In computer vision, the existing studies on gesture recognition primarily deal with non-

affective gestures such as sign language. There has been few investigations into affective body posture and gesture analysis. This is probably due to the high variability of the possible posture and gesture that can be displayed. Recently some tentative attempts have been made [7, 60]. However, there are some limitations with these studies. For example, studies were carried out on very limited data (for instance, only 27 video sequences from 4 subjects were processed in [60]). Feature extraction and representation are rather simple, for example, the neutral and expressive frames are manually selected for analysis in [60], which is unsuitable in real-world scenarios. In this thesis, we investigate visual affective body gesture analysis on a large dataset by exploiting spatial-temporal features, which makes few assumptions about the observed data, such as background and occlusion.

Human emotional and interpersonal states are not conveyed by a single indicator rather by a set of cues. It is the combination of movements of the face, arms, legs and other body parts, as well as voice and touching behaviours, that yields an overall display [6]. To more accurately simulate the human ability to assess affect, an automatic affect recognition system should make use of multimodal data. However, most of the previous work has relied on a single modality [122]. Existing work combining different modalities for affect analysis investigated mainly the combination of facial and vocal signals [120].

Recently Meeren *et al.* [100] showed that the recognition of facial expressions is strongly influenced by the concurrently presented emotional body language, and that the affective information from the face and the body start to interact rapidly, and the integration is a mandatory automatic process occurring early in the processing stream. Indeed, in social interpersonal communication, human face is usually perceived not as an isolated object but as an integrated part of a whole body. Therefore, fusing facial expression and body gesture in video sequences provides a potential way to accomplish improved affect analysis. However, there is little effort on visual human affect analysis by combining face and body gestures [122]. Kapoor and Picard [79] presented a multi-sensor affect recognition system for classifying interest (or disinterest) in children trying to solve a puzzle on the computer. The multimodal information from face expressions and postures are sensed through a camera and a chair respectively, which are combined with the state of the puzzle. Their approach

generates separate class labels corresponding to each individual modality, which are probabilistically combined for final classification. The multimodal method is shown to outperform classification using the individual modalities. Recently Balomemos *et al.* [7] and Gunes and Piccardi [60] made tentative attempt to analyze emotions from user facial expressions and body gestures on very small datasets. Most of these studies fuse face and body cues at the decision level. However, to accomplish a human-like analysis of multiple input signals, the signals cannot be considered mutually independently and combined at the end of the intended analysis but, on the contrary, the input data should be processed in a joint feature space [73]. In this thesis, we exploit statistical techniques to fuse the two modalities at the feature level by deriving a semantic joint feature space.

2.3 Summary

Although vision-based facial expression and body language analysis has been extensively addressed for decades, there still exist many limitations. In this thesis, we mainly address the following challenging problems:

- **Facial feature representation and selection**: In the existing work, Gabor-wavelet features have been widely adopted to represent appearance changes of faces, showing superior performance to geometric features. However, it is computational expensive to extract Gabor features. It is crucial to identify a low-computation discriminative feature space for expression analysis. To this end, we comprehensive investigate appearance features based on Local Binary Patterns. Extensive experiments illustrate that LBP features are effective and efficient for facial expression discrimination, and capable of robust performance over a rang of image resolutions. We further investigate feature selection methods to derive the most effective LBP features for better facial representation.

- **Manifold analysis of facial expression**: Facial dynamics is an important factor in interpreting facial expression precisely, and one way to capture explicitly expression dynamics is to map expression images to low dimensional manifolds. As stated in Section

2.1.2, there are several noticeable limitations in the existing work on expression manifold learning. To address these problems, we exploit Locality Preserving Projections to learn the expression manifold in a dense appearance feature space. We also present to align manifolds of different subjects on a generalized expression manifold. A Bayesian framework is also formulated for person-independent dynamic expression recognition employing the derived manifold representation. More crucially, our approaches are evaluated with a large number of subjects.

- **Correlation analysis of facial parts**: Facial muscles are contracted in unison to display natural facial expressions, so different facial regions have strong spatio-temporal correlations. Most of existing methods treats human face as a whole, ignoring these correlations. In this work, we employ Canonical Correlation Analysis to model correlations among facial parts. To address the limitations of classical CCA for image data, we introduce a Matrix-based Canonical Correlation Analysis for better correlation analysis of 2D image or matrix data in general.

- **Multimodal facial and body language analysis**: We investigate two important problems that are understudied in the existing work: gender discrimination by gait and affective body gesture analysis. By adopting a simple silhouette-based gait representation, we study gait-based gender discrimination on a large database. On affective gesture analysis, spatial-temporal features are exploited and evaluated on a large dataset. Considering each modality, face or body, in isolation has its inherent weakness and limitations, we further present to fuse the two modalities at the feature level by deriving semantic joint feature spaces for improved gender discrimination and affect analysis.

3 Feature Selection and Representation

Deriving an effective feature representation from original images is a vital step for successful facial and body language understanding. This includes identifying an appropriate feature representation scheme and selecting the most discriminative features. With regard to facial representation, as reviewed in Chapter 2, geometric feature-based facial representations require accurate and robust facial feature detection and tracking, and usually cannot encode changes in skin texture that are critical for facial expression modeling. In contrast, appearance features, especially Gabor-wavelet features, have shown superior performance in facial representation. However, it is both time and memory inefficient to convolve face images with a bank of Gabor filters to extract features.

In this chapter, we study facial representation based on statistical local features, in particular, Local Binary Patterns (LBP). Compared to Gabor wavelets, LBP features can be extracted faster in a single scan through the raw image and lie in a much lower dimensional space, whilst still retaining facial information by representing salient micro-patterns. With LBP features, we examine different machine learning methods for person-independent facial expression recognition on several public databases, including low-resolution settings. We also investigate feature selection methods to derive the most effective LBP features for better facial representation. In addition to utilizing AdaBoost [55], we present and evaluate a Conditional Mutual Information based learning procedure, which allows much faster training. Our experiments show that the best recognition performance is obtained by using SVM classifiers with the selected LBP features.

3.1 Local Binary Patterns

The original LBP operator was introduced by Ojala *et al.* [111], and has been proved a powerful means of texture description. The operator labels the pixels of an image by thresholding a 3×3 neighborhood of each pixel with the center value and considering the results as a binary number (see Figure 3.1 for an illustration), and the 256-bin histogram of LBP labels computed over a region is used as a texture descriptor. The derived binary numbers (called local binary patterns or LBP codes) codify local primitives including different types of curved edges, spots, flat areas, etc. (as shown in Figure 3.2), so each LBP code can be regarded as a micro-texton [63].
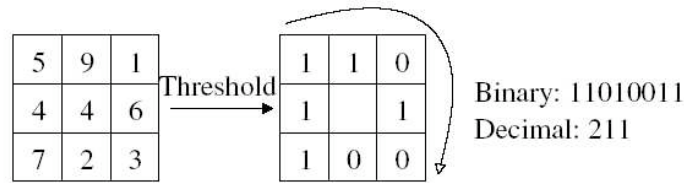
Figure 3.1: The basic LBP operator [1].
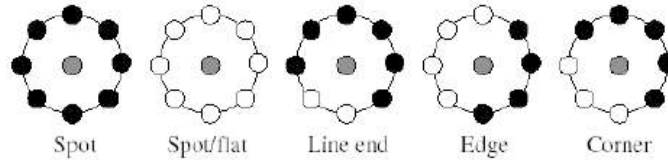
Figure 3.2: Examples of texture primitives which can be detected by LBP (white circles represent ones and black circles zeros) [63].

The limitation of the basic LBP operator is its small 3×3 neighborhood which can not capture dominant features with large scale structures. Hence the operator was later extended to use neighborhoods of different sizes [112]. Using circular neighborhoods and bilinearly interpolating the pixel values allow any radius and number of pixels in the neighborhood. Figure 3.3 shows examples of the extended LBP operator, where the notation (P, R) denotes a neighborhood of P equally spaced sampling points on a circle of radius of R that form a circularly symmetric neighbor set.

The LBP operator $LBP_{P,R}$ produces 2^P different output values, corresponding to the 2^P

(P=8,R=1.0)   (P=12,R=1.5)   (P=16,R=2.0)
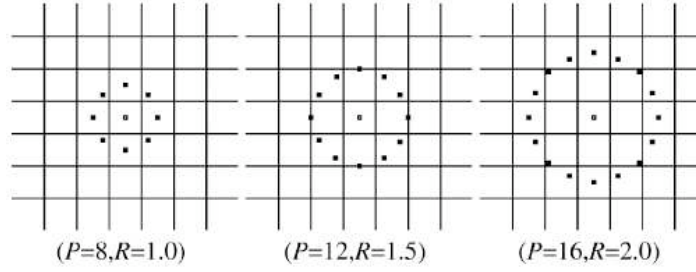
Figure 3.3: Three examples of the extended LBP [112]: the circular $(8,1)$ neighborhood, the circular $(12, 1.5)$ neighborhood, and the circular $(16, 2)$ neighborhood respectively.

different binary patterns that can be formed by the P pixels in the neighbor set. It has been shown that certain bins contain more information than others [112]. Therefore, it is possible to use only a subset of the 2^P local binary patterns to describe the texture of images. Ojala *et al.* [112] called these fundamental patterns as uniform patterns. A local binary pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For example, 00000000, 001110000 and 11100001 are uniform patterns. It is observed that uniform patterns account for nearly 90% of all patterns in the $(8,1)$ neighborhood and for about 70% in the $(16, 2)$ neighborhood in texture images [112]. Accumulating the patterns which have more than 2 transitions into a single bin yields an LBP operator, denoted $LBP_{P,R}^{u2}$, with less than 2^P bins. For example, the number of labels for a neighborhood of 8 pixels is 256 for the standard LBP but 59 for LBP^{u2}.

After labeling a image with an LBP operator, a histogram of the labeled image $f_l(x,y)$ can be defined as

$$H_i = \sum_{x,y} I(f_l(x,y) = i), \qquad i = 0, \ldots, n-1 \tag{3.1}$$

where n is the number of different labels produced by the LBP operator and

$$I(A) = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases} \tag{3.2}$$

This LBP histogram contains information about the distribution of the local micro-patterns, such as edges, spots and flat areas, over the whole image, so can be used to statistically

describe image characteristics.

Each face image can be seen as a composition of micro-patterns which can be effectively detected by the LBP operator. Therefore, it is intuitive to use LBP features to represent face images [1, 63]. A LBP histogram computed over the whole face image encodes only the occurrences of the micro-patterns without any indication about their locations. To also consider the shape information of faces, face images can be equally divided into m small regions $R_0, R_1, \ldots, R_m$ to extract LBP histograms [1] (as shown in Figure 3.4). The LBP features extracted from each sub-region are then concatenated into a single, spatially enhanced feature histogram defined as:

$$H_{i,j} = \sum_{x,y} I\big(f_l(x,y) = i\big) I\big((x,y) \in R_j\big) \tag{3.3}$$

where $i = 0, \ldots, n-1, j = 0, \ldots, m-1$.
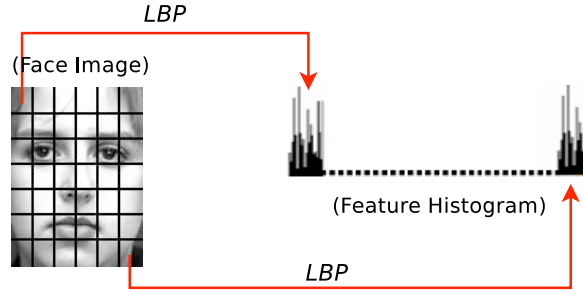
Figure 3.4: A face image is divided into small regions from which LBP histograms are extracted and concatenated into a single, spatially enhanced feature histogram.

The extracted feature histogram represents the local texture and global shape of face images. In this histogram, the face is described on three different levels of locality: the labels for the histogram contain the pixel-level patterns, the labels are summed over a small region to produce information on a regional level and the regional histograms are concatenated to build a global description of the face [1]. We call the above LBP features Uniform-LBP, in contrast to boosted LBP features discussed in Section 3.5.

Figure 3.5: Sample facial expression images from the Cohn-Kanade database.

3.2 Facial Expression Recognition Using Uniform-LBP

We study person-independent facial expression recognition using Uniform-LBP features. Different machine learning techniques, including template matching, SVM, and the linear programming technique, are examined for the task of recognizing expressions.

3.2.1 Facial Expression Data

Following most of recent work on facial expression analysis [97, 28, 176, 95], we conduct experiments on prototypic emotional expression recognition. We mainly utilize the Cohn-Kanade database [78], one of the most comprehensive databases in the current facial-expression-research community. The database consists of 100 university students aged from 18 to 30 years, of which 65% were female, 15% were African-American, and 3% were Asian or Latino. Subjects were instructed to perform a series of 23 facial displays, six of which were based on description of prototypic emotions. Image sequences from neutral to target display were digitized into 640×490 pixel arrays with 8-bit precision for grayscale values. Figure 3.5 shows some sample images from the Cohn-Kanade database.

For our experiments, we selected 320 image sequences from the database. The only selection criterion was that a sequence could be labeled as one of the six basic emotions. The sequences come from 96 subjects, with 1 to 6 emotions per subject. For each sequence, the neutral face and three peak frames were used for prototypic emotional expression recognition. To evaluate the generalization performance to novel subjects, we adopted a 10-fold cross-validation testing scheme in our experiments. More precisely, we partitioned the data set randomly into ten groups of roughly equal numbers of subjects. Nine groups were used as the training data to train classifiers, while the remaining group was used as the test data. The above process was repeated ten times for each group in turn to be omitted from the training process. We report the average recognition results on the test data.

Following Tian [152], we normalized the faces to a fixed distance between the two eyes. We manually labeled the eyes' location, to evaluate LBP features in the condition of no face registration errors. Automatic face registration can be achieved by face detection [165] and eye localization [152], which will be addressed in our future work. Facial images of 110×150 pixels were cropped from original frames based on the two eyes location. No further registration such as alignment of mouth [182] was performed in our approaches. As the faces in the database are at a frontal view, we did not consider head pose changes. For realistic sequences with head pose variation, head pose estimation [152] needs to be adopted to detect faces of frontal or near frontal view. Illumination changes exist in the database, but we made no attempt to remove illumination changes [152] in our experiments, due to LBP's gray-scale invariance. Figure 3.6 shows an example of the original face image and the cropped image.

Some parameters can be optimized for better Uniform-LBP feature extraction. One is the LBP operator, and the other is the number of regions divided. Following the setting in [1], we selected the 59-bin $LBP_{8,2}^{u2}$ operator, and divided 110×150 pixels face images into 18×21 pixels regions, giving a good trade-off between recognition performance and feature vector length. Thus face images were divided into $42(6\times7)$ regions as shown in Figure 3.7, and represented by LBP histograms with the length of $2,478(59\times42)$.

Figure 3.6: The original face image and the cropped image.

3.2.2 Template Matching

To evaluate the effectiveness of LBP based representation, we first adopt a template matching approach [1] to classify facial expressions for its simplicity. In training, LBP histograms of expression images in a given class are averaged to generate a template for this class. In testing, the input image is matched to the closest template using a nearest-neighbor classifier.
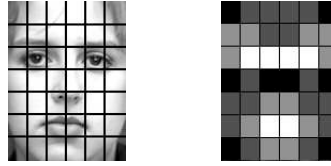
Figure 3.7: (Left) A face image divided into 6×7 sub-regions. (Right) A weights set for weighted dissimilarity measure. Black squares indicate weight 0.0, dark gray 1.0, light gray 2.0 and white 4.0.

Ahonen *et al.* [1] compared different dissimilarity measures for LBP histograms, and found the Chi square statistic (χ^2) is most effective. So we select the Chi square statistic as the dissimilarity measure here:

$$\chi^2(\mathbf{S}, \mathbf{M}) = \sum_i \frac{(S_i - M_i)^2}{S_i + M_i} \qquad (3.4)$$

where $\mathbf{S}$ and $\mathbf{M}$ are two LBP histograms. It is observed that facial features contributing to facial expressions mainly lie in regions such as eye and mouth regions, and these regions contain more discriminative information for expression classification. Therefore, a weight can be set for each region based on its importance, as shown in Figure 3.7. This particular weight

set is designed empirically based on the observation. The weighted χ^2 statistic is then given as

$$\chi_w^2(\mathbf{S}, \mathbf{M}) = \sum_{i,j} w_j \frac{(S_{i,j} - M_{i,j})^2}{S_{i,j} + M_{i,j}} \tag{3.5}$$

where w_j is the weight for region j.

Our template matching achieved a generalization performance of 84.5% for the 6-class basic emotional expression recognition and 79.1% for the 7-class expression recognition (6 emotions plus neutral face). We compared the results with those reported in [28], where Cohen *et al.* adopted Bayesian network classifiers to classify 7-class emotional expressions based on the tracked geometric facial features (eyebrows, eyelids, and mouth). They carried out 5-fold cross-validation on a subset of 53 subjects from the Cohn-Kanade database, and obtained the best performance of 73.2% by using Tree-Augmented-Naive Bayes (TAN) classifiers. Although it is impossible to have a direct comparison due to different experimental setups, comparison in Table 3.1 indicates that our simple template matching using LBP features provides slightly better overall performance. The confusion matrix of 7-class recognition of our method is shown in Table 3.2. We can observe that Joy and Surprise can be recognized with high accuracy (around 90-92%), while Anger and Fear are easily confused with others.

Methods (Feature + Classifier)	7-Class Recognition	6-Class Recognition
LBP Features + Template Matching	79.1%	84.5%
Geometric Features + TAN [28]	73.2%	-

Table 3.1: Comparisons between the geometric features based TAN [28] and our LBP-based template matching.

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Anger	**58.7%**	5.5%	0	0	26.7%	0	9.1%
Disgust	3.3%	**85.0%**	2.5%	0	2.5%	0	6.7%
Fear	1.0%	0	**61.7%**	24.0%	10.3%	0	3.0%
Joy	0	0	6.0%	**90.4%**	0	0	3.6%
Sadness	4.9%	0	0	0	**72.4%**	1.7%	21.0%
Surprise	0	0	1.3%	0	2.7%	**92.4%**	3.6%
Neutral	2.0%	0.8%	0.4%	0.8%	25.7%	0	**70.3%**

Table 3.2: Confusion matrix of 7-class facial expression recognition using template matching with LBP features.

46

3.2.3 Support Vector Machine

A popular technique to facial expression classification is SVM [12, 10, 162, 159]. SVM is an optimal discriminant method based on the Bayesian learning theory. For the cases where it is difficult to estimate the density model in high-dimensional space, the discriminant approach is preferable to the generative approach. SVM [163] performs an implicit mapping of data into a higher dimensional feature space, and then finds a linear separating hyperplane with the maximal margin to separate data in this higher dimensional space.

Given a training set of labeled examples $\{(x_i, y_i), i = 1, \ldots, l\}$ where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, a new test example x is classified by the following function:

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b) \qquad (3.6)$$

where α_i are Lagrange multipliers of a dual optimization problem that describe the separating hyperplane, $K(\cdot, \cdot)$ is a kernel function, and b is a threshold parameter of the hyperplane. The training sample x_i with $\alpha_i > 0$ is called *support vectors*, and SVM finds the hyperplane that maximizes the distance between the support vectors and the hyperplane. Given a non-linear mapping Φ that embeds the input data into the high dimensional space, kernels have the form of $K(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle$. SVM allows domain-specific selection of the kernel function. Though new kernels are being proposed, the most frequently used are the linear, polynomial, and Radial Basis Function (RBF) kernels.

SVM makes binary decisions, so the multi-class classification here is accomplished by using the one-against-rest technique, which trains binary classifiers to discriminate one expression from all others, and outputs the class with the largest positive output of binary classification. With regard to the parameter selection of SVM, as suggested in [71], we carried out a grid-search on the hyper-parameters in the 10-fold cross-validation. The parameter setting producing best cross-validation accuracy was picked. Each dimension of the training and testing vector was scaled to be between -1 and 1. We used the SVM implementation in the publicly available machine learning library SPIDER[1] in our experiments. The generalization

---

[1] *http://www.kyb.tuebingen.mpg.de/bs/people/spider/index.html*

performances achieved with different kernels are shown in Table 3.3, where the degree of the polynomial kernel is 1, and the standard deviation for the RBF kernel is 2^{13} for 7-class recognition and 2^{11} for 6-class recognition. The confusion matrices of 6-class and 7-class recognition with the RBF kernel are shown in Table 3.4 and Table 3.5. It is observed that, Disgust, Joy, Surprise, and Neutral can be recognized with high accuracy (90-98%), while the recognition rates for Fear and Sadness are much lower (68-69%). Compared to the recognition results of template matching in Table 3.2, the recognition performance for every expression is increased except Sadness. For the 6-class problem, the number of support vectors of the linear/polynomial SVMs were 18-29 percents of the total number of training samples, while the RBF SVMs employed 18-31 percents. For the 7-class problem, the linear/polynomial SVMs employed 15-30 percents, while the RBF SVMs employed 16-35 percents.

	6-Class Recognition	7-Class Recognition
SVM (Linear)	91.5%	88.1%
SVM (Polynomial)	91.5%	88.1%
SVM (RBF)	92.6%	88.9%

Table 3.3: Recognition performance of LBP-based SVM with different kernels.

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger	**89.7%**	2.7%	0	0	7.6%	0
Disgust	0	**97.5%**	2.5%	0	0	0
Fear	0	2.0%	**73.0%**	22.0%	3.0%	0
Joy	0	0.4%	0.7%	**97.9%**	1.0%	0
Sadness	10.3%	0	0.8%	0.8%	**83.5%**	4.6%
Surprise	0	0	1.3%	0	0	**98.7%**

Table 3.4: Confusion matrix of 6-class facial expression recognition using SVM (RBF).

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Anger	**85.0%**	2.7%	0	0	4.8%	0	7.5%
Disgust	0	**97.5%**	2.5%	0	0	0	0
Fear	0	2.0%	**68.0%**	22.0%	1.0%	0	7.0%
Joy	0	0	0.7%	**94.7%**	1.1%	0	3.5%
Sadness	8.6%	0	0	0	**69.5%**	2.3%	19.6%
Surprise	0	0	1.3%	0	0	**98.2%**	0.5%
Neutral	1.6%	0.4%	0	1.6%	6.0%	0.4%	**90.0%**

Table 3.5: Confusion matrix of 7-class facial expression recognition using SVM (RBF).

We further compare LBP features with Gabor-wavelet features for facial expression recog-

nition using SVMs. Following Bartlett *et al.* [12, 10], we converted face images into a Gabor magnitude representation using a bank of Gabor filters at 8 orientations and 5 spatial frequencies (9:36 pixels per cycle at 1/2 octave steps). To reduce the length of the feature vector, the outputs of the 40 Gabor filters were downsampled by a factor of 16 [44], so the dimensionality of the Gabor feature vector is $42,650(40 \times 110/4 \times 150/4)$. We show the generalization performance of Gabor-wavelet features in Table 3.6. It is evident that LBP features consistently outperform Gabor features when using SVM classifiers.

Bartlett *et al.* [12, 10] recently conducted similar experiments using the Gabor-wavelet representation with SVMs on the Cohn-Kanade database. They selected 313 image sequences from the database, which came from 90 subjects, with 1 to 6 emotions per subject. The facial images were converted into a Gabor magnitude representation using a bank of 40 Gabor filters. They divided the subjects randomly into ten groups of roughly equal size and did "leave one group out" cross-validation [12]. SVMs with linear, polynomial, and RBF kernels were used to classify 7-class expressions. Linear and RBF kernels performed best, achieving recognition rates of 84.8% and 86.9% respectively. We also include their recognition results in Table 3.6, although they are obtained with different experimental setups. In their more recent work [10], they reported 88.0% (Linear) and 89.1% (RBF) in Leave-one-subject-out experiments.

	6-Class		7-Class		
	LBP	Gabor	LBP	Gabor	Gabor [12]
SVM (Linear)	91.5%	89.4%	88.1%	86.6%	84.8%
SVM (Polynomial)	91.5%	89.4%	88.1%	86.6%	worse than RBF/linear
SVM (RBF)	92.6%	89.8%	88.9%	86.8%	86.9%

Table 3.6: Comparisons between LBP features with Gabor-filter features for facial expression recognition using SVMs.

	LBP	Gabor	Gabor [12]
Memory (Feature Dimension)	2,478	42,650	92,160
Time (Feature Extraction Time)	0.03s	30s	-

Table 3.7: Time and memory costs for extracting LBP features and Gabor-filter features.

Comparisons summarized in Table 3.6 show that the LBP-based SVMs perform slightly better than the Gabor-wavelet based SVMs. More crucially though, the advantage of LBP features lies at that the simplicity of LBP features allows very fast feature extraction, without

complex analysis in extracting a large set of Gabor features. We compare the time and memory costs of the feature extraction process (Matlab implementation) between LBP features with Gabor features in Table 3.7, where the Gabor-filter convolutions were calculated in spatial domain. It is observed that LBP features bring significant speed benefit, and, compared to the high dimensionality ($O(10^5)$) of the Gabor features, LBP features lie in a much lower dimensional space, reducing the memory space by a order of 17.

3.2.4 Linear Programming

Guo and Dyer [62] adopted a linear programming technique to perform simultaneous feature selection and classifier training for facial expression recognition. Similarly Feng *et al.* [53] recently presented an approach that uses LBP features with a linear programming technique for facial expression recognition. However, their studies were carried out on a very small database (JAFFE). Here we examine LBP features using the linear programming technique on a large dataset.

Given two sets of data samples $\mathcal{A}$ and $\mathcal{B}$ in $\mathbb{R}^n$, we seek a linear function such that $f(x) > 0$ if $x \in \mathcal{A}$, and $f(x) \leq 0$ if $x \in \mathcal{B}$. This function is given by $f(x) = \mathbf{w}^T x - \gamma$, and determine a plane $\mathbf{w}^T x = \gamma$ with normal $\mathbf{w} \in \mathbb{R}^n$ that separates $\mathcal{A}$ from $\mathcal{B}$. Let the set of m samples in $\mathcal{A}$ be represented by a matrix $A \in \mathbb{R}^{m \times n}$ and the set of k samples in $\mathcal{B}$ be represented by a matrix $B \in \mathbb{R}^{k \times n}$. After normalization, we want to satisfy

$$A\mathbf{w} \geq e\gamma + e, \qquad B\mathbf{w} \leq e\gamma - e \tag{3.7}$$

where e is a vector of all 1s with appropriate dimension. Practically, because of the overlap between the two classes, one has to minimize some norm of the average error in Eqn. (3.7) [62]:

$$\min_{\mathbf{w},\gamma} f(\mathbf{w}, \gamma) = \min_{\mathbf{w},\gamma} \left\{ \frac{1}{m} \|(-A\mathbf{w} + e\gamma + e)_+\|_1 + \frac{1}{k} \|(B\mathbf{w} - e\gamma + e)_+\|_1 \right\} \tag{3.8}$$

where x_+ denotes the vector with components satisfying $(x_+)_i = \max\{x_i, 0\}, i = 1, \ldots, n$, and $\| \cdot \|_1$ denotes the 1-norm. Eqn. (3.8) can be modeled as a robust linear programming

problem [62]:

$$\min_{w,\gamma,y,z} \quad \frac{e^T y}{m} + \frac{e^T z}{k} \tag{3.9}$$

$$\text{subject to} \quad \begin{cases} -A\mathbf{w} + e\gamma + e \leq y, \\ B\mathbf{w} - e\gamma + e \leq z, \\ y \geq 0, z \geq 0 \end{cases}$$

which minimizes the average sum of misclassification errors. We use Eqn. (3.9) to solve the classification problem.

Following Feng *et al.* [53], multi-class facial expression recognition was decomposed into one-to-one pairs of binary classification, where each binary classifier was produced by the linear programming technique. Binary classifiers were combined with a voting scheme to output the final recognition result. To reduce the length of the LBP feature vector, we also discarded the dimensions whose occurrence frequency is lower than a threshold [53].

In our 10-fold cross-validation experiments, the linear programming technique produces a generalization performance of 82.3% for 7-class recognition and 89.6% for 6-class recognition. Compared with that of SVM (linear) as shown in Table 3.8, the linear programming technique produces inferior performance to SVM (linear). This indicates that, in the input expression image space, it is hard for the linear decision surface to discriminate expression with high confidence, as expression images contain complex variations and significant overlapping among different classes. In contrast, SVM is more effective when the class distributions are not Gaussian, so SVM may be better suited to expression classification.

	7-Class Recognition	6-Class Recognition
Linear Programming	82.3%	89.6%
SVM (Linear)	88.1%	91.5%

Table 3.8: Comparison between the linear programming technique and SVM (linear) for facial expression recognition.

3.3 Low-Resolution Facial Expression Recognition

In real-world environments such as smart meeting and visual surveillance, only low-resolution video input is available. Figure 3.8 shows a real-world image recorded in a smart meeting scenario. How to derive a discriminative facial representation from low-resolution images is a critical problem for real-world applications. In this section, we investigate LBP features for low-resolution facial expression recognition. We first examine LBP features on different image resolutions, then perform experiments on real-world compressed low-resolution video sequences.
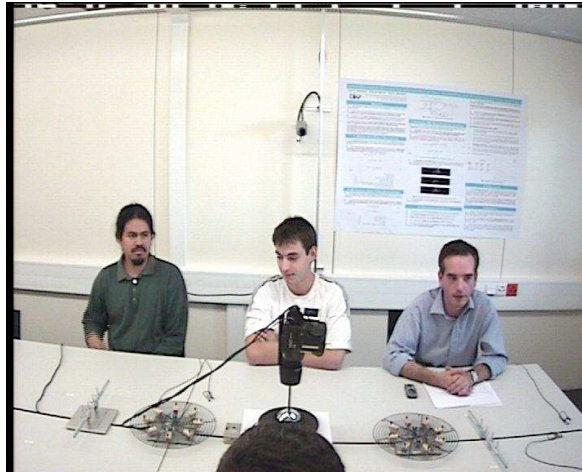
Figure 3.8: An example of low-resolution facial expressions recorded in real-world environments. (from PETS 2003 data set)

3.3.1 Evaluation on Different Resolutions

As shown in Table 3.9, in total six different resolutions of the face region were studied (110×150, 55×75, 36×48, 27×37, 18×24, and 14×19 pixels) based on the Cohn-Kanade database. The lower resolution images were down-sampled from the original images. For LBP feature extraction, lower resolution face images were divided into 10×10 pixels regions (which may overlap with each other in the small face images). We adopted a 4-neighborhood LBP operator $LBP_{4,1}$ for each sub-region.

We conducted experiments on 6-class basic expression recognition using SVM with a RBF

	110×150	55×75	36×48	27×37	18×24	14×19
LBP	92.6%	89.9%	87.3%	84.3%	79.6%	76.9%
Gabor	89.8%	89.2%	86.4%	83.0%	78.2%	75.1%
Gabor [152]	92.2%	91.6%	-	77.6%	-	68.2%
Geometric features (tracking) [152]	91.8%	91.6%	-	N/A	-	N/A
Geometric features (detection) [152]	73.8%	72.9%	-	61.3%	-	N/A

Table 3.9: Recognition performance in low-resolution images with different methods.

kernel. We show the recognition results in Table 3.9, where the standard deviation of RBF kernels were 2^{11}, 2^9, 2^7, 2^8, 2^6 and 2^8 respectively. Besides LBP features, we also carried out experiments with the Gabor-magnitude representation by convolving images with a bank of 40 Gabor filters at 8 orientations and 5 spatial frequencies. The generalization performances of the Gabor-wavelet representation are also shown in Table 3.9.

Recently Tian also evaluated the effects of different image resolutions for facial expression analysis. In her experiments, 375 image sequences were selected from the Cohn-Kanade database for 6-class expression classification. Tian extracted two types of facial features: geometric features and appearance features. Geometric features were derived by feature tracking [154] and feature detection [153] respectively. For appearance features, a bank of 40 Gabor filters were applied to the difference images to extract facial appearance changes, where the difference images were obtained by subtracting a neutral expression for each image. A three-layer Neural Network was adopted to recognize expressions. Recognition results of Tian's methods are also summarized in Table 3.9[2], although we cannot have a direct comparison due to different experimental setups and classifiers.

We can draw the following conclusions from the experimental results shown in Table 3.9: (1) Geometric features are not available for lower resolutions, while appearance features such as Gabor wavelets and LBP features can be extracted from lower resolutions. It is difficult to detect or track facial components such as mouth, eyes, brows and nose in lower resolution images, so geometric features are not reliable in low-resolution images. On the contrary,

---

[2]In [152], different resolutions of the head region were 144×192, 72×96, 36×48, 18×24 pixels, which are comparable to the resolutions of the face region 110×150, 55×75, 27×37, 14×17 pixels in our experiments.

appearance features are robust in presenting appearance changes of faces such as wrinkles and furrows in lower resolutions. (2) The LBP features perform slightly better than the Gabor-wavelet representation on low-resolution expression recognition. Recently Liao *et al.* [93] also compared LBP features with Gabor-filter features on the JAFFE database, and their experiments demonstrated that LBP features provide better performance for low-resolution face images, which reinforces our finding. (3) The LBP features perform robustly and stably over a useful range of low resolutions. This reinforce the superiority of LBP features for face detection and recognition in low-resolution images reported in [63]. So LBP features are very promising for real-world applications where only low-resolution video input is available.

3.3.2 Evaluation on Real-world Video Sequences

We further conducted experiments on compressed low-resolution image sequences recorded in a real environment. We used the smart meeting data set (scenario A, camera 1) in the PETS 2003 evaluation data sets[3]. In this scenario, each person enters a conference room one after the other, goes to his place, presents himself to the frontal camera, and sits down. Then each person looks at the other people with different expressions. Figure 3.8 shows an example frame in the video sequence. Three facial expressions, neutral, anger and joy, are available in the data set.

Real-world video sequences typically contain full range of head motion. In Tian's previous work [153], the head pose was first estimated based on the detected head, and then for frontal and near frontal views of the face, facial features were extracted to perform expression recognition. Since our focus was investigating the validity of LBP features in compressed video inputs, we did not consider pose estimation. We cropped the face region in frontal and near frontal views based on the location of two eyes from the input image sequence, then performed recognition on the cropped face images. Figure 3.9 shows face regions cropped in an example frame.

It is very difficult, even for humans, to recognize facial expressions at low resolution. Following Tian *et al.* [153], we conducted experiments on showing frames of facial expressions

---

[3]*http://www.cvg.cs.rdg.ac.uk/PETS-ICVS/pets-icvs-db.html*
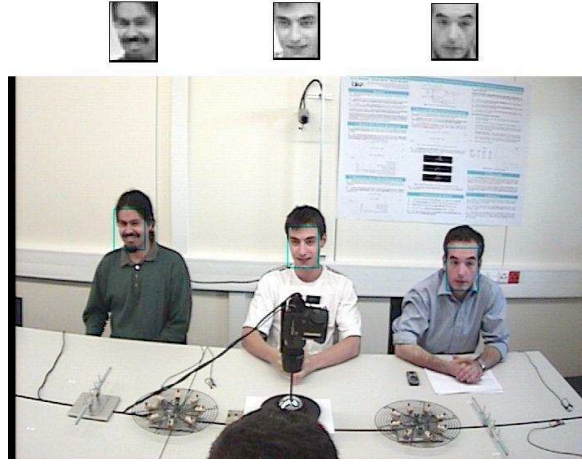
Figure 3.9: We cropped the face region in frontal and near frontal view based on the location of two eyes from the input image sequence (Frame 17130).

at low resolution to a small set of human observers (in this instance five researchers in our lab) resulting in many who could not perform recognition against the ground truth provided by the PETS data set (original GT). Tian *et al.* modified the ground truth based on the majority. Here we also generated a new ground truth (modified GT) for some frames based on human observations. Examples of modified GT vs original GT are shown in Table 3.10.

Original GT	Neutral	Joy	Neutral	Neutral
Modified GT	Sideview	Sideview	Joy	Joy

Table 3.10: Examples of modified GT vs original GT.

A total of 1209 images from the Cohn-Kanade database were used to train a SVM classifier. Since face regions in PETS data set are around 40×50 pixels, the training images were downsampled from the original images to 38×48 pixels. Our trained classifier recognized five expressions: neutral, joy, angry, surprise, and others (including fear, sadness, and disgust).

Modified GT	Joy	Joy	Neutral	Neutral
Test Results	Others	Others	Joy	Others

Table 3.11: Examples of failed recognition.

Our method performed well with these real-world image sequences. The overall recognition rate on frames from 18000 to 18190 was 91.5%, which is comparable to results reported in Tian's work [153]. Table 3.11 shows some failed examples. We observed that some frames of near frontal view were incorrectly classified because our training data includes only frontal view expressions. Additionally, the training images were captured when subjects exaggerating their facial expressions, whilst the test images were natural facial expressions without any deliberate exaggerated posing. This difference in data also brings some classification errors.

3.4 LBP Feature Selection

In the above Uniform-LBP feature extraction, face images are equally divided into small sub-regions from which LBP histograms are extracted and concatenated into a single feature vector. The effectiveness of these uniform-LBP features depends heavily on how the sub-regions were divided, and this LBP feature extraction scheme suffers from fixed sub-region size and positions. By shifting and scaling a sub-window over face images, many more sub-regions can be obtained, bringing many more LBP histograms, which yield a more complete description of face images. To minimize a very large number of LBP histograms necessarily introduced by shifting and scaling a sub-window, feature selection techniques such as AdaBoost [136] can be exploited to learn those LBP histograms that containing the most discriminative information. Zhang *et al.* [180] presented an approach for face recognition by boosting LBP-based classifiers, where the distance between corresponding LBP histograms of two face images was used as a discriminative feature, and AdaBoost was used to learn a few of most efficient features. Here we present to learn the most discriminative LBP histograms (or regions) for facial expression recognition.

3.4.1 AdaBoost

AdaBoost, introduced by Freund and Schapire [55, 136], provides a simple yet effective approach for stagewise learning of a nonlinear classification function. AdaBoost learns a small number of weak classifiers whose performance are just better than random guessing, and

- Given the instance set $S = \{(x_1, y_1), \ldots, (x_m, y_n)\}$ where $y_i = 0, 1$ for negative and positive examples respectively, and the size of the final strong classifier T.

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

- For $t = 1, \ldots, T$

  1. Normalize the weights
  $$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \qquad (3.10)$$
  so that w_t is a probability distribution.

  2. For each feature, j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to $w_t, \epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$.

  3. Choose the classifier, h_t, with the lowest error ϵ_t.

  4. Update the weights:
  $$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i} \qquad (3.11)$$
  where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$.

- Output the final strong classifier:
$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases} \qquad (3.12)$$

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

Figure 3.10: The AdaBoost algorithm [165].

boosts them iteratively into a strong classifier of higher accuracy. The process of AdaBoost maintains a distribution on the training samples. At each iteration, a weak classifier which minimizes the weighted error rate is selected, and the distribution is updated to increase the weights of the misclassified samples and reduce the importance of the others. AdaBoost has been used widely for face detection [165] and face recognition [75].

The AdaBoost algorithm is shown in Figure 3.10. On selecting a weak classifier, we adopt the simple template matching described in Section 3.2.2. In training, LBP histograms in a given class are averaged to generate a template for this class. In recognition, a nearest-neighbor classifier is adopted to match the input histogram with the closest template. We also

used the Chi square statistic (χ^2) as the dissimilarity measure for histograms (Eqn. (3.4)).

Recently Li and Zhang [91] have shown that a strong classifier learned by AdaBoost is suboptimal in terms of the error rate, and proposed FloatBoost for improved performance. FloatBoost incorporates Floating Search into AdaBoost, using a backtrack mechanism after each iteration of AdaBoost to delete the weak classifiers that are ineffective in reducing error rate. Compared to AdaBoost, FloatBoost leads to a strong classifier consisting of fewer weak classifiers, while improving the classification performance. However, as FloatBoost removes unfavorable weak classifiers after each iteration, this increases training time massively. FloatBoost costs five times longer training time than that of AdaBoost [91]. AdaBoost itself usually requires very expensive training time already. To address this problem, we present in the following an efficient learning method that both avoids selecting ineffective weak classifiers in each iteration of learning and allows for very fast training.

3.4.2 Conditional Mutual Information based Boosting

CMI based Feature Selection

Mutual Information is a basic concept in information theory [35]. It estimates the quantity of information shared between random variables. For two random variables U and V, their mutual information $I(U;V)$ is defined as follows:

$$I(U;V) = H(U) - H(U|V) = H(V) - H(V|U) \tag{3.13}$$

where $H(\cdot)$ is the entropy of a random variable. Entropy $H(U)$ quantifies the uncertainty of U. For a discrete random variable U, $H(U)$ is defined as

$$H(U) = -\sum_{u \in U} p(u) \log p(u) \tag{3.14}$$

Here $p(u)$ represents the marginal probability distribution of U. The conditional entropy $H(U|V)$ quantifies the remaining uncertainty of U, when V is known.

Given M samples with N features $X_1, \ldots, X_N$, and the target classification variable Y,

feature selection is to find K features $X_{\nu(1)}, \ldots, X_{\nu(K)}$ that optimally characterizes Y. Mutual information based feature selection [14] is to select features $\nu(1), \ldots, \nu(K)$ which individually maximize the mutual information $I(Y; X_{\nu(l)})$. However, feature selection based on such a criterion cannot ensure weak dependency among features, and can lead to redundant and poorly informative families of features. Recently Fleuret [54] proposed a Conditional Mutual Information maximization criterion to select binary features. Its essence is that a feature X can be discarded if there is one feature X_ν already picked such that X and Y are conditionally independent given X_ν. Conditional Mutual Information is defined as

$$I(U; V|W) = H(U|W) - H(U|W, V) \tag{3.15}$$

that measures the information shared between U and V when W is known. If V and W carry the same information about U, the two terms on the right are equal, and the CMI is zero, even if both V and W are individually informative. On the contrary if V brings information about U which is not already contained in W, the difference is large.

For feature selection, a feature X' is good only if $I(Y; X'|X)$ is large for every X already picked. This means that X' is good only if it carries information about Y, and if this information has not been caught by any of the X already picked. An iterative procedure for a CMI based feature selection can be defined as

$$\nu(1) = \arg\max_{n \in N} I(Y; X_n) \tag{3.16}$$

$$\forall k, 1 \leq k < K, \nu(k+1) = \arg\max_{n \in N} \left\{ \min_{l \leq k} I(Y; X_n | X_{\nu(l)}) \right\} \tag{3.17}$$

$I(Y; X_n | X_{\nu(l)})$ is small either if X_n contains no information about Y, or if such information was already in $X_{\nu(l)}$. A similar criterion was also proposed independently in [164])

CMI based Boosting

The above CMI based feature selection works on binary features. Here we extend it to learn weak classifiers from a large classifier pool, and combine the learned weak classifiers into

a strong classifier. Specifically, we regard the output of the weak classifier as a random variable, a "feature" for the candidate class, and employ the CMI maximization criterion to select the effective "features", i.e. the characterizing weak classifiers. In this way, a sequence of weak classifiers is learned which maximize their mutual information about a candidate class, conditional to the response of any weak classifier already selected. So a weak classifier similar to those that were already learned will not be selected, even if it is individually powerful as it does not carry additional information about the candidate class.

After learning weak classifiers, a strategy is needed to perform final classification by combining the learned weak classifiers. We adopt the Naive-Bayes to make the final decision based on outputs of the weak classifiers, contrary to the voting procedure used in AdaBoost. A Naive-Bayes classifier is simple but highly effective if the features can be assumed to be largely independent for a given class. As the weak classifiers learned based on CMI are by their very nature weakly dependent, it is reasonable to use Naive-Bayes to combine them for final classification. If using c to represent the value of the class variable, and $x_1, \cdots, x_k$ for the features, a Naive Bayesian classifier is defined as

$$\hat{c} = \arg \max_c p(c) \prod_{i=1}^{k} p(x_i|c) \tag{3.18}$$

As the presented method finds a highly accurate classifier by combining many weak classifiers, each of which is only moderately accurate, we call it CMI based "Boosting" (CMIB). But CMIB has little relationship with typical boosting algorithms as there is no reweighting of the samples. The CMIB algorithm is summarized in Figure 3.11. CMIB learns weak classifiers that are both individually informative and weakly dependent.

Speed up CMIB

AdaBoost usually requires very expensive training time. The improved performance of Float-Boost [91] also pays the price for 5 times longer training time than AdaBoost. In contrast, CMIB promises very fast training. Fleuret presented in [54] the way to speed up the CMI based feature selection, which is also suitable for our CMIB. We reintroduce the speed-up

Given a training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $y_i \in \{-1, 1\}$ and the size of the final strong classifier T:

1. Train weak classifiers $H_1(x), \ldots, H_N(x)$ based on training samples, where N is the total number of the candidate weak classifiers

2. For $t = 1, \ldots, T$,

   - If $t = 1$, choose $H_t(x) = H_j(x)$, such that

   $$j = \arg\max_{n \in N} I(Y; H_n(x)) \qquad (3.19)$$

   - If $t > 1$, choose $H_t(x) = H_j(x)$, such that

   $$j = \arg\max_{n \in N} \left\{ \min_{l < t} I(Y; H_n(x)|H_l(x)) \right\} \qquad (3.20)$$

3. Output the final strong classifier as

$$f(x) = \arg\max_y p(y) \prod_{i=1}^{T} p(H_i(x)|y) \qquad (3.21)$$

Figure 3.11: Conditional Mutual Information based Boosting.

method here (please refer to [54] for details).

The most expensive part in CMIB is Step 2 in Figure 3.11, where T weak classifiers $H_1, \ldots, H_T$ are learned. The straight-forward implementation of Step 2 keeps a score vector s which contains for every weak classifier H_n the score $s[n] = \min_{l \leq t} I(Y; H_n(x)|H_l(x))$, after the choice of $H_t(x)$. The score vector is initialized with the values $I(Y; H_n(x))$. At each iteration the weak classifier j with the highest score is selected, and then score $s[n]$ is updated by taking the minimum of $s[n]$ and $I(Y; H_n(x)|H_j(x))$. This is illustrated in Algorithm 1 given below, where `MI(n)` computes $I(Y; H_n(x))$ and `CMI(n, m)` computes $I(Y; H_n(x)|H_m(x))$:

We can speed up the computation based on the fact that because the score vector can only decrease in the process, bad scores do not need to be updated. This means `CMI(n, m)` only need be called for good weak classifiers. Now, for every weak classifier $H_n(x)$, we store a partial score $ps[n]$, which is the minimum over a few of the conditional mutual information in Eqn. (3.20). Another vector $idx[n]$ contains the index of the last weak classifier taken into

---

Algorithm 1:

for $n = 1 \dots N$ **do**
  $s[n] \leftarrow$ `MI`(n)
end
for $t = 1 \dots T$ **do**
  $sel[t] = \arg\max_n s[n]$
  for $n = 1 \dots N$ **do**
    $s[n] \leftarrow \min(s[n], $`CMI`$(n, sel[t]))$
  end
end

---

account in the computation of $ps[n]$. Thus, we have

$$ps[n] = \min_{l \leq idx[n]} I(Y; H_n(x)|H_l(x)) \tag{3.22}$$

At each iteration, the algorithm goes through all candidates and updates its score only if the best one found so far in that iteration is not better compared to it. For example, the best score (*bests*) of the first k weak classifiers is 0.2, and the $k + 1$ weak classifier's score was 0.05, it is not necessary to update the $k + 1$ weak classifier's score. This gives us a speed-up implementation illustrated in Algorithm 2 below.

Algorithm 2:

for $n = 1 \dots N$ **do**
  $ps[n] \leftarrow$ `MI`(n)
  $idx[n] \leftarrow 0$
end
for $t = 1 \dots T$ **do**
  $bests \leftarrow 0$
  for $n = 1 \dots N$ **do**
    while $ps[n] > bests$ **and** $idx[n] < t - 1$ **do**
      $idx[n] \leftarrow idx[n] + 1$
      $ps[n] \leftarrow \min(ps[n], $`CMI`$(n, sel[idx[n]]))$
    end
    if $ps[n] > bests$ **then**
      $bests \leftarrow ps[n]$
      $sel[t] \leftarrow n$
    end
  end
end
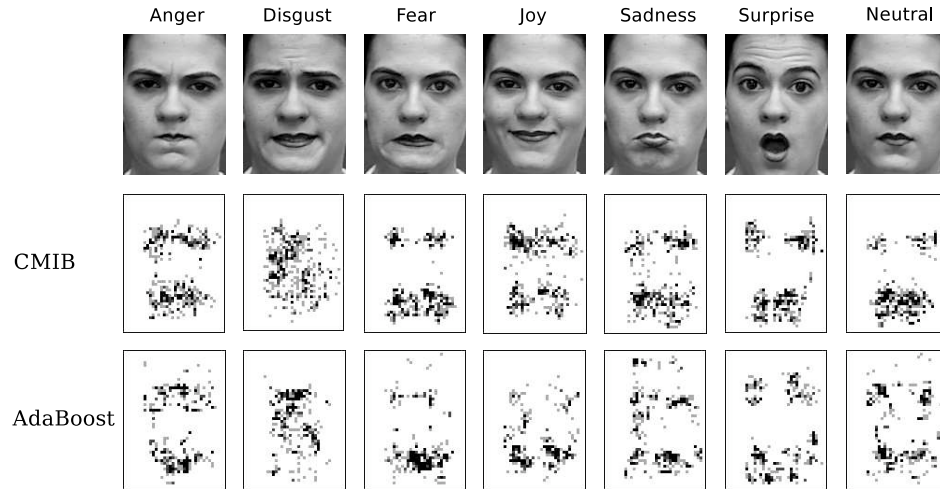
Figure 3.12: Distributions of top 50 sub-regions (LBP histograms) selected by CMIB and AdaBoost for each expression.

3.5 Boosting LBP for Facial Expression Recognition

In order to compare quantitatively the effectiveness of CMIB with AdaBoost, we adopt them in this section to learn a small subset of the most discriminative LBP features for facial expression recognition. As each LBP histogram is calculated from a sub-region, boosting methods are actually used to find the sub-regions that contain more discriminative information for expression recognition in terms of LBP histogram. By shifting and scaling a sub-window, 16,640 sub-regions, i.e. 16,640 LBP histograms in total were extracted from each face image. As AdaBoost and CMIB both work on two-class problems, the multi-class problem here is accomplished by using the one-against-rest technique, which trains AdaBoost/CMIB between one expression against all others. For each AdaBoost/CMIB learner, the images of one expression were positive samples, while the images of all other expressions were negative samples. We first present comparisons between AdaBoost and CMIB, and then perform expression recognition with the learned LBP features.

3.5.1 Comparison Experiments of AdaBoost and CMIB

We plot in Figure 3.12 the spatial locations of top sub-regions (i.e. the centers of the sub-regions) that corresponded by the top 50 LBP histograms selected by AdaBoost and CMIB

for each expression. In images of the bottom two rows of Figure 3.12, the gray scale of each small grid is proportional to the number of sub-regions selected in that grid. We can see that there are many common features selected by CMIB and AdaBoost, though there are also some different features selected by each as different criterion adopted. It is observed that different expressions have different key discriminant features, and the discriminant features are mainly distributed in the eye and mouth regions. We compare CMIB and AdaBoost from the following three different perspectives.

Training Computational Complexity — We plot the average training time of CMIB and AdaBoost as a function of the number of weak classifiers in Figure 3.13. Our experiments were run on a standard 2.0Ghz PC with the Matlab implementation. It can be seen that CMIB performs significantly faster than AdaBoost, especially when the number of learned weak classifiers increases. For example, CMIB selects top 100 weak classifiers with an average time of 1.5t, while AdaBoost needs 38.3t for the task. The variation in AdaBoost running time was due to the network or system load, since we conducted experiments with Matlab installed in a central server. Even with such variations, we can still observe that the training time of AdaBoost is linear to the number of feature selection rounds.

The fast training of CMIB is very significant for any incremental or adaptive learning when the size of the initial available training data is small but accumulating over time, which exists in most real-world vision problems.

Classification Accuracy — We conducted expression recognition using strong classifiers boosted by CMIB and AdaBoost. The generalization performance in 6-class and 7-class recognition are shown in Figure 3.14, as a function of the number of weak classifiers. Here a strong classifier is composed of up to 200 weak classifiers.

We can observe in Figure 3.14 that the generalization performance is clearly improved for both the 6-class and 7-class recognition tasks by boosting LBP-based classifiers over that of Uniform-LBP (84.5% and 79.1% respectively as shown in Table 3.1). It is seen that AdaBoost performs better when using less than 40-60 weak classifiers, while CMIB achieves as good or
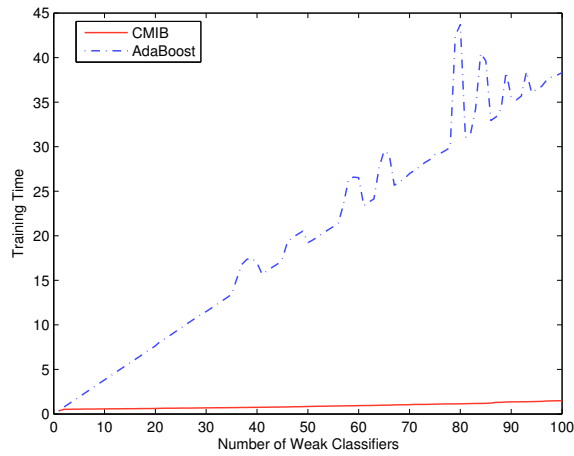
Figure 3.13: Training time of CMIB and AdaBoost, as a function of the number of weak classifiers.
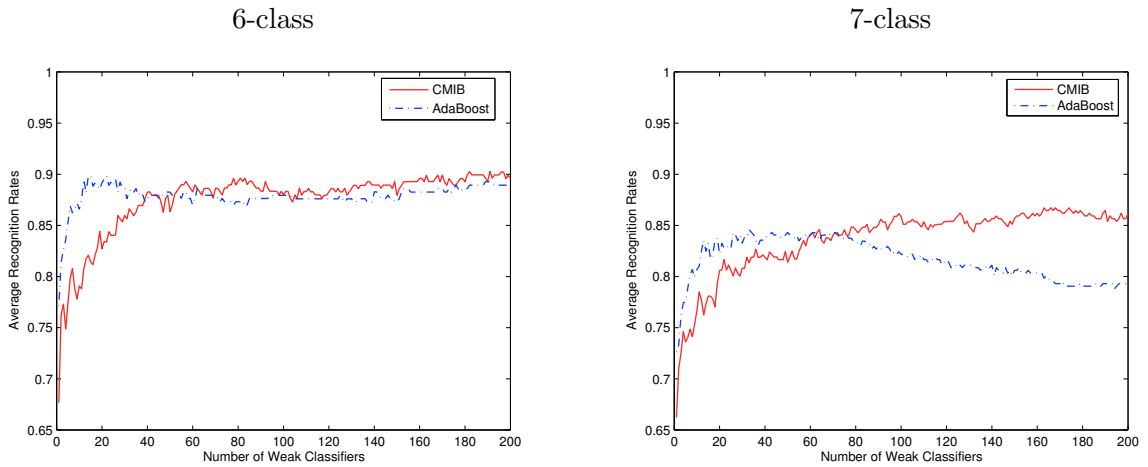
6-class 7-class

Figure 3.14: Generalization performance of the boosted classifier using CMIB and AdaBoost, as a function of the number of weak classifiers. (Left) 6-class; (Right) 7-class.

better recognition results when more weak classifiers are included.

Between-Class Discriminative Robustness — We show the outputs of boosted classifiers of different expressions for samples "Joy", "Surprise", and "Neutral" in Figure 3.15. It can be seen that different weak classifiers being learned by CMIB and AdaBoost have some impact on not only the recognition accuracy, but also the robustness of recognition. Weak classifiers learned by CMIB seems to provide better discriminative ability in between-class
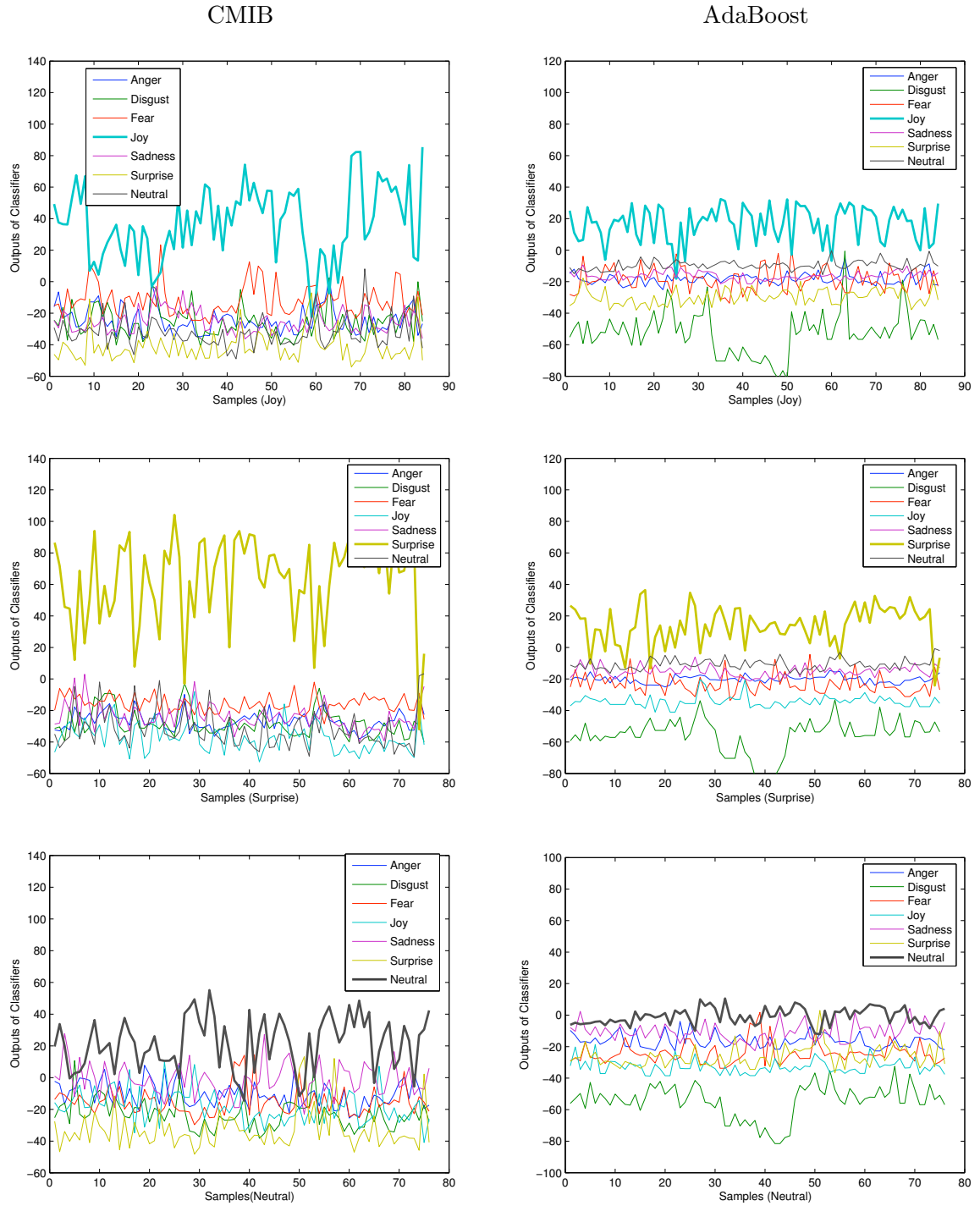
Figure 3.15: (Best viewed in color) Outputs of classifiers for samples Joy, Surprise, and Neural. The left column: CMIB; the right column: AdaBoost.

separation than that of AdaBoost, resulting in more robust recognition. The above comparative experiments illustrate that CMIB enables much faster training and yields a classifier of improved overall accuracy and discrimination over that of AdaBoost. However, with the first several ten weak classifiers, AdaBoost provides superior performance.

3.5.2 Facial Expression Recognition using Boosted-LBP

As AdaBoost performs better than CMIB when using the first several tens of weak classifiers, in this section, we learn the most discriminative LBP histograms using AdaBoost (Boosted-LBP) for facial representation.

We performed facial expression recognition using the strong classifier boosted by AdaBoost, and outputs the class with the largest positive output of binary classifiers. In our experiments, AdaBoost training continued until the classifier output distribution for the positive and negative samples were completely separated, so the number of LBP histograms selected for each expression was not pre-defined, but automatically determined by the AdaBoost learning itself. In the 10-fold experiments, the number of selected LBP histogram ranges 49-52 for 6-class expressions and 65-70 for 7-class expressions. Figure 3.16 shows selected sub-regions (LBP histograms) for each basic expression in one trial of the 10-fold cross-validation. We can observe that the selected sub-regions have variable sizes and positions. Moreover, while the weights of sub-regions in the template matching in Section 3.2.2 were chosen empirically, the weights in boosted classifiers were learned automatically by AdaBoost. The generalization performance of the boosted classifiers is 84.6% for 7-class recognition and 89.8% for 6-class recognition respectively. As shown in Table 3.12, compared to the Uniform-LBP based template matching in Section 3.2.2, AdaBoost (boosted-LBP) provides improved performance. We also show the confusion matrix of 7-class recognition using AdaBoost in Table 3.13, where Disgust, Joy, Surprise, and Neutral can be recognized with high accuracy. It can be seen that AdaBoost's performance is inferior to that of SVM (RBF) reported in Table 3.5 for most expressions except Fear and Neutral.

We further combine feature selection by AdaBoost with classification by SVM. In particular, we train SVM with the Boosted-LBP features. In each trial of the 10-fold cross-validation,
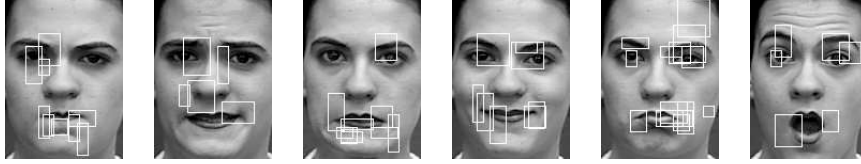
Figure 3.16: The sub-regions (LBP histograms) selected by AdaBoost for each emotion. from left to right: Anger, Disgust, Fear, Joy, Sadness, Surprise.

	7-Class Recognition	6-Class Recognition
AdaBoost (Boosted-LBP)	85.0%	89.8%
Uniform-LBP + Template matching	79.1%	84.5%

Table 3.12: Recognition performance of Boosted-LBP vs Uniform-LBP.

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Anger	**66.6%**	3.7%	2.0%	0	7.3%	0	20.4%
Disgust	0	**92.5%**	2.5%	0	0	0	5.0%
Fear	0	0	**70.0%**	17.0%	3.0%	0	10.0%
Joy	0	0	2.5%	**90.1%**	0	0	7.4%
Sadness	6.4%	0	0	0	**61.2%**	0.8%	31.6%
Surprise	0	0	1.3%	0	0.5%	**92.5%**	5.7%
Neutral	0	0	0.8%	0.4%	3.6%	0	**95.2%**

Table 3.13: Confusion matrix of 7-class facial expression recognition using AdaBoost (Boosted-LBP).

we applied AdaBoost to learn the discriminative LBP histograms for each expression, and then utilized the union of the selected LBP histograms as the input for SVMs. For example, in Figure 3.16, the union of all sub-regions selected resulted in a total of 51 LBP histograms. The generalization performance of Boosted-LBP based SVM is summarized in Table 3.14, where the degree of the polynomial kernel is 1 and the standard deviation for the RBF kernel is 2^{11}. For comparison, we also include the recognition performance of SVMs with the Uniform-LBP (in Section 3.2) in Table 3.14. We observe that Boosted-LBP based SVMs outperform SVMs using Uniform-LBP by around 2.5-3.5 percent points. The 7-class expression recognition result of 91.4% is very encouraging, compared to the state of the art [28]. Bartlett *et al.* [10] obtained the best performance 93.3% by selecting a subset of Gabor filters using AdaBoost and then training SVM on the outputs of the selected filters. With regard to the 6-class recognition, the result of 95.1% is, to our best knowledge, the best recognition rate reported so far in the published literature on this database. Previously Tian [152]

achieved 94% performance using a three-layer neural networks when combining geometric features and Gabor wavelet features. The confusion matrix of 7-class expression recognition using Boosted-LBP based SVM (RBF) is shown in Table 3.15. We can observe that, Disgust, Joy, and Surprise can be recognized with very high accuracy (more than 97%), and Sad is the easiest confused expression with recognition accuracy around 75%. We also re-conducted the experiments on low-resolution face images in Section 3.3 using the boosted-LBP features, and the recognition rates are all increased by 3-5%.

	7-Class		6-Class	
	Boosted-LBP	Uniform-LBP	Boosted-LBP	Uniform-LBP
SVM (Linear)	91.1%	88.1%	95.0%	91.5
SVM (Polynomial)	91.1%	88.1%	95.0%	91.5
SVM (RBF)	91.4%	88.9%	95.1%	92.6

Table 3.14: Recognition performance of Boosted-LBP based SVMs vs Uniform-LBP based SVMs.

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Anger	**85.1%**	2.7%	0	0	8.6%	0	3.6%
Disgust	0	**97.5%**	0.8%	1.7%	0	0	0
Fear	0	1.0%	**79.9%**	11.0%	3.1%	1.0%	4.0%
Joy	0	0	0	**97.5%**	0.4%	0	2.1%
Sadness	12.0%	0	0.8%	0	**74.7%**	0	12.5%
Surprise	0	0	1.3%	0.9%	0	**97.3%**	0.5%
Neutral	1.2%	0	0.8%	3.6%	2.4%	0	**92.0%**

Table 3.15: Confusion matrix of 7-class facial expression recognition using Boosted-LBP based SVM.

3.6 Generalization to Other Datasets

We evaluated the Boosted-LBP based SVM approach on the MMI database [123] and the JAFFE database [97]. The MMI database includes more than 20 subjects of both sexes (44% female), ranging in age from 19 to 62, having either a European, Asian, or South American ethnic background. Subjects were instructed to display 79 series of facial expressions, six of which are prototypic emotions. Image sequences have neutral faces at the beginning and the end, and were digitized into 720×576 pixels. Some sample images from the MMI database are shown in Figure 3.17. Although the original data in the MMI database are
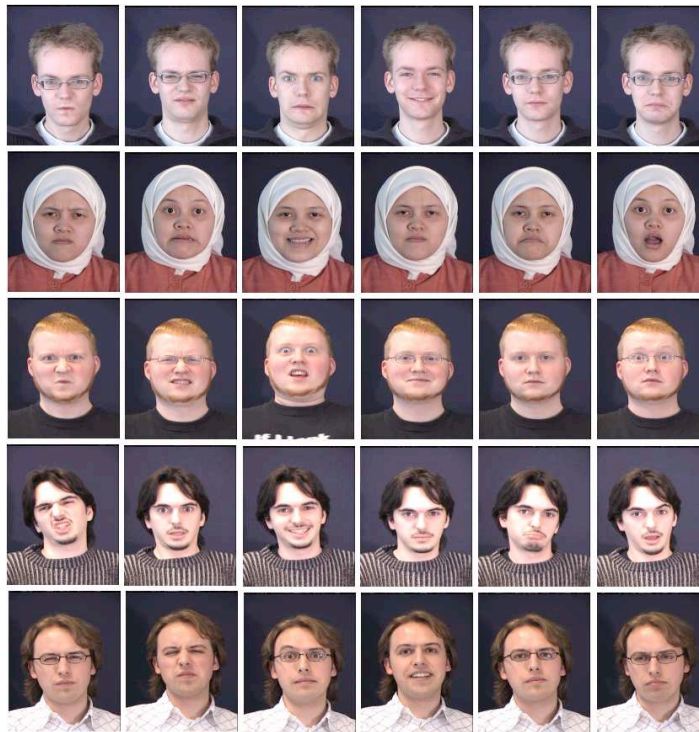
Figure 3.17: Sample face expression images from the MMI database.

color images, in our experiment, we converted them to 8-bit grayscale images. As can be seen, the subjects displayed facial expressions with and without glasses, which makes facial expression recognition more difficult. The JAFFE database consists of 213 images of Japanese female facial expressions. Ten expressers posed 3 or 4 examples for each of the seven basic expressions (six emotional expressions plus neutral face). The image size is 256×256 pixels. Figure 3.18 shows some sample images from the JAFFE database. As we did on the Cohn-Kanade database (in Section 3.2.1), we normalized the faces of these two databases to a fixed distance between the two eyes, and cropped facial images of 110×150 pixels from original frames based on the two eyes location.

In our experiments, 96 image sequences were selected from the MMI database. The only selection criterion is that a sequence can be labeled as one of the six basic emotions. The sequences come from 20 subjects, with 1 to 6 emotions per subject. The neutral face and three peak frames of each sequence (hence, 384 images in total) were used for 7-class expression recognition. All 213 images of the JAFFE database were used for 7-class expression
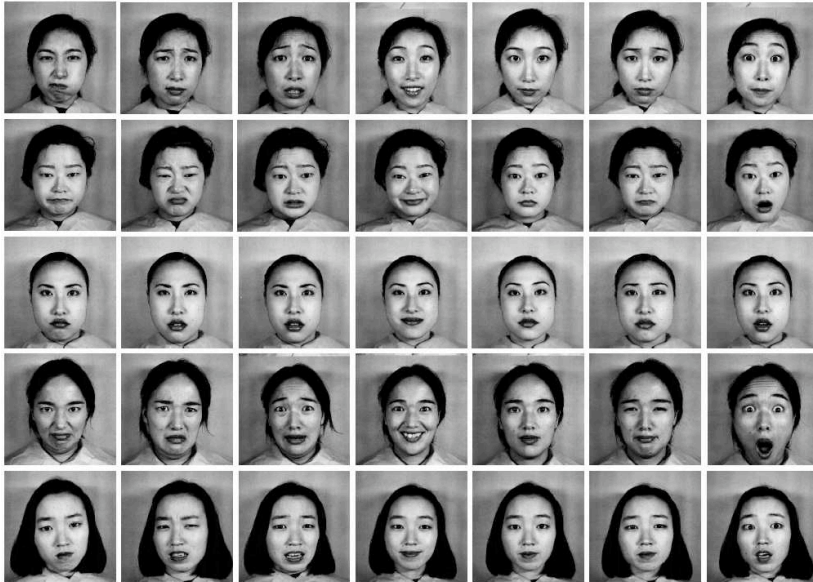
Figure 3.18: Sample face expression images from the JAFFE database.

recognition.

We first performed 10-fold cross-validation on each data set, and the recognition rates are shown in the top two rows of Table 3.16, where the degree of the polynomial kernel is 1 and the standard deviation for the RBF kernel is 2^{15} for the MMI database and 2^8 for the JAFFE database. The best recognition performance of 86.9% on the MMI database is inferior to that on the Cohn-Kanade database. This is possibly because that there are fewer images in the data set, and subjects are wearing glasses. The performance on the JAFFE is worst overall compared to that of the Cohn-Kanade database and the MMI database, and this may be also due to a much small data set. With LBP features and the linear programming technique, Feng *et al.* [53] reported the performance of 93.8% on the JAFFE database. They preprocessed face images to exclude nonface area with an elliptical mask. Liao *et al.* [93] recently reported the recognition performance of 85.6% on the JAFFE database, but they did not conducted 10-fold cross-validation.

We then performed across-dataset experiments, i.e. we performed LBP feature selecting and SVM training on the Cohn-Kanade database, and then tested the classifier on the MMI database and the JAFFE database respectively. Recognition results are shown in the bottom

two rows of Table 3.16, where the degree of the polynomial kernel is 1 and the standard deviation for the RBF kernel is 2^{14} for the MMI database and 2^{11} for the JAFFE database. We observe that generalization performance across data sets was much lower, around 50% on the MMI database and around 40% on the JAFFE database. These results actually reenforce Bartlett *et al.* 's recent finding [95], where they trained selected Gabor-wavelet features based SVMs on the Cohn-Kanade database and tested them on another Pictures of Facial Affect database, and obtained 56-60% performance. As we preprocessed face images of different databases in the same way, the only difference between the data is that they were collected under different controlled environments. So the current expression classifier trained on a single dataset with uniformly controlled environment works well only within that dataset. This suggests that, in order to generalize across image collection environments, we have to collect large training datasets with variations in image conditions [95].

	SVM (Linear)	SVM (Polynomial)	SVM (RBF)
MMI	86.7%	86.7%	86.9%
JAFFE	79.8%	79.8%	81.0%
Train:Cohn-Kanade Test:MMI	50.8%	50.8%	51.1%
Train:Cohn-Kanade Test:JAFFE	40.4%	40.4%	41.3%

Table 3.16: Generalization performance of Boosted-LBP based SVM on other datasets.

3.7 Summary

In this chapter, we presented a comprehensive empirical study of facial representation based on Local Binary Patterns. The key issues can be summarized as follows:

1. Compared to Gabor wavelets, LBP features can be extracted faster in a single scan through the raw image and lie in a much lower dimensional space, bringing significant time and space benefit. Among different classification techniques examined, SVM classifiers provide best performance for person-independent facial expression recognition.

2. One challenging problem is recognizing facial expressions at low resolutions, as only compressed low-resolution video input is available in real-world applications. We also investigate LBP features on low-resolution images, and our experiments demonstrate

that LBP features perform stably and robustly over a useful range of low resolutions of face images.

3. We study feature selection techniques to extract the most discriminative LBP features for better expression recognition. To address the problem that AdaBoost requires very expensive training time, we present an efficient procedure based on Conditional Mutual Information. The best recognition performance is obtained by using SVM classifiers with Boosted-LBP features. However, this method has limited generalization ability to other datasets.

All the recognition experiments are performed on static images without considering temporal behaviours of facial expressions. The psychological experiments [13] suggest that temporal dynamics is a critical factor for successful interpretation of facial expressions. This is especially true for spontaneous facial expressions without any deliberate exaggerated posing [2]. In next chapter, we present a method to capture and represent temporal dynamics of facial expression by discovering the underlying low-dimensional manifold.

4 Manifold Analysis of Facial Expression

The temporal dynamics of human behavior is a critical factor for successful interpretation of an observed behavior [117]. The differences between facial (or bodily) expressions are often conveyed more powerfully by dynamic transitions between different stages of expressions rather than any single state represented by a still image. In this chapter, we present a method to capture and represent facial expression dynamics by discovering the underlying low-dimensional manifold.

To address the limitations of the existing work [25, 26, 72], we exploit Locality Preserving Projections (LPP) to learn the expression manifold in the LBP based appearance feature space. Compared to LLE [135] and Isomap embedding [151] adopted previously in [25, 72, 89], LPP provides an explicit mapping from the input space to the reduced space, so is better suited to facial expression recognition. The LBP based appearance features offer advantages over shape models using sparse 2D feature points [25, 26, 72] in describing detailed facial deformations that are important to facial expression modeling. One challenging problem in expression manifold learning is to obtain a generalized representation for facial expressions from different subjects. By deriving a universal discriminant expression subspace using a supervised LPP (SLPP), we effectively align manifolds of different subjects on a generalized expression manifold. We also comprehensively evaluate different linear subspace methods in expression subspace learning.

We further formulate a Bayesian framework to examine both the temporal and appearance characteristics for dynamic facial expression recognition employing the derived manifold representation. Our method is person-independent, and provides superior performance to both static frame-based methods and HMM based models in comparison experiments presented

in Section 4.5. As facial expressions vary in intensity, it is helpful to estimate the expression intensity for quantitative assessment of facial expression. We also present a method for expression intensity estimation using the Fuzzy K-Means method using the derived expression manifold.

4.1 Expression Manifold Learning

A number of nonlinear dimensionality reduction techniques have been proposed for manifold learning, e.g. Isomap [151], LLE [135], and Laplacian Eigenmaps [16]. However, these techniques yield mappings defined only on the training data, and do not provide explicit mappings from the input space to the reduced space. Therefore, they may not be suitable for facial expression recognition tasks. In particular, Chang *et al.* [25] investigated LLE for facial expression manifold learning and their experiments show that LLE is better suited to visualizing expression manifolds but fails to provide good expression classification. Alternatively, He and Niyogi [66, 67] proposed a general manifold learning method called Locality Preserving Projections (LPP). LPP builds a graph model which reflects the intrinsic geometric structure of a given data space, and finds optimal projections that preserves locality with respect to the graph structure. Although it is still a linear technique, LPP is shown to recover important aspects of nonlinear manifold structure. More crucially, LPP is defined everywhere in the ambient space rather than just on the training data, therefore it has a significant advantage over other techniques in explaining the test data in the reduced subspace. In this chapter, we examine LPP for learning facial expression manifold.

Given a data set $x_1, x_2, \ldots, x_m$ in $\mathbb{R}^n$, LPP finds a transformation matrix W to map it to $y_1, y_2, \ldots, y_m$ in $\mathbb{R}^l (l \ll n)$, such that $y_i = W^T x_i$. Let $\mathbf{w}$ denote the transformation vector, LPP derives the optimal projections preserving locality by the objective function [16]:

$$\min_{\mathbf{w}} \sum_{i,j} \left(\mathbf{w}^T x_i - \mathbf{w}^T x_j \right)^2 S_{ij} \tag{4.1}$$

where S_{ij} evaluates a local structure of the data space, and is defined as:

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{t}} & \text{if } x_i \text{ and } x_i \text{ are ``close''}, \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

or in a simpler form as

$$S_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_i \text{ are ``close''}, \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

where "close" can be defined as $\|x_i - x_j\|^2 < \epsilon$, where ϵ is a small constant, or x_i is among k nearest neighbors of x_j or x_j is among k nearest neighbors of x_i. The objective function with symmetric weights $S_{ij}(S_{ij} = S_{ji})$ incurs a heavy penalty if neighboring points x_i and x_j are mapped far apart. Minimizing their distance is therefore to ensure that if x_i and x_j are "close", $y_i(= \mathbf{w}^T x_i)$ and $y_j(= \mathbf{w}^T x_j)$ are also "close". The objective function of Eqn. (4.1) can be reduced to:

$$\begin{aligned} \frac{1}{2}\sum_{ij}(\mathbf{w}^T x_i - \mathbf{w}^T x_j)^2 S_{ij} &= \sum_i \mathbf{w}^T x_i D_{ii} x_i^T \mathbf{w} - \sum_{ij} \mathbf{w}^T x_i S_{ij} x_j^T \mathbf{w} \\ &= \mathbf{w}^T X (D - S) X^T \mathbf{w} \\ &= \mathbf{w}^T X L X^T \mathbf{w} \end{aligned} \tag{4.4}$$

where $X = [x_1, x_2, \ldots, x_m]$ and D is a diagonal matrix whose entries are column (or row, since S is symmetric) sums of S, $D_{ii} = \sum_j S_{ji}$. $L = D - S$ is a Laplacian matrix. D measures the local density on the data points. The bigger the value D_{ii} is (corresponding to y_i), the more "important" is y_i. Therefore, a constraint is imposed as follows:

$$\mathbf{y}^T D \mathbf{y} = 1 \Rightarrow \mathbf{w}^T X D X^T \mathbf{w} = 1 \tag{4.5}$$

The transformation vector $\mathbf{w}$ that minimizes the objective function is given by the minimum

eigenvalue solution to the following generalized eigenvalue problem:

$$XLX^T\mathbf{w} = \lambda XDX^T\mathbf{w} \tag{4.6}$$

Suppose that a set of vectors $\mathbf{w}_0, \ldots, \mathbf{w}_{l-1}$ is the solution, ordered according to their eigenvalues, $\lambda_0, \ldots, \lambda_{l-1}$, the transformation matrix is then derived as $W = [\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{l-1}]$. The obtained projections are the optimal linear approximation to the eigenfunctions of the Laplace Beltrami operator on the manifold [66].

For appearance-based facial expression analysis, the dimension of the feature space (n) is often much larger than the number of samples in a training set (m). Thus, matrix XDX^T is likely to be singular. To overcome this problem, PCA can be first applied to project the training data into a low-dimensional subspace. If the transformation matrix of PCA is denoted as W_{PCA}, the final transformation matrix is

$$W = W_{PCA}W_{LPP}, \quad W_{LPP} = [\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{l-1}]. \tag{4.7}$$

The optimal data set for expression manifold learning might contain $O(10^2)$ subjects, and each subject has $O(10^3)$ images that cover basic expressions. However, until now, there is no such a database that can meet this requirement. Chang *et al.* [25, 26] conducted experiments on a small data set, e.g. only two subjects (one male and one female) were used in [25]. Here we conduct experiments on the Cohn-Kanade database [78] which consists of 100 subjects, though each subject only has several tens of frames of basic expressions. In our experiments, 316 image sequences (5,876 images in total) of basic expressions were selected from the database, which come from 96 subjects, with 1 to 6 emotions per subject. Each sequence begins with the neutral face and ends with a typical facial expression at apex, and the duration of each expression varied.

We first learn the expression manifold of each individual. We randomly selected six subjects from the data set, each of which has six image sequences corresponding to six basic expressions, and then applied LPP to embed image sequences of each subject into a low-
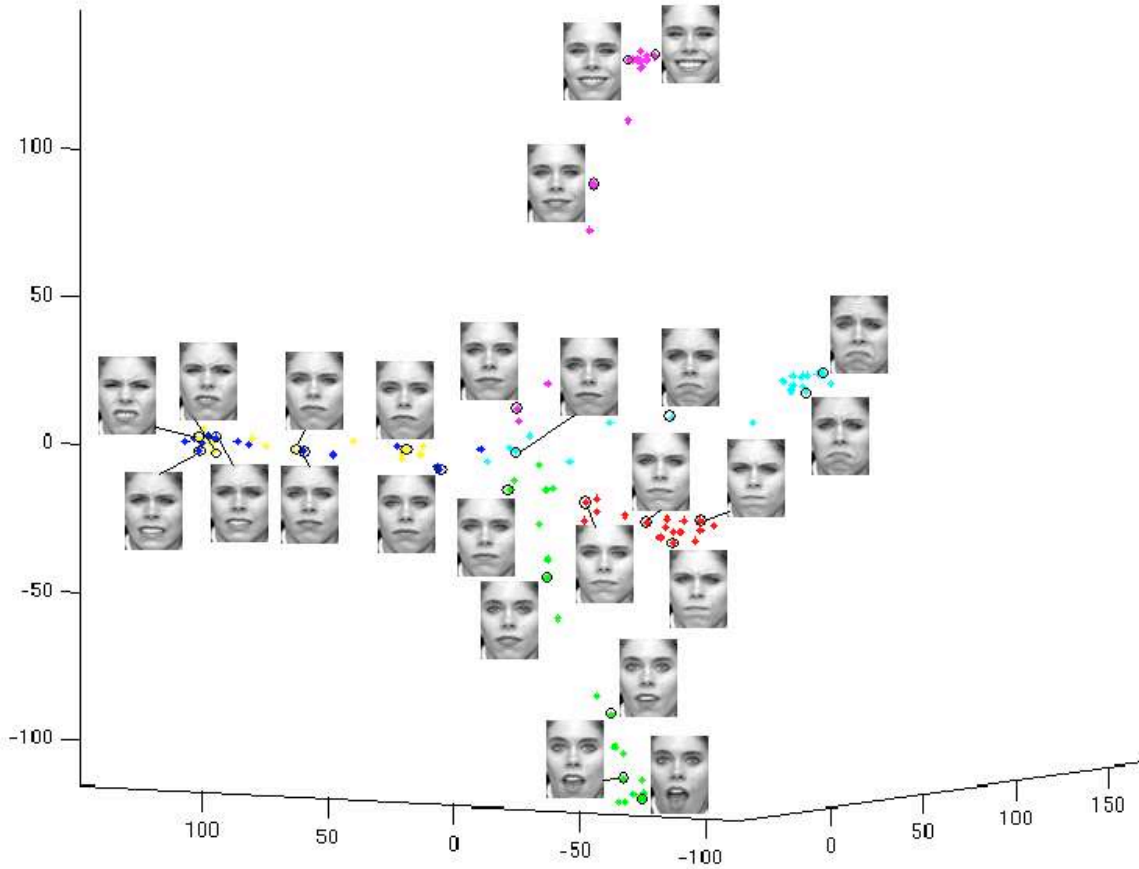
Figure 4.1: (Best viewed in color) Six image sequences of basic expressions of an individual are mapped into the embedding space described by the first three coordinates of LPP. Representative faces are shown next to circled points in different parts of the space. Different expressions are color coded as: Anger (red), Disgust (yellow), Fear (blue), Joy (magenta), Sadness (cyan), and Surprise (green). (Note: these color codes remain the same in all figures throughout the rest of this chapter.)

dimensional subspace respectively. By applying LPP to the LBP appearance feature space, image sequences of facial expressions of an individual are mapped into the embedded space as shown in Figure 4.1. The embedded manifolds of another five subjects are shown in Figure 4.2. It is observed that expression images of each individual were embedded as a smooth manifold in low-dimensional subspace, and every image sequence is mapped to a curve on the manifold that begins from the neutral face and extends in distinctive directions with varying intensity of facial expression.

Following this, we applied LPP to images sequences of multiple subjects to derive expression
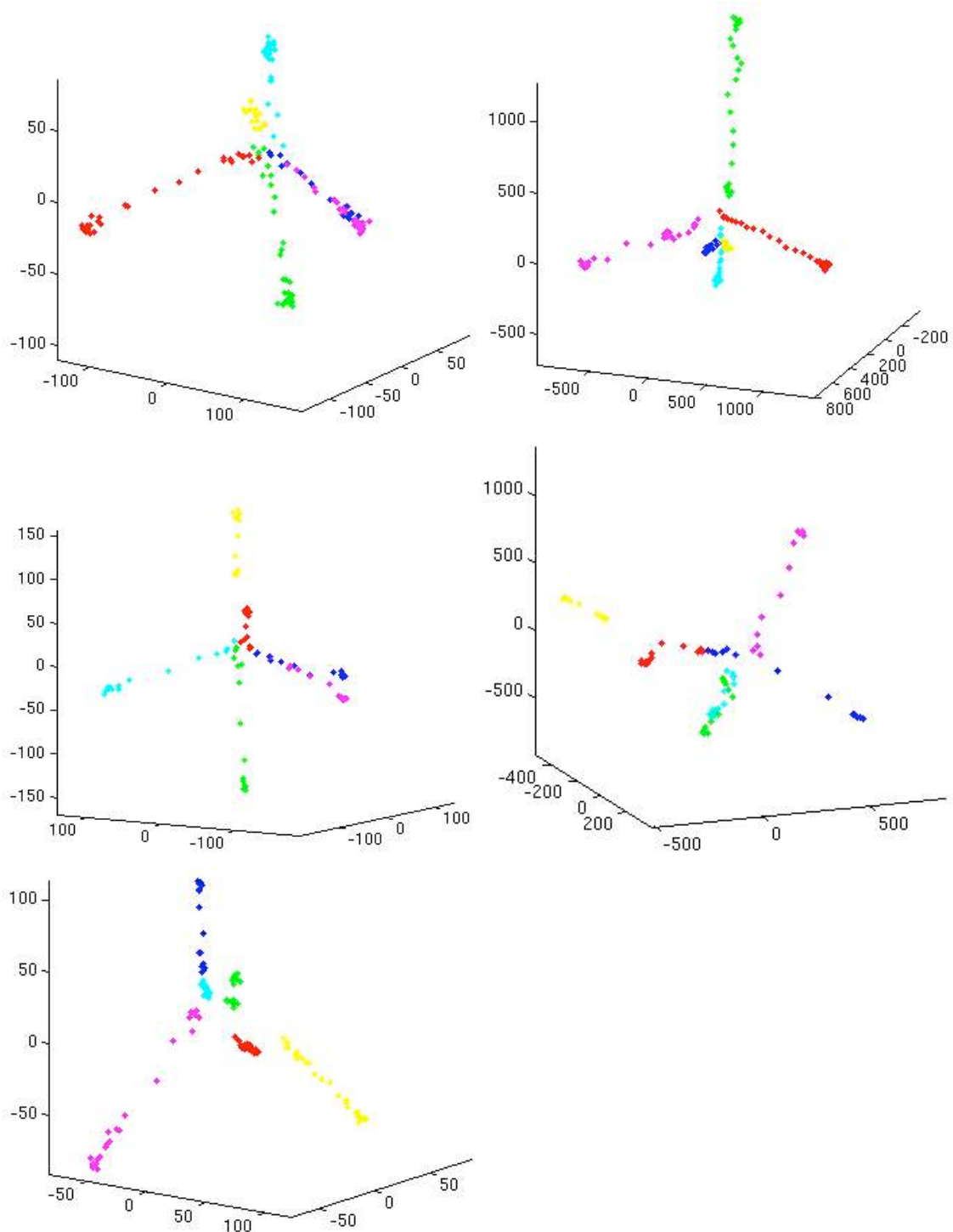
Figure 4.2: (Best viewed in color) 3-D visualization of expression manifolds of five subjects.
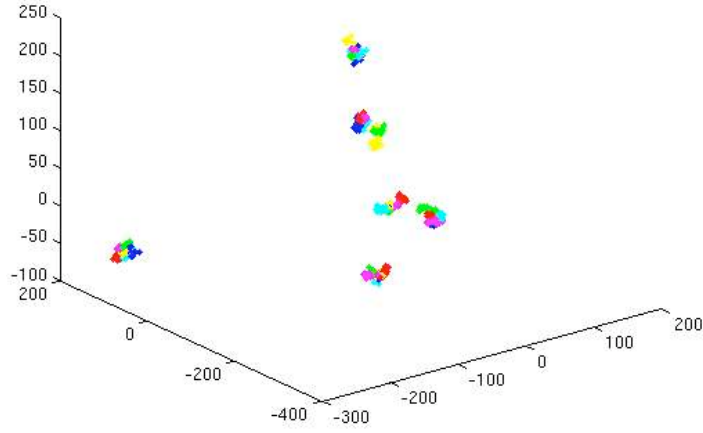
Figure 4.3: (Best viewed in color) Image sequences of six subjects mapped into the embedding space described by the first three coordinates of LPP.

manifold of different subjects. Figure 4.3 shows the embedded manifold of image sequences from six subjects, which demonstrates that different subjects correspond to different clusters, and the manifolds of different subjects vary a great deal in the covered regions and the stretching directions. We further show the embedded manifold of image sequences from all 96 subjects in Figure 4.4. It can be observed that the embedding of the same facial expression of different subjects is quite different. There are two types of variations in the data set: different subjects and varying expression intensity. Generally, LPP can cope with intensity variation, to map each image sequence to a curve on the manifold. As an unsupervised learning algorithm that constructs the low-dimensional manifold based on the neighborhood graph, it is harder for LPP to keep images of similar expression but from different subjects in the near region on the manifold, when facial appearance variation across different subjects is greater than that due to facial expressions.

4.2 A Generalized Expression Manifold

Our aim is to derive a subject-independent manifold-based unified representation of facial expressions. Due to the significant variations of appearance and facial deformation across different subjects, the manifolds of different subjects vary a great deal (Figure 4.3 and Figure 4.4). But conceptually all the manifolds (for the same expression) are the same, and there
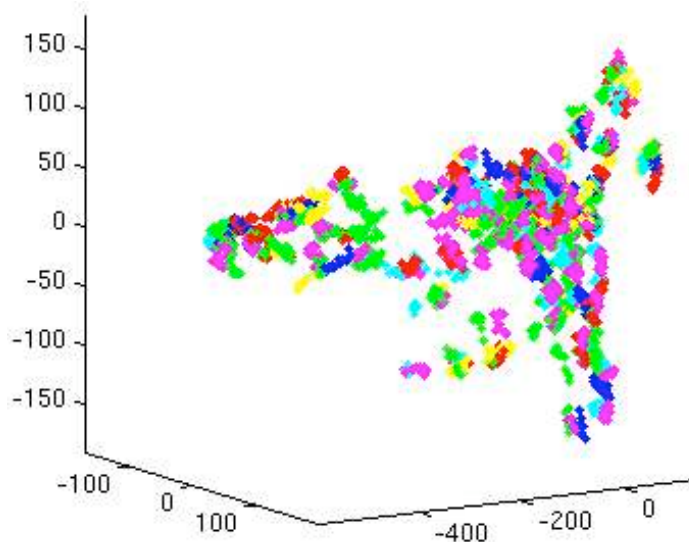
Figure 4.4: (Best viewed in color) Image sequences of 96 subjects mapped into the embedding space described by the first three coordinates of LPP.

should be a universal space which provides a unified subject-independent representation of the expression manifold. In this universal space, manifolds of different subjects are aligned in a way that the images from different subjects with semantic similarity are mapped to the same region. Here we propose to align manifolds of different subjects in a universal space, and derive a generalized expression manifold for a large number of subjects.

As shown in Figure 4.1 and Figure 4.2, every image sequence representing facial expression with increasing intensity is embedded as a curve on the manifold, from the neutral face to a typical expression. If we define a global coordinate space, in which different typical expressions (including neutral faces and six basic expressions) from multiple subjects are well clustered and separated, the image sequences from different subjects with the same expression will be embedded as curves between the same two clusters: neutral faces and the typical expression. In this way, the manifolds of different subjects will be aligned on a generalized manifold. So the problem of aligning manifolds is reduced to deriving a universal discriminant subspace from static images of typical expressions.

For the data set containing images of typical expressions from different subjects, as appearance varies greatly across different subjects, there is significant overlapping among different

expression classes. The original LPP, which performs in an unsupervised manner, fails to embed the data set into a discriminant subspace in which different classes are well clustered and separated (see Section 4.5.1 for details). This demonstrates that a representation derived from unsupervised learning is not always optimal for discrimination. The problem is shared by most existing manifold learning techniques that focus on unsupervised learning of a compact low-dimensional representation of the input data space. To address this problem in the case of LLE, Ridder et al. [132] introduced a supervised LLE for improved classification. However, this supervised method still does not provide explicit mapping from the input space to the reduced space. Although it was suggested in [67] that LPP could be performed either unsupervised or supervised, no explicit supervised LPP has been given in the published work. By incorporating a *priori* knowledge about class into LPP, we present here a Supervised LPP (SLPP) to learn discriminant subspace of facial expressions.

The principle of SLPP is to preserve class information when constructing a neighborhood graph such that the local neighborhood of a sample x_i from class c is composed of samples from class c only. This can be achieved by increasing the distances between samples belonging to different classes, but leaving them unchanged if they are from the same class. Let $Dis(i, j)$ denote the distance between x_i and x_j, the distance after incorporating class information is

$$SupDis(i, j) = Dis(i, j) + \alpha M \delta(i, j) \qquad \alpha \in [0, 1] \tag{4.8}$$

where $M = \max_{i,j} Dis(i, j)$, and $\delta(i, j) = 1$ if x_i and x_j belong to different classes, and 0 otherwise. SLPP introduces an additional parameter α to quantify the degree of supervised learning. When $\alpha = 0$, one obtains the unsupervised LPP; when $\alpha = 1$, the result is the fully supervised LPP. For the fully supervised LPP, distances between samples in different classes will be larger than the maximum distance in the entire data set. This implies that neighbors of a sample will always be picked from the same class. Varying α between 0 and 1 gives a partially supervised LPP, where an embedding is found by introducing some separation between classes. We summarize the SLPP algorithm in Figure 4.5. SLPP utilizes the class information to construct a better graph with respect to discriminant information, so encodes

1. **Preprocessing**: The training set $\{x_i\}$ is first projected into a PCA subspace, to ensure matrix XDX^T is nonsingular, We denote the transformation matrix of PCA as W_{PCA}.

2. **Constructing the supervised adjacency graph**: Let G denotes a graph with m nodes, an edge is put between nodes i and j if x_i and x_j are "close" (ϵ-neighborhoods or k-nearest neighbors). The distance between nodes is $SupDis$ defined by Eqn. (4.8).

3. **Choosing weights**: S is a sparse symmetric matrix with S_{ij} having the weight of the edge joining nodes i and j, and 0 if there is no such edge. Two variations for weighting edges are shown in Eqn. (4.2) and Eqn. (4.3).

4. **Eigenmaps**: Let the vectors $\mathbf{w}_0, \ldots, \mathbf{w}_{l-1}$ be the solution to Eqn.(4.6), ordered according to their eigenvalues, the embedding is as follows:

$$x \to y = W^T x, \tag{4.9}$$

$$W = W_{PCA} W_{SLPP}, \quad \text{and} \quad W_{SLPP} = [\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{l-1}] \tag{4.10}$$

Figure 4.5: The algorithmic procedure of SLPP.

more discriminative power than the original LPP for improved classification capability.

We selected neutral face and typical expressions at its apex of every sequence to build a data set of 7-class basic expression images. SLPP ($\alpha = 1$) was explored to embed the data set into a subspace shown in Figure 4.6. We can observe that different expressions are well clustered and separated in the derived subspace (see Section 4.5.1 for more evaluation). This subspace provides global coordinates for expression manifolds of different subjects. Figures 4.7 and 4.8 plot aligned manifolds in the universal space. Compared to the aligned manifolds in Figures 4.3 and 4.4, the manifolds of different subjects are aligned on a generalized manifold, which provides a unified representation of the expression manifold[1].

Although only image sequences of basic expressions are discussed here, the generalized expression manifold provides a global semantic representation for all possible facial expressions. For example, blends of expression will lie between the curves of basic expressions, so can be analyzed with reference to the basic curves. Expression intensity can also be quantified easily on the manifold. Therefore, more effective facial expression analysis can be facilitated using

---

[1] A video demonstration manifold_align.avi is available at *http://www.dcs.qmul.ac.uk/~cfshan/demos.*
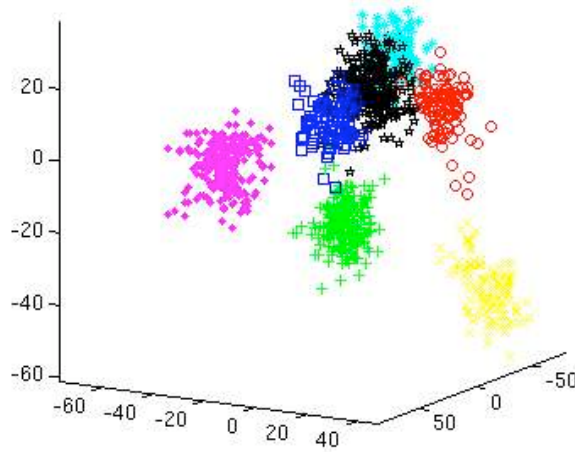
Figure 4.6: (Best viewed in color) The global coordinate space derived by SLPP from images of typical expressions. Different expressions are coded as: Anger (red circle), Disgust (yellow x-mark), Fear (blue square), Joy (magenta point), Sadness (cyan star), Surprise (green plus), Neutral (black pentagram).
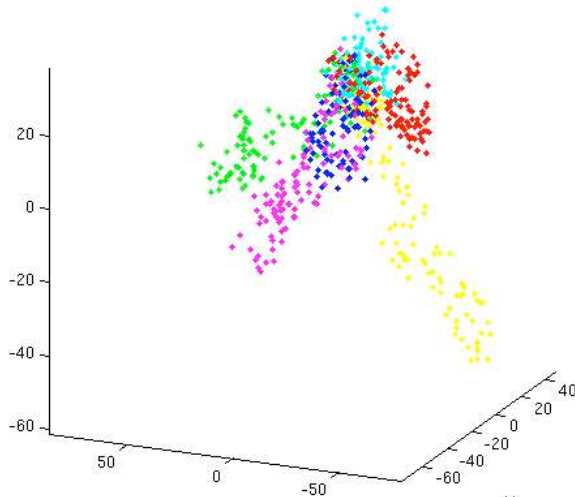
Figure 4.7: (Best viewed in color) The aligned manifolds of the six subjects.

the generalized manifold.

4.3 Dynamic Expression Recognition

Anderson and McOwan [5] recently presented a real-time automated system for facial expression recognition, where SVMs were used to classify expressions using motion signatures pro-
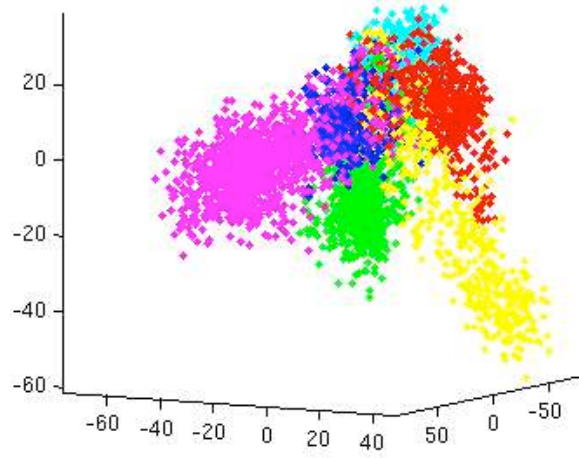
Figure 4.8: (Best viewed in color) The aligned manifolds of 96 subjects.
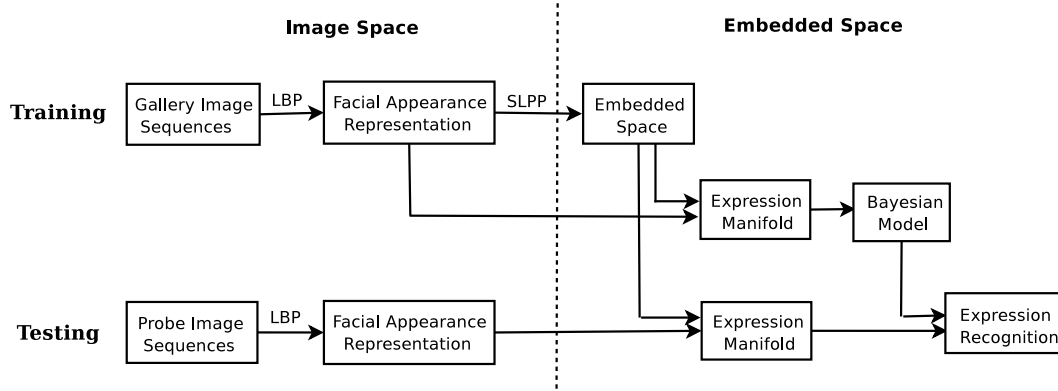
Figure 4.9: Manifold based dynamic facial expression recognition.

duced by optimal flow analysis. They did not model facial expression dynamics, and the SVM classifier performed recognition on a batch of data, e.g. four consecutive frames. Motivated by Bayesian tracking such as particle filters, in this section we formulate a Bayesian temporal model to exploit temporal information in the embedded manifold. Our Bayesian model can robustly recognize each frame based accumulated observation, which is more attractive than batch processing for online expression recognition. Figure 4.9 shows the framework of our recognition approach.

Given a probe image sequence mapped into the universal embedding subspace $Z_t, t = 0, 1, 2, ...$, the labeling of its corresponding facial expression class can be represented as a temporally accumulated posterior probability at time t, $p(X_t|Z_{0:t})$, where the state variable

X represents the class label of a facial expression. If we consider seven basic expression classes including Neutral, Anger, Disgust, Fear, Joy, Sadness and Surprise, $X = x_i, i = 1, ..., 7$. From a Bayesian perspective,

$$p(X_t|Z_{0:t}) = \frac{p(Z_t|X_t)p(X_t|Z_{0:t-1})}{p(Z_t|Z_{0:t-1})} \tag{4.11}$$

where

$$p(X_t|Z_{0:t-1}) = \int p(X_t|X_{t-1})p(X_{t-1}|Z_{0:t-1}) \, dX_{t-1} \tag{4.12}$$

Hence

$$p(X_t|Z_{0:t}) = \int p(X_{t-1}|Z_{0:t-1}) \frac{p(Z_t|X_t)p(X_t|X_{t-1})}{p(Z_t|Z_{0:t-1})} \, dX_{t-1} \tag{4.13}$$

Note in Eqn.(4.12), we use the Markov property to derive $p(X_t|X_{t-1}, Z_{0:t-1}) = p(X_t|X_{t-1})$. So the problem is reduced to how to estimate the prior $p(X_0|Z_0)$, the transition model $p(X_t|X_{t-1})$, and the observation model $p(Z_t|X_t)$.

The prior $p(X_0|Z_0) \equiv p(X_0)$ can be learned from a gallery of expression image sequences. The expression class transition probability from time $t-1$ to t given by $p(X_t|X_{t-1})$ can be estimated as

$$p(X_t|X_{t-1}) = p(X_t = x_j|X_{t-1} = x_i) = \begin{cases} \varepsilon & T_{i,j} = 0 \\ \alpha T_{i,j} & \text{otherwise} \end{cases} \tag{4.14}$$

where ε is a small empirical number we set between 0.02 - 0.05 typically, α is a scale coefficient, and $T_{i,j}$ is a transition frequency measure, defined by

$$T_{i,j} = \sum I(X_{t-1} = x_i \text{ and } X_t = x_j) \qquad i = 1, ..., 7, j = 1, ..., 7$$

where

$$I(A) = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases}$$

$T_{i,j}$ can be easily estimated from the gallery of image sequences. Parameters ε and α are selected such that $\sum_j p(x_j|x_i) = 1$.

The expression manifold derived by SLPP preserves optimally local neighborhood infor-

mation in the data space. To take the advantage of such a locality preserving structure, we define a likelihood function $p(Z_t|X_t)$ according to statistical analysis of local neighborhood. For example, given an observed frame Z_t, if there are more neighbors labeled as "Anger" (we denote "Anger" as x_1), there is less ambiguity for the observation Z_t to be classified as "Anger", so the observation has a higher value for $p(Z_t|X_t = x_1)$.

More precisely, let $\{N_j, j = 1, ..., k\}$ be the k-nearest neighbors of frame Z_t, we compute a neighborhood distribution measure as

$$M_i = \sum_j I(N_j = x_i) \qquad j = 1, ..., k, i = 1, ..., 7$$

and further define a probability distribution as

$$p_i = \begin{cases} \tau & M_i = 0 \\ \beta M_i & \text{otherwise} \end{cases} \qquad i = 1, ..., 7$$

where τ is a small empirical number and is set between 0.05 - 0.1 typically, β is a scale coefficient. Parameters τ and β are selected such that $\sum_i p_i = 1$. A likelihood function $p(Z_t|X_t)$ is then defined as

$$p(Z_t|X_t) = p(Z_t|X_t = x_i) = p_i \tag{4.15}$$

Given the prior $p(X_0)$, the transition model $p(X_t|X_{t-1})$, and the likelihood function $p(Z_t|X_t)$, the posterior $p(X_t|Z_{0:t})$ can be computed straightforwardly using Eqn.(4.13). This provides us with a probability distribution measure of all seven candidate expression classes in the current frame, given an input image sequence. This Bayesian temporal model exploits the expression dynamics represented in the expression manifold, so potentially it provides better recognition performance and improved robustness against the static frame-based model.

4.4 Expression Intensity Estimation

Facial expressions vary in intensity. A distinct emotional state of an expresser can not be correctly perceived unless the expression intensity exceeds a certain level. Methods that work for intense expressions may generalize poorly to subtle expressions with low intensity. It has been experimentally shown that the expression decoding accuracy and the perceived intensity of the underlying affective state vary linearly with the physical intensity of a facial display [68]. Explicit analysis of expression intensity variation is also essential for distinguishing between spontaneous and posed facial behavior [115]. So expression intensity estimation is necessary and helpful for accurate assessment of facial expressions. It is desirable to decide the expression intensity from the face data without human labeling. Some methods have been presented to automatically quantify expression intensity variation in emotional expressions [83, 4] and action units [94].

Here we describe a method for expression intensity estimation using the derived manifold. As shown in Figures 4.7 and 4.8, expression intensity variation can be extracted from the expression manifold, where image sequences are embedded as continuous curves; the stretching direction indicates the expression type and the distance from the origin (neutral faces) shows the degree of expressions. So it is intuitive to estimate the expression intensity from its location in the embedded space. We adopt an unsupervised clustering technique to automatically generate the spectrum of the expression intensity. The expression intensity variation is gradual instead of abrupt, so disjoint classes poorly fit this problem. An approach with fuzzy classes seems more appropriate. Specifically, we employ the Fuzzy K-Means (FKM) clustering algorithm [18]. Compared to the K-Means clustering algorithm, KFM's strength is that it yields the data's membership in each of clusters.

Fuzzy K-Means is based on the minimization of the following objective function:

$$J = \sum_{i=1}^{m} \sum_{j=1}^{c} u_{ij}^{\phi} d^2(x_i, c_j) \qquad 1 \le \phi < \infty \qquad (4.16)$$

where m is the number of data, c is the number of clusters (or classes), $u_{ij} \in [0, 1]$ is the

degree of membership of data x_i in cluster j, c_j is the centroid of cluster j, $d(x_i, c_j)$ is the distance from data x_i to centroid c_j. ϕ is the fuzzy exponent, which determines the degree of fuzziness of the resulting clusters, that is, the degree of overlap between classes. In our study, ϕ was set to 2.0. Our experimental results on expression intensity estimation are presented in Section 4.5.3.

4.5 Experiments

In order to derive the generalized expression manifold, we first adopt the SLPP algorithm to learn a discriminant universal subspace from static images of prototypic expressions. In this section, we first evaluate SLPP and other linear subspace methods in facial expression subspace learning, and then present facial expression recognition and expression intensity estimation on the learned expression manifold.

4.5.1 Facial Expression Subspace Learning

Appearance-based linear subspace analysis is one of successful approaches to facial expression recognition [44]. Recently a number of graph-based linear subspace techniques have been proposed. However, these techniques have not been investigated for the task of facial expression analysis, and it still unknown which technique is most suitable for expression subspace learning. Here we extensively evaluate and compare different linear subspace methods, which include traditional PCA and LDA, and the recent proposed LPP, SLPP, Orthogonal Neighborhood Preserving Projections (ONPP) [84], and Locality Sensitive Discriminant Analysis (LSDA) [24]. We use implementations of LPP[1], ONPP and LSDA provided by their authors. On facial feature representation, in addition to LBP features extracted from equally divided sub-regions, we also consider raw image data (IMG) and Boosted LBP features (BLBP).

Four data sets were constructed: (1) **S1**: 320 image sequences were selected from the Cohn-Kanade Database. The sequences come from 96 subjects, with 1 to 6 emotions per subject. The three peak frames of each sequence were used for 6-class expression analysis, resulting

---

[1]http://people.cs.uchicago.edu/~xiaofei/LPP.m

Data Set	Images	Subjects	Expressions
S1	960	96	six
S2	1280	96	seven
S3	384	20	seven
S4	213	10	seven

Table 4.1: Four data sets for facial expression subspace analysis.

in 960 images (108 Anger, 120 Disgust, 99 Fear, 282 Joy, 126 Sadness, and 225 Surprise). (2) **S2**: the neutral face of each sequence was further included for 7-class expression analysis, resulting in 1,280 images (960 emotional images plus 320 neutral faces). (3) **S3**: 96 image sequences were selected from the MMI Database. The sequences come from 20 subjects, with 1 to 6 emotions per subject. The neutral face and three peak frames of each sequence were used (384 images in total). (4) **S4**: all 213 images of the JAFFE database were used. The four data sets are summarized in Table 4.1.

Comparative Evaluation on Subspace Learning

In the six methods examined, PCA and LPP are unsupervised techniques, while LDA, SLPP, ONPP, and LDSA perform in a supervised manner. The 2D visualization of embedded subspaces are shown in Figures 4.10 to 4.13. It is evident that the classes of different expressions are heavily overlapped in 2D subspaces generated by unsupervised methods PCA and LPP (with all three facial representations), therefore are poorly represented. The projections of PCA are spread out since PCA aims at maximizing the variance. In the cases of LPP, although it preserves local neighborhood information, as expression images contain complex variations and significant overlap among different classes, it is difficult for LPP to yield meaningful projections in the absence of class information. For supervised methods, it is surprising to observe that different expressions are still heavily overlapped in the 2D subspace derived by ONPP. In contrast, the supervised methods LDA, SLPP and LSDA yield much meaningful projections since images of the same class are mapped close to each other. SLPP provides evidently the best projection since different classes are well separated and the clusters appear cohesive. This is because SLPP preserves the locality and class information simultaneously in the projections. On the other hand, LDA discovers only the Euclidean structure therefore
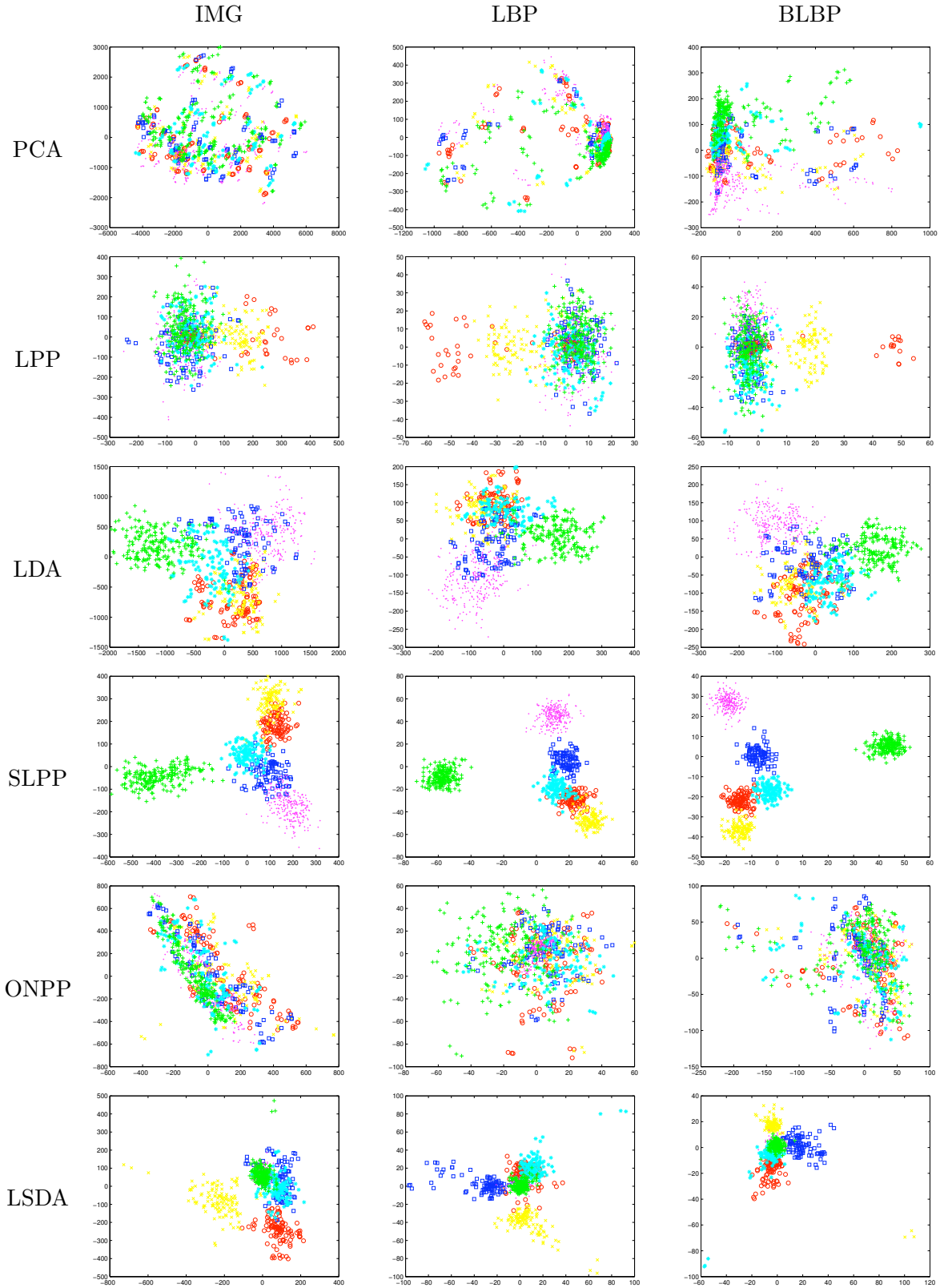
Figure 4.10: (Best viewed in color) Images of data set **S1** are mapped into 2D embedding spaces.
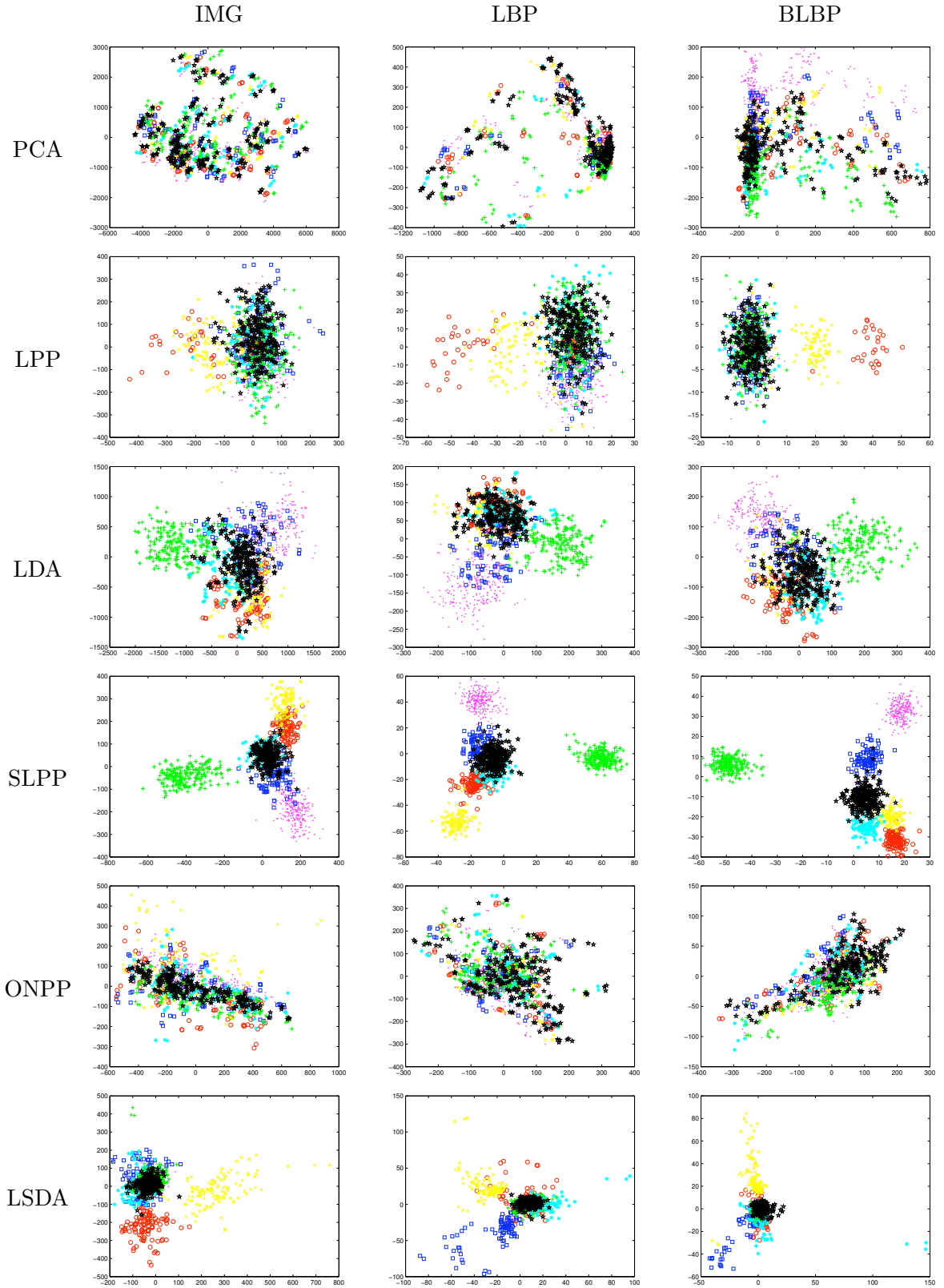
Figure 4.11: (Best viewed in color) Images of data set **S2** are mapped into 2D embedding spaces. Neutral expression is color coded as black.
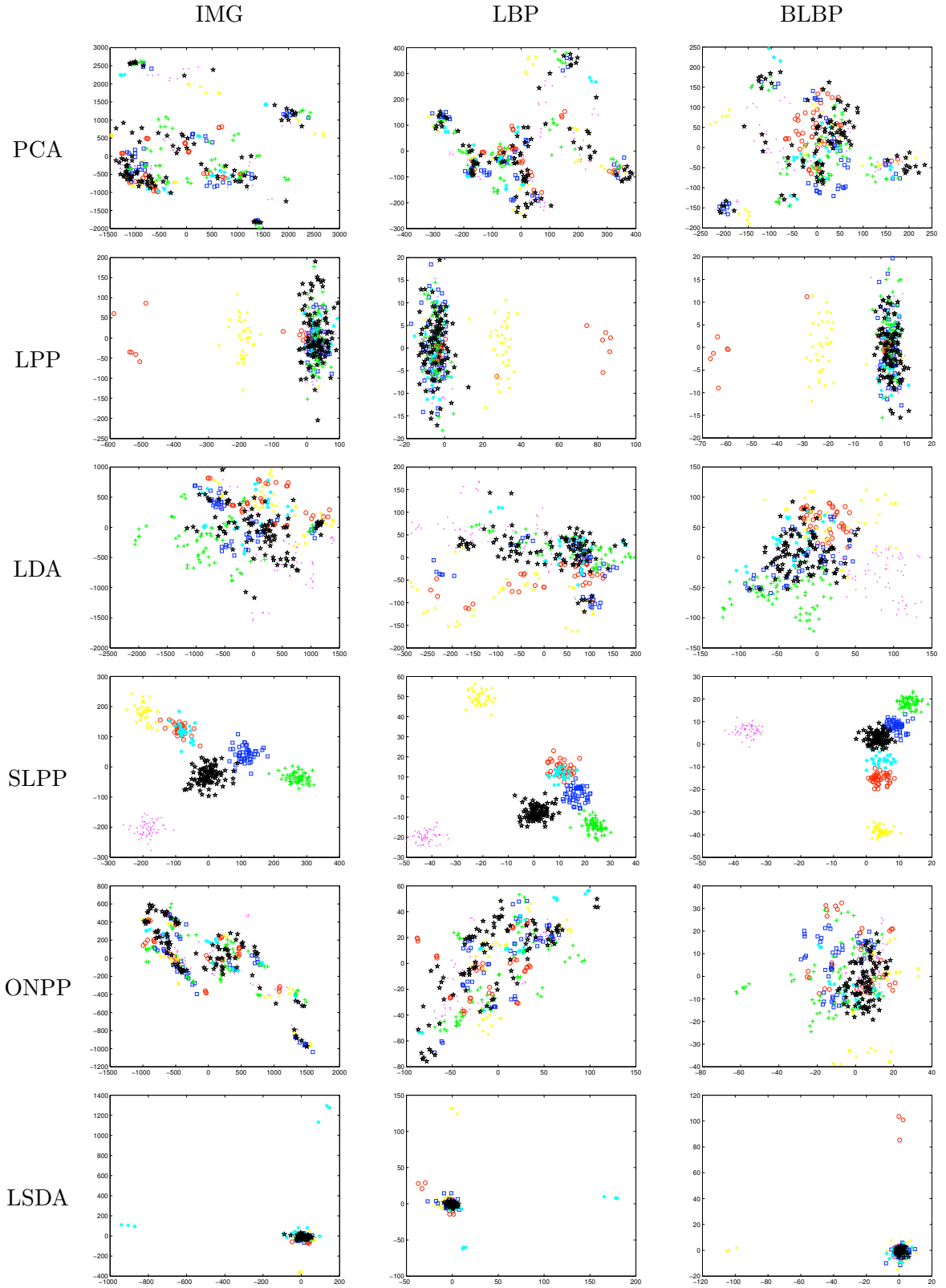
Figure 4.12: (Best viewed in color) Images of data set **S3** are mapped into 2D embedding spaces.
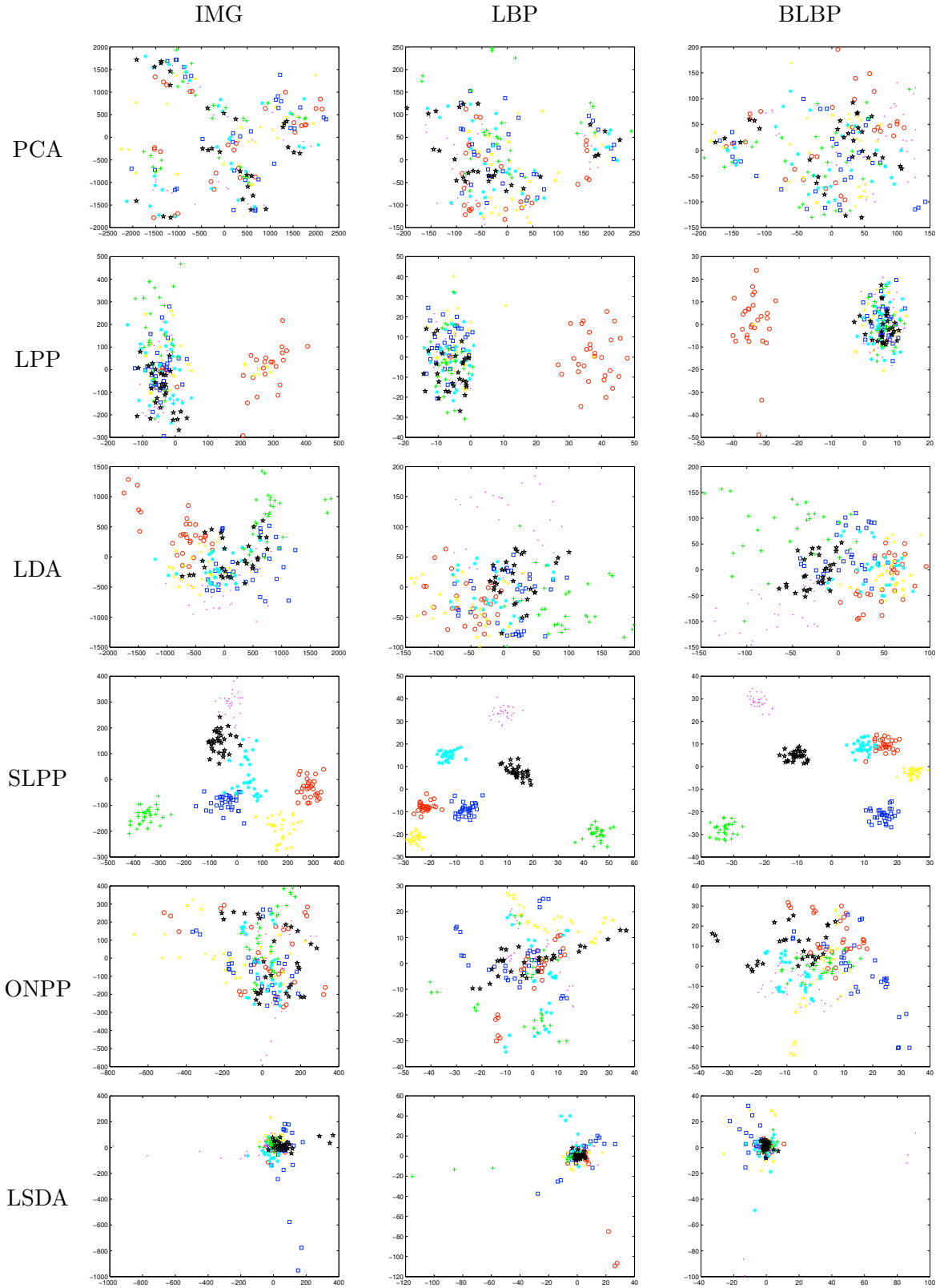
Figure 4.13: (Best viewed in color) Images of data set **S4** are mapped into 2D embedding spaces.

fails to capture accurately any underlying nonlinear manifold that expression images lie on, resulting in its discriminating power being limited. LSDA obtains better projections than LDA on datasets **S1** and **S2**. The results obtained by SLPP also reflect human observation that Joy and Surprise can be clearly separated, but Anger, Disgust, Fear and Sadness are easily confused. This reenforces the findings of other published work [152]. Notice that in the SLPP subspace, after including neutral faces, Anger, Disgust, Fear, Sadness, and Neutral are easily confused, while Joy and Surprise still can be clearly separated. On comparing facial representation, BLBP provides evidently the best performance with projected classes more cohesive and clearly separable in the SLPP subspace, and IMG is worst.

For a quantitative evaluation of the derived subspaces, following Li *et al.* 's methodology [92], we investigate the histogram distribution of within-class pattern distance and between-class pattern distance of different techniques. The former is the distance between expression patterns of the same expression type, while the latter is the distance between expression patterns belonging to different expression types. Obviously, for a good representation, the within-class distance distribution should be dense, close to the origin, having a high peak value, and well-separated from the between-class distance distribution. We plot in Figure 4.14 the results of different methods on **S1**. It is observed that SLPP consistently provides the best distributions for different facial representations, while those of PCA, LPP, and ONPP are worst. The average within-class distance d_w and between-class distance d_b are shown in Table 4.2. To ensure the distance measures from different methods are comparable, we compute a normalized difference between the within- and between-class distances of each method as $dif = \frac{d_b - d_w}{d_w}$, which can be regarded as a relative measure on how widely the within-class patterns are separated from the between-class patterns. A high value of this measure indicates success. It is evident in Table 4.2 that SLPP has the best separating power whilst PCA, LPP and ONPP are the poorest. The separating power of LDA and LSDA is inferior to that of SLPP, but always outperform those of PCA, LPP, and ONPP. Both Figure 4.14 and Table 4.2 reinforce the observation in Figures 4.10 to 4.13.
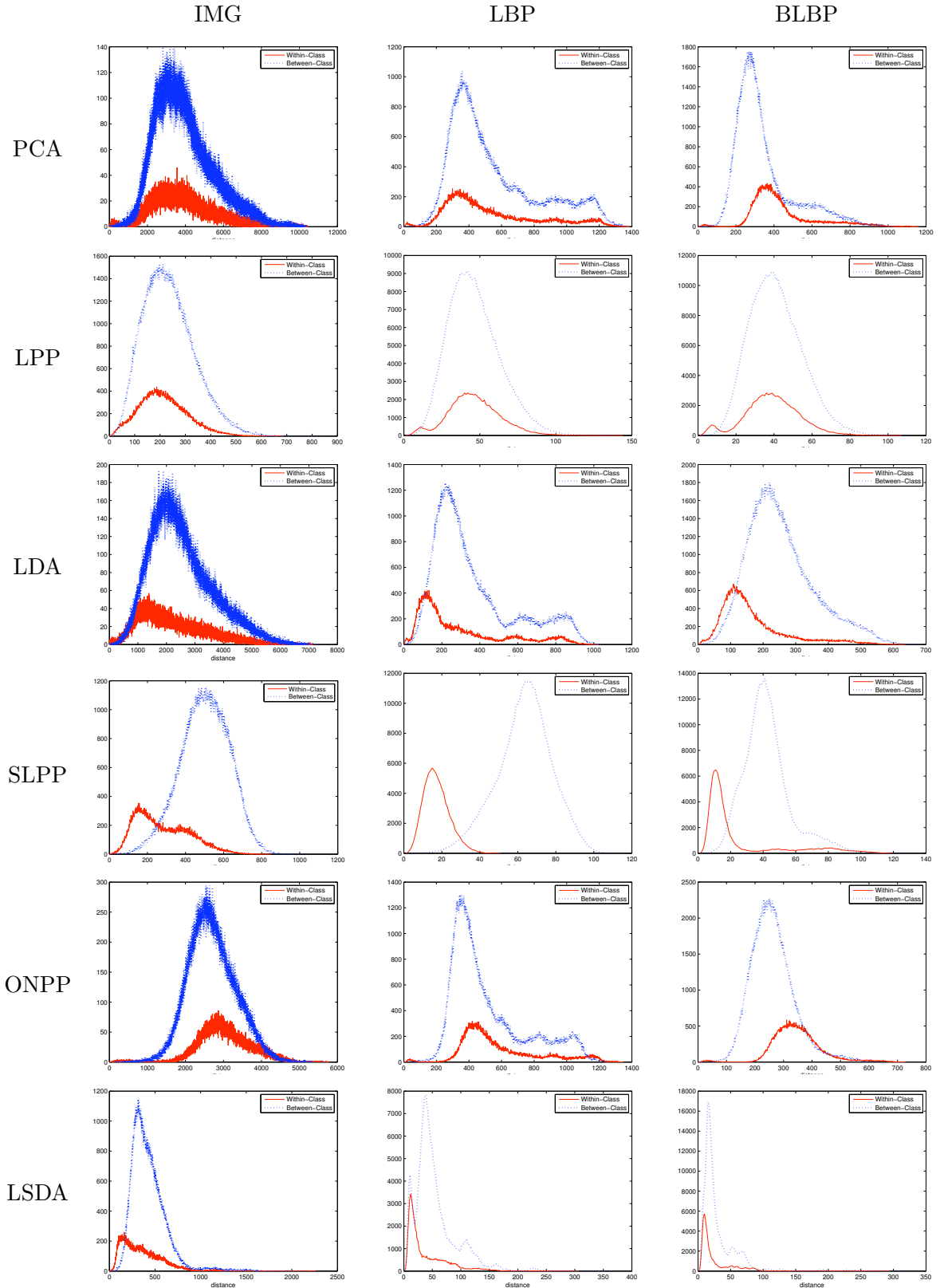
Figure 4.14: (Best viewed in color) Histogram distribution of within-class pattern distance (solid red lines) and between-class pattern distances (dotted blue line) on data set **S1**

	IMG			LBP			BLBP		
	d_w	d_b	dif	d_w	d_b	dif	d_w	d_b	dif
PCA	3921.8	4219.9	*0.0760*	542.7	591.3	*0.0897*	480.12	532.55	*0.1092*
LPP	216.3	241.4	*0.1164*	48.7	52.2	*0.0723*	42.304	46.016	*0.0877*
LDA	2195.3	2684.3	*0.2228*	288.5	377.1	*0.3071*	170.01	279.44	*0.6436*
SLPP	284.6	587.4	***1.0636***	18.1	86.4	***3.7741***	26.032	68.837	***1.6443***
ONPP	3069.8	3317.8	*0.0808*	633.3	673.5	*0.0636*	396.07	442.80	*0.1180*
LSDA	361.9	561.7	*0.5520*	44.3	76.1	*0.7189*	29.1	45.6	*0.5665*

Table 4.2: The average within-class and between-class distance and their normalization difference values on data set **S1**.

Comparative Evaluation on Expression Recognition

To further compare different methods, we also performed image-based expression recognition in the derived subspace. We adopted a nearest-neighbor classifier for its simplicity. The Euclidean metric was used as the distance measure. The number of nearest neighbors was set according to the size of the training set. To evaluate the algorithms' generalization ability, we adopted a 10-fold cross-validation test scheme. We show the average recognition results (with the standard deviation) here.

The recognition performance of these subspace learning techniques varies with the dimensionality of subspace (Note that the dimension of the reduced subspace of LDA is at most $c - 1$, where c is the number of classes). Moreover, the graph-based techniques rely on the parameter k, the number of nearest neighbors, and how to set the parameter is still an open problem. In our cross-validation experiments, we tested different combinations of parameter k with the subspace dimensionality, and the best performance obtained are shown in Tables 4.3 to 4.6. It is evident that supervised methods outperform the unsupervised ones. For unsupervised methods, PCA performs better than LPP, with all three facial features. It is observed that SLPP consistently produces the best recognition performance, having a clear margin of superiority over LDA (12-38% better on **S1** and **S2**, 19-50% better on **S3**, 14-38% better on **S4** and LSDA (6-13% better on **S1** and **S2**, 16-33% better on **S3**, 22-38% better on **S4**). On datasets **S1** and **S2**, LSDA and LDA perform better than ONPP, and LSDA outperforms LDA. The recognition results reinforce our early observations shown in Figures 4.10 to 4.14 and Table 4.2.

	PCA (%)	LPP (%)	LDA (%)	SLPP (%)	ONPP (%)	LSDA (%)
IMG	49.3±7.1	48.9±6.2	67.5±7.6	**86.4**±5.0	61.9±6.8	80.3±5.2
LBP	49.6±8.9	42.6±5.5	73.4±7.1	**90.4**±2.9	63.1±9.4	82.0±4.8
BLBP	74.9±9.1	48.9±9.0	84.2±6.7	**94.7**±3.5	75.9±6.9	87.2±5.4

Table 4.3: Averaged recognition rates (with the standard deviation) of 6-class facial expression recognition on data set **S1**.

	PCA (%)	LPP (%)	LDA (%)	SLPP (%)	ONPP (%)	LSDA (%)
IMG	41.4±7.0	37.9±8.4	59.5±5.4	**82.2**±6.2	50.0±6.8	77.3±5.7
LBP	50.6±6.4	36.6±5.0	67.9±5.5	**87.1**±3.0	57.8±9.4	76.9±6.5
BLBP	65.9±8.8	44.0±7.3	75.9±5.8	**92.0**±3.9	68.3±6.9	82.3±5.2

Table 4.4: Averaged recognition rates (with the standard deviation) of 7-class facial expression recognition on data set **S2**.

Recognition performance on **S4** is much poorer than that on **S1**, **S2** and **S3**, as there are fewer images in the data set resulting in a poor sampling of the underlying latent space. The effect of the training set size is also reflected on the standard deviation of 10-fold cross-validation. The standard deviations of **S4** are much larger than those of **S1**, **S2** and **S3**, and the standard deviations of **S3** are larger than those of **S1** and **S2** as well. So the recognition performance of linear subspace methods on the small training sets is not robust and reliable. On comparing the standard deviation of 10-fold cross validation on **S1** and **S2**, SLPP always produce the smallest deviation. This demonstrates that SLPP is much more robust than other methods on relatively large data sets.

To clearly compare recognition rates of different methods with different facial representations, we plot bar graphes of recognition rates in Figure 4.15. On comparing feature representation, it is clearly observed that, in most cases, BLBP features perform better than LBP and IMG features and LBP have better or comparable performance with IMG features.

We show in Figure 4.16 the averaged recognition rates versus dimensionality reduction by different subspace schemes using BLBP features. As the dimension of the reduced LDA subspace is at most $c - 1$, where c is the number of classes, we plot only the best achieved recognition rate by LDA across various values of the dimension of subspace. We observe that SLPP outperforms the other methods. The performance difference between SLPP and LDA is conspicuous when the dimension of subspace is small. But when the dimension increases,

	PCA (%)	LPP (%)	LDA (%)	SLPP (%)	ONPP (%)	LSDA (%)
IMG	54.3±17.2	38.8±8.7	55.0±15.2	**81.2**±10.1	54.3±16.1	62.2±13.6
LBP	59.5±16.6	38.8±10.7	63.7±14.1	**80.7**±9.1	60.2±15.4	69.6±8.4
BLBP	60.2±15.3	35.1±7.3	71.4±11.7	**84.6**±8.8	66.2±12.1	63.5±8.7

Table 4.5: Averaged recognition rates (with the standard deviation) of 7-class facial expression recognition on data set **S3**.

	PCA (%)	LPP (%)	LDA (%)	SLPP (%)	ONPP (%)	LSDA (%)
IMG	39.5±6.3	31.6±14.5	51.2±11.0	**69.2**±15.7	48.5±12.2	56.3±22.8
LBP	51.8±13.2	31.0±10.5	56.4±16.1	**72.4**±18.1	54.2±16.0	52.3±16.6
BLBP	62.6±17.5	32.4±9.8	65.4±15.5	**74.2**±16.1	66.8±13.7	54.9±14.5

Table 4.6: Averaged recognition rates (with the standard deviation) of 7-class facial expression recognition on data set **S4**.
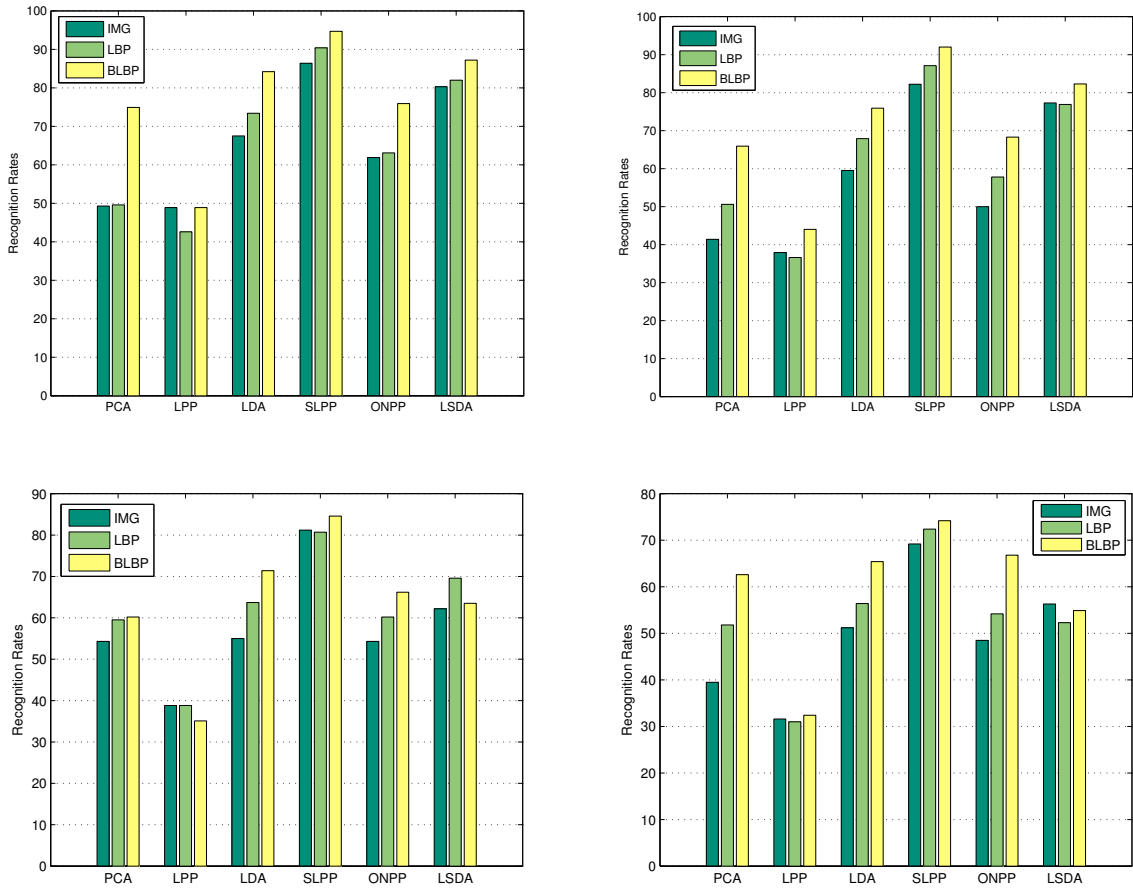
Figure 4.15: Comparison of recognition rates using different subspace methods with different features. From top to bottom, from left to right: **S1**, **S2**, **S3** and **S4**.
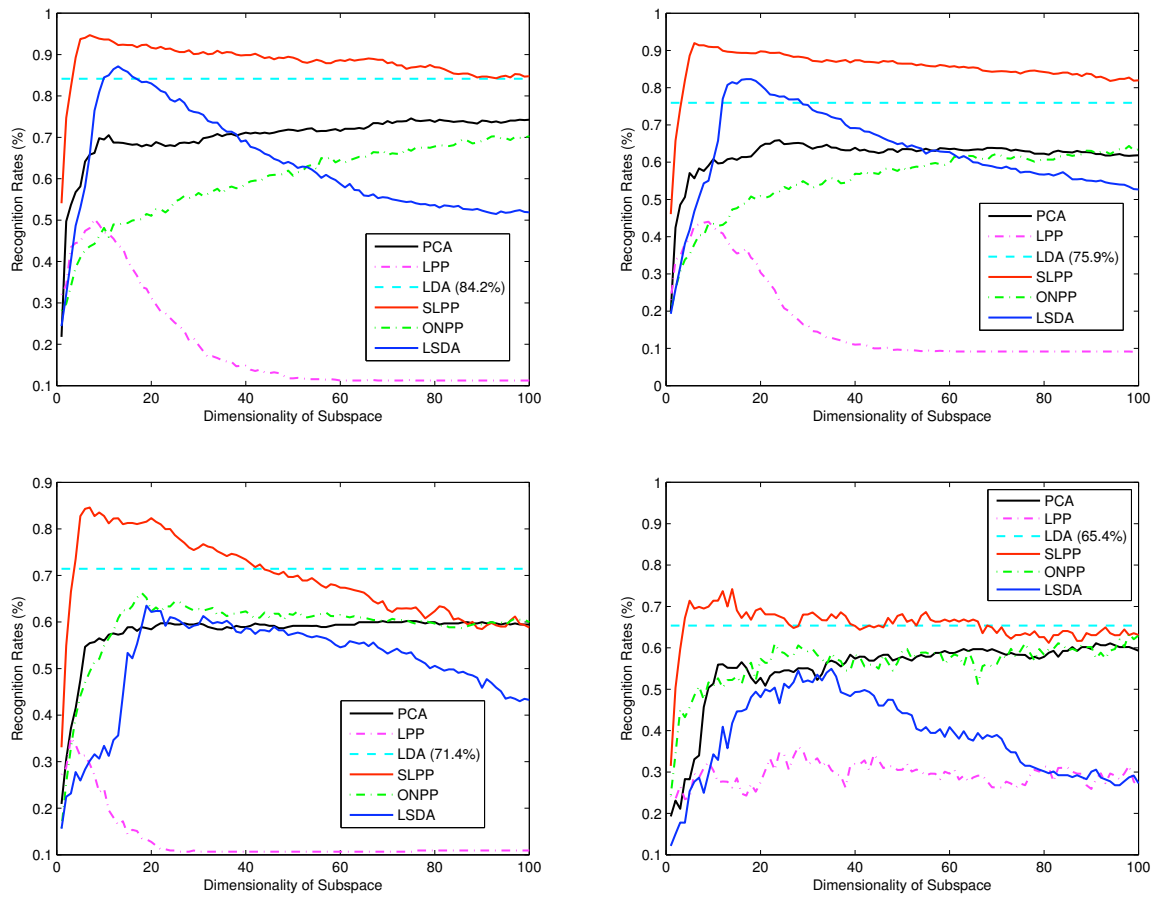
Figure 4.16: (Best viewed in color) Averaged recognition accuracy versus dimensionality reduction (with BoostLBP features). From top to bottom, from left to right: **S1**, **S2**, **S3** and **S4**.

their performances become rather similar. The performances of PCA, LPP, and ONPP is inferior to that of LDA consistently across all values of the subspace dimension. LSDA has similar trend with SLPP, but much worse performance. The performance of PCA and ONPP eventually become stable and similar when the dimension increases. On the other hand, the performance of LPP degrades when the dimension increases, and is the worst overall. The plots for **S4** shows greater variations compared to those of **S1** and **S2**. This may also be due to the small size of the training set.

The best result of 94.7% in 6-class facial expression recognition on the Cohn-Kanade database, achieved by BLBP based SLPP, is to our best knowledge the best recognition

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Anger	85.2%	2.8%	0	0	7.4%	0	4.6%
Disgust	0	97.5%	2.5%	0	0	0	0
Fear	0	0	81.8%	11.1%	1.0%	1.0%	5.1%
Joy	0	0	0	97.5%	0	0	2.5%
Sadness	4.0%	0	0.8%	0	84.9%	0	10.3%
Surprise	0	0	2.7%	0	0	96.9%	0.4%
Neutral	0.8%	0	0.4%	2.8%	5.2%	0.4%	90.4%

Table 4.7: Confusion matrix of 7-class expression recognition on data set **S2**.

rate reported so far on the database in the published literature. Previously Tian achieved 94% performance using a three-layer neural networks when combining geometric features and Gabor wavelet features [152]. With regard to 7-class facial expression recognition, BLBP based SLPP achieves the best performance of 92.0%, which is also very encouraging given that previously published 7-class recognition performance on this database were 81-83% [27]. The confusion matrix of 7-class facial expression in data set **S2** is shown in Table 4.7, which illustrates that most confusion occurs between Anger, Fear, Sadness, and Neutral.

4.5.2 Dynamic Expression Recognition

The above experiments clearly show that SLPP is superior in deriving the discriminant expression subspace. In this section, we perform dynamic facial expression recognition using the proposed Bayesian temporal manifold model (BTMM). A total of 316 image sequences of basic expressions were selected from the database. The sequences come from 96 subjects, with 1 to 6 emotions per subject. To test our approach's generalization to novel subjects, we partitioned the 316 image sequences randomly into ten groups of roughly equal numbers of subjects, and adopted 10-fold cross-validation to perform manifold learning and expression recognition. Figure 4.17 shows the learned manifold from one of the trials. The left sub-figure displays the embedded manifold of the gallery image sequences, while the right sub-figure shows the embedded results of the probe image sequences. (NOTE: for the sake of illustration, in Figure 4.17, each image sequence is labeled by one color. However, in our recognition experiments, the latent expression state is not fixed for each sequence, i.e., there are expression variations in each sequence, for example, from the neutral face to the emotional
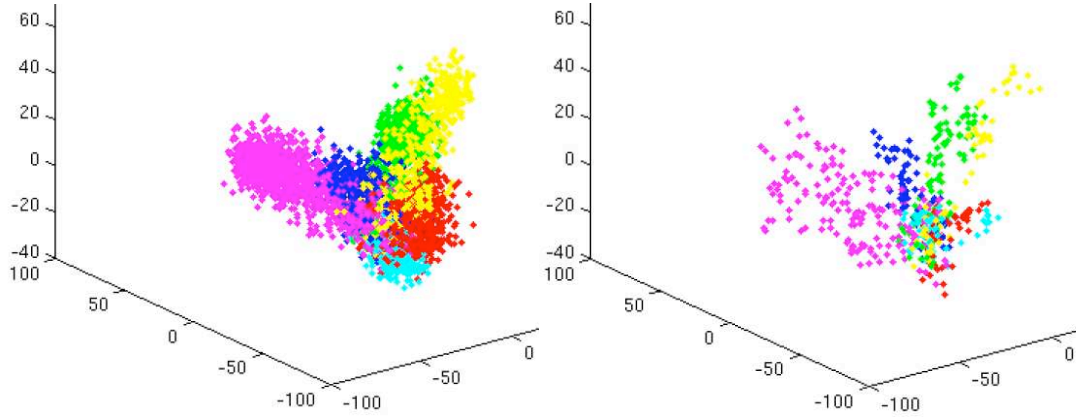
Figure 4.17: (Best viewed in color) Image sequences are mapped into the learned manifold space. *Left*: the gallery set; *Right*: the probe set.

expression.)

We compared BTMM with another three methods for expression recognition on the learned expression manifold. (1) k-NN: to verify the benefit of exploiting temporal information in recognition, we also adopted a k-NN classifier to recognize each frame based on static frame features. k was set as 11 in our experiments. (2) Bayesian: we further used a common Bayesian model [28] to perform recognition on static frame features, which computes the posterior probability of class labels given the observed frame features, and then classify the features with the most probable class label. (3) HMM: HMMs are effective for the modeling of temporal behaviours, so we also performed recognition using HMMs. The HMM we used is a fully ergodic model based on Gaussian emission probabilities having diagonal covariance matrix for each state. The parameters of the model (the emission probability density and the state transition matrix) were learned from the training data using Maximum Likelihood.

We show the average frame-level recognition results of different methods in Table 4.8. Since there is no clear boundary between neutral faces and typical expressions in sequences, we manually labeled neutral faces, which introduced some noise in our recognition. We observe that by exploiting the temporal information, BTMM and HMM provide superior performance to static frame features based k-NN and Bayesian model. Moreover, BTMM slightly outperforms HMM in this case.

	Overall	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
BTMM	83.1%	70.5%	78.5%	44.0%	94.5%	55.0%	94.6%	90.7%
HMM	81.7%	63.2%	80.7%	58.7%	85.7%	55.1%	90.8%	92.9%
k-NN	79.0%	66.1%	77.6%	51.3%	88.6%	54.4%	90.0%	81.7%
Bayesian	79.8%	58.1%	79.6%	58.2%	88.7%	61.8%	91.0%	82.0%

Table 4.8: Frame-level facial expression recognition on the Cohn-Kanade database.

We also performed sequence-level expression recognition by following the frame-level recognition by a voting scheme, which classifies each sequence according to the most common expression detected in the sequence. We report the average recognition rates in Table 4.9, which shows that BTMM also provides the best performance. We also compared our model to Yeasin et al.' two-stage approach [176]: k-NN classifiers were first used on consecutive frames to produce characteristic temporal signature, then HMMs were used to model the temporal signatures of each expression. They obtained the average result of 90.9% in 5-fold cross-validation on the Cohn-Kanade database. They also used k-NN classifier followed by a voting scheme, and achieved the performance of 75.3%. The comparisons summarized in Table 4.9 illustrate that our approach provides superior performance. In Figures 4.18 to 4.19, we present some examples of facial expression recognition in live image sequences. We plotted the probability distribution for example sequences. The recognition results consistently confirm that the dynamic aspect of our BTMM approach can lead to a more robust expression recognition in image sequences[1].

Method	Average Recognition Performance
BTMM	91.8%
HMM	88.9%
k-NN	86.3%
Bayesian	87.6%
k-NN based HMM [176]	90.9%
k-NN [176]	75.3%

Table 4.9: Sequence-level facial expression recognition on the Cohn-Kanade database.

---

[1]A video demonstration manifold_rcg.avi is available at *http://www.dcs.qmul.ac.uk/∼cfshan/demos.*

Anger

Disgust
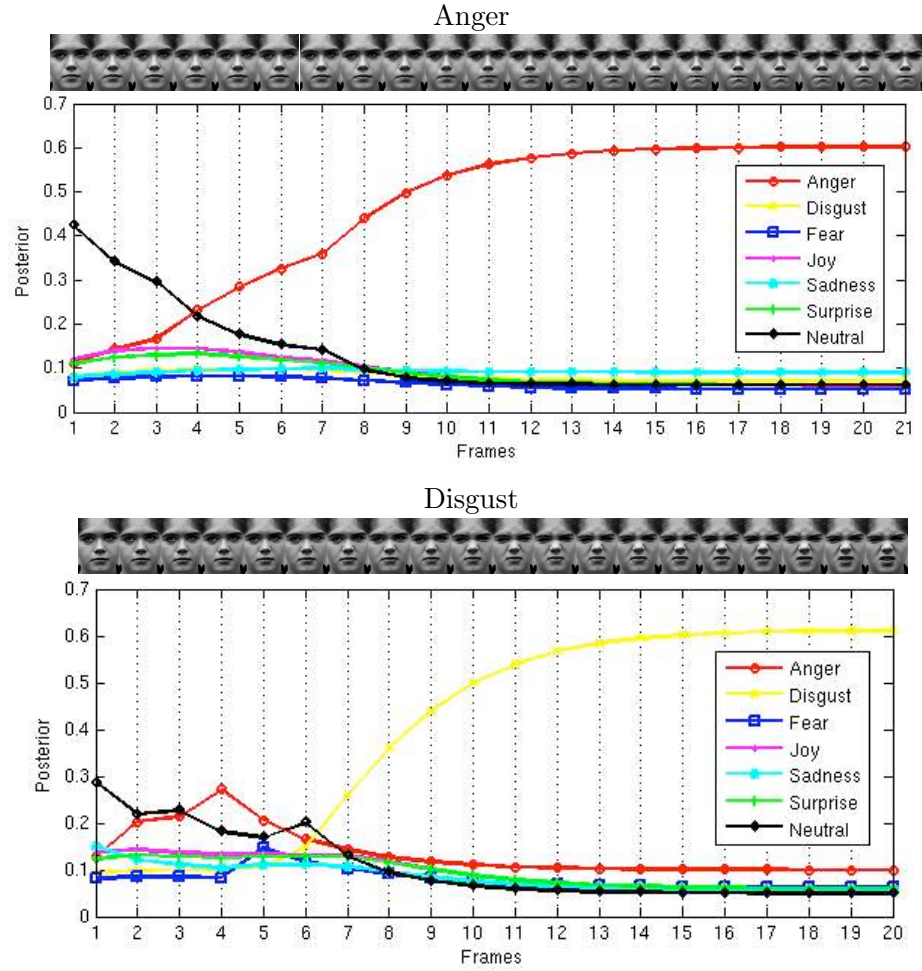
Figure 4.18: (Best viewed in color) Expression recognition using BTMM on two example image sequences.

4.5.3 Expression Intensity Estimation

We adopt a 3-grade intensity scale (*Low*, *Moderate*, and *High*) to describe intensity variation of facial expression. Fuzzy K-Means were applied on the training set to derive three fuzzy clusters in the embedded space, and then the expression intensity of test images was estimated by classifying them to different clusters, where the cluster memberships were mapped to expression degrees. We show the estimation results of some test sequences in Figures 4.21 to 4.20. we observe an orderly mapping to clusters as the face moves from the neutral expression to a typical expression at apex. In the FKM clustering, the degree of *Moderate* is much transient than that of *Low* and *High*. Our experiments demonstrate that the expression

Joy

Surprise
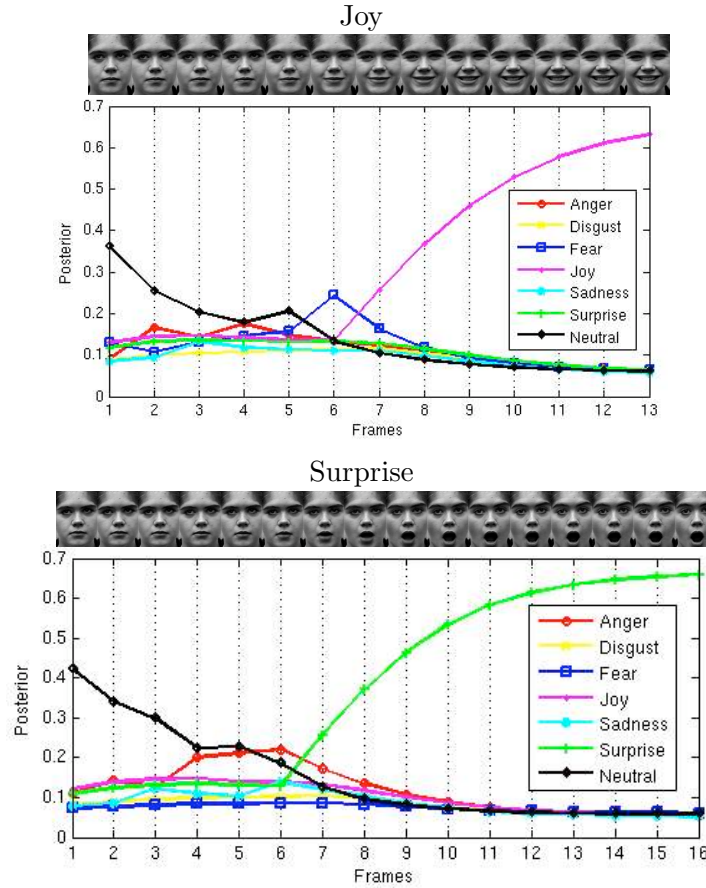
Figure 4.19: (Best viewed in color) Expression recognition using BTMM on two example image sequences.

intensity can be quantified in the expression manifold.

4.6 Summary

We introduce an appearance-based approach to capture and present facial expression dynamics by discovering the underlying low-dimensional manifold. Extensive experiments demonstrate our manifold-based representation facilitates facial expression analysis such as dynamic expression recognition and expression intensity estimation. Compared with the existing works on manifold-based facial expression analysis, our work enjoys several favorable properties:

1. We learn expression manifold in a dense appearance feature space, so that the detailed facial deformations that are important to facial expression modeling can be better
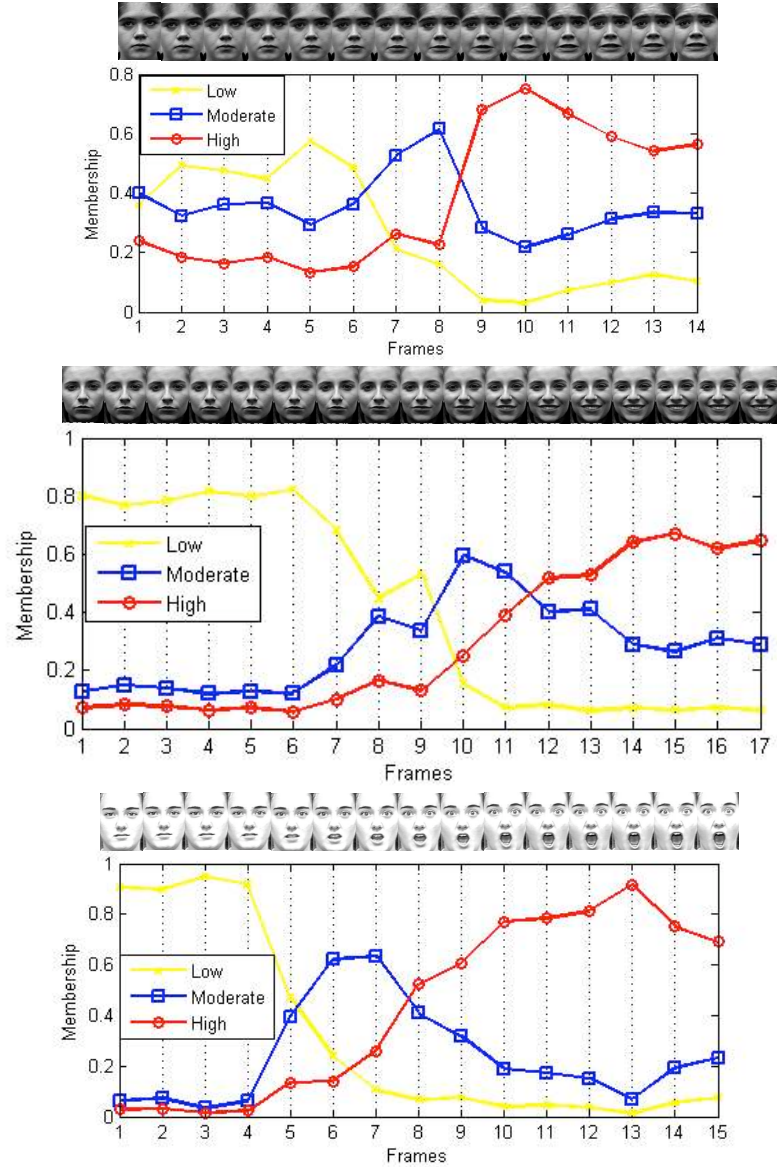
Figure 4.20: (Best viewed in color) Expression intensity estimation on three example image sequences.

described.

2. We present a method to align manifolds of different subjects in a discriminant universal subspace. Moreover, the derived generalized expression manifold was verified on a large number of subjects.

3. By employing the derived manifold representation, we present a Bayesian temporal
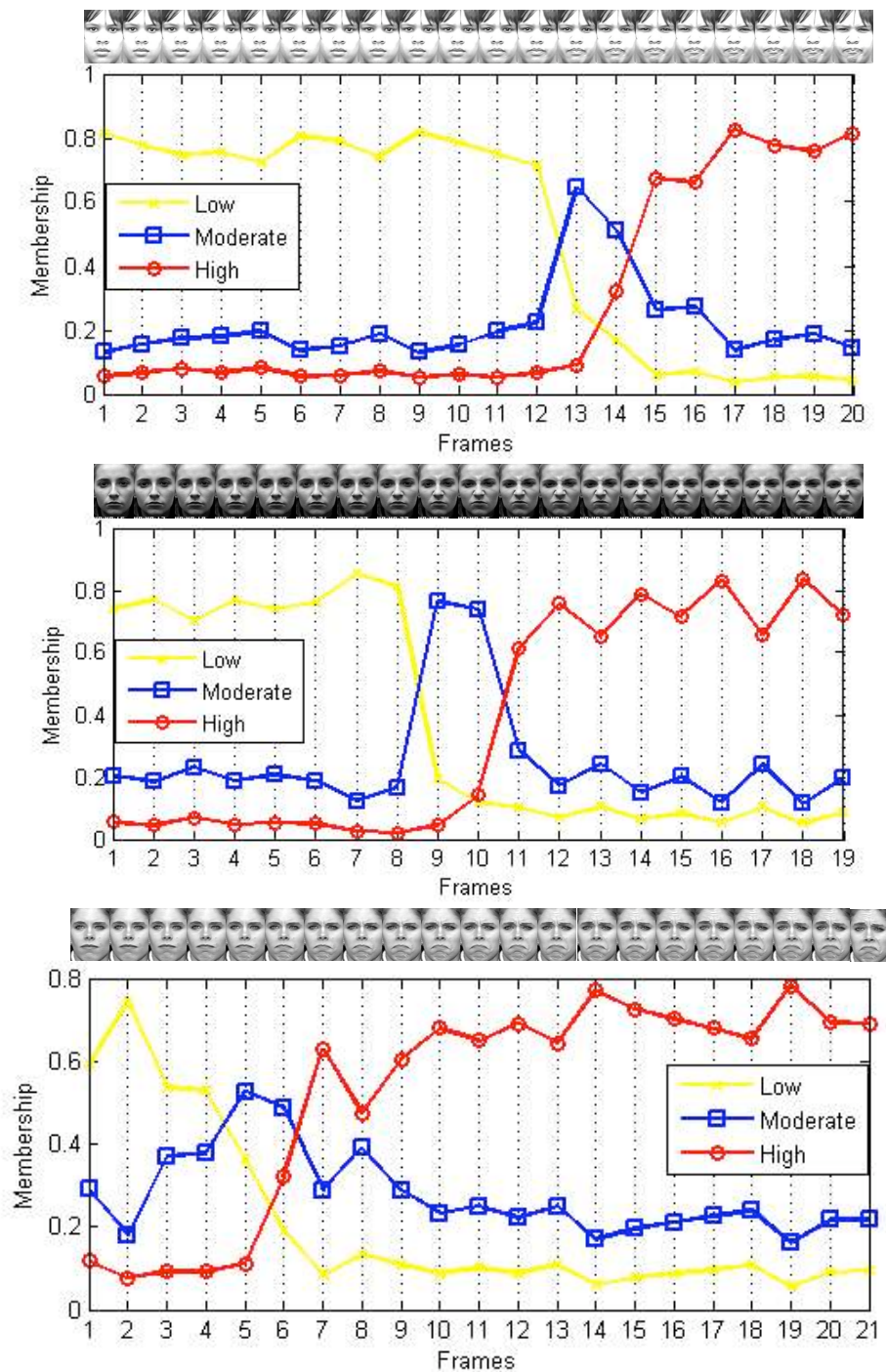
Figure 4.21: (Best viewed in color) Expression intensity estimation on three example image sequences.

model for dynamic facial expression recognition, and estimate the expression intensity using the Fuzzy K-Means.

As facial muscles are contracted in unison to display facial expressions, different facial parts have strong spatial-temporal correlations. Analyzing the correlations among facial parts is important for modeling facial expressions precisely. Our manifold based representation only captures temporal correlations of the whole face image, and does not consider the correlations among different facial regions. To have a closer look at the correlations among facial parts, in next chapter, we investigate correlations among facial parts employing a statistical technique.

5 Correlation Analysis of Facial Parts

Capturing and analyzing the correlations among facial parts is important for modeling facial expressions precisely. Most of the existing work does not explicitly model these correlations. In this chapter, we employ Canonical Correlation Analysis (CCA) [70], a statistical technique that is well suited for relating two sets of signals, to model correlations of facial parts.

When applying CCA to image data, the original two-dimensional images have to be reshaped into one-dimensional vectors, as the traditional CCA is based on the vector-space model. However, this matrix-to-vector operation leads to some problems. For example, the intrinsic 2D structure of image matrices is removed, so the spatial information stored therein is discarded. To address these problems, we introduce a novel Matrix-based Canonical Correlation Analysis (MCCA) for better correlation analysis of 2D image or matrix data in general. MCCA takes a 2D matrix based data representation rather than the 1D vector based representation in classical CCA. MCCA seeks canonical factors in two dimensions to maximize the correlations between two sets of matrices. Unlike classical CCA, there is no closed-form solution for the optimization problem in MCCA. Instead, we propose an iterative solution with a convergence proof.

5.1 Canonical Correlation Analysis

CCA was developed by H. Hotelling [70] for measuring linear relationships between two vector variables. It finds pairs of base vectors (i.e. canonical factors) for two variables such that the correlations between the projections of the variables onto these canonical factors are mutually maximized. The directions of canonical factors capture functional relations of the two variables.

Given two zero-mean random variables $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, CCA finds pairs of directions $\mathbf{w}_x$ and $\mathbf{w}_y$ that maximize the correlation between the projections $x = \mathbf{w}_x^T \mathbf{x}$ and $y = \mathbf{w}_y^T \mathbf{y}$ (x and y are called *canonical variates*). More formally, CCA maximizes the function:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^T \mathbf{x}\mathbf{y}^T \mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^T \mathbf{x}\mathbf{x}^T \mathbf{w}_x]E[\mathbf{w}_y^T \mathbf{y}\mathbf{y}^T \mathbf{w}_y]}}$$
$$= \frac{\mathbf{w}_x^T \mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx}\mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy}\mathbf{w}_y}} \tag{5.1}$$

where $\mathbf{C}_{xx} \in \mathbb{R}^{m \times m}$ and $\mathbf{C}_{yy} \in \mathbb{R}^{n \times n}$ are the *within-set covariance matrices* of $\mathbf{x}$ and $\mathbf{y}$, respectively, while $\mathbf{C}_{xy} \in \mathbb{R}^{m \times n}$ denotes their *between-sets covariance matrix*. A number of at most $k = \min(m, n)$ canonical factor pairs $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ can be obtained by successively solving $\arg\max_{\mathbf{w}_x^i, \mathbf{w}_y^i}\{\rho\}$ subject to $\rho(\mathbf{w}_x^j, \mathbf{w}_x^i) = \rho(\mathbf{w}_y^j, \mathbf{w}_y^i) = 0$ for $j = 1, \ldots, i - 1$, i.e., the next pair of $\langle \mathbf{w}_x, \mathbf{w}_y \rangle$ are orthogonal to the previous ones.

The maximization problem can be solved by setting the derivatives of Eqn. (5.1), with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, equal to zero, resulting in eigenvalue equations as:

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x = \rho^2 \mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y = \rho^2 \mathbf{w}_y \end{cases} \tag{5.2}$$

Matrix inversions need to be performed in Eqn. (5.2), leading to numerical instability if $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are rank deficient. Alternatively, $\mathbf{w}_x$ and $\mathbf{w}_y$ can be obtained by computing principal angles (see [82] for details), as CCA is the statistical interpretation of principal angles between two linear subspace [57].

Recently CCA has been applied to computer vision and pattern recognition problems [19, 102, 65, 82, 45]. Borga [19] adopted CCA to find corresponding points in stereo images. Melzer *et al.* [102] applied CCA to model the relation between an object's poses with raw brightness images for appearance-based 3D pose estimation. Hardoon *et al.* [65] presented a method using CCA to learn a semantic representation to web images and their associated text. CCA has also been used for image set matching [82], where each set is represented by a linear subspace and principal angles (canonical correlations) between two subspaces are

exploited as a similarity measure. Like PCA and LDA, CCA also reduces the dimensionality of original variables, since only a few factor pairs are normally needed to represent the relevant information. However, they serve different purposes: whilst PCA aims to minimize the reconstruction error and LDA derives a discriminant function that maximizes between-class scatter and minimize within-class scatter, CCA seeks directions for two sets of variables to maximize their correlations, so it is better suited for regression tasks. It has been evidently shown that CCA outperforms PCA for regression tasks [102]. Compared to other linear regression methods such as Partial Least Squares and Multivariate Linear Regression, CCA has some attractive properties. For example, it is invariant to affine transformations of the input variables [45]. Donner *et al.* [45] introduced a fast Active Appearance Model search algorithm, which uses reduce-rank regression estimates obtained by CCA, instead of standard linear least-square regression estimates. Reiter *et al.* [131] recently presented to predict 3D depth maps of faces and near-infrared face texture from color face images using CCA.

5.2 Matrix-based Canonical Correlation Analysis

In the existing work, when applying CCA to image data, the original two-dimensional images have to be reshaped into one-dimensional vectors. However, this matrix-to-vector operation leads to two main problems. Firstly, the intrinsic 2D structure of image matrices is removed, so the spatial information stored therein is discarded. CCA based on these vectors can not fully capture correlations among the original 2D image data. Secondly, each image sample is modeled as a high-dimensional vector so that a large number of training samples is needed to yield a reliable estimation of the underlying data distribution. However, in reality, limited number of training data is usually available. Actually these problems are shared by other subspace methods such as PCA and LDA. Recently some methods have been proposed to extend these vector-based methods to 2D matrices or high-order tensors [174, 175, 37, 173]. However, all these existing matrix-based methods were developed for learning in one set of variables, and they are not suited for measuring relationships between two set of variables.

To address these problems, we introduce a novel Matrix-based Canonical Correlation Anal-

ysis (MCCA) for better correlation analysis of 2D image or matrix data in general. MCCA takes a 2D matrix based data representation rather than the 1D vector based representation in classical CCA. So the collection of data is represented as a set of matrices, instead of a single large matrix. We notice that more recently Zou *et al.* [185] introduced a 2DCCA by simply replacing the image vector with image matrix in computing the variance matrices. Their approach is different to ours both in concept and algorithmic design. Moreover, they addressed the correlations between image sets and their label matrices, instead of two sets of images.

Given two matrix variables $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{j \times k}$ (we assume the variables are both zero-mean), MCCA finds pairs of directions $\mathbf{v}_a \in \mathbb{R}^m$, $\mathbf{w}_a \in \mathbb{R}^n$, $\mathbf{v}_b \in \mathbb{R}^j$ and $\mathbf{w}_b \in \mathbb{R}^k$ that maximize the correlation between the projections $a = \mathbf{v}_a^T \mathbf{A} \mathbf{w}_a$ and $b = \mathbf{v}_b^T \mathbf{B} \mathbf{w}_b$. Mathematically, we can formulate this as the following maximization problem: find optimal $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ that maximize

$$\rho = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^T \mathbf{A} \mathbf{w}_a \mathbf{w}_b^T \mathbf{B}^T \mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T \mathbf{A} \mathbf{w}_a \mathbf{w}_a^T \mathbf{A}^T \mathbf{v}_a]E[\mathbf{v}_b^T \mathbf{B} \mathbf{w}_b \mathbf{w}_b^T \mathbf{B}^T \mathbf{v}_b]}} \tag{5.3}$$

Here $\mathbf{v}_a$ ($\mathbf{v}_b$) and $\mathbf{w}_a$ ($\mathbf{w}_b$) are canonical factors in two dimensions, acting as a two-sided linear transformation on the data in matrix form. To our knowledge, there is no closed-form solution for the maximization problem in Eqn. (5.3). A key observation, which leads to an iterative algorithm for the computation of $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$, is stated in the following Lemma:

Lemma 1 *Let $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ be the optimal solution to the maximization problem in Eqn. (5.3), then*

(1) Given $\mathbf{w}_a$ and $\mathbf{w}_b$, $\mathbf{v}_a$ and $\mathbf{v}_b$ can be obtained as canonical factors of two variables $\mathbf{a}' \in R^m$ and $\mathbf{b}' \in R^j$, where $\mathbf{a}' = \mathbf{A} \mathbf{w}_a$ and $\mathbf{b}' = \mathbf{B} \mathbf{w}_b$.

(2) Given $\mathbf{v}_a$ and $\mathbf{v}_b$, $\mathbf{w}_a$ and $\mathbf{w}_b$ can be obtained as canonical factors of two variables $\mathbf{a}'' \in R^n$ and $\mathbf{b}'' \in R^k$, where $\mathbf{a}'' = \mathbf{A}^T \mathbf{v}_a$ and $\mathbf{b}'' = \mathbf{B}^T \mathbf{v}_b$.

Proof (1) $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ maximize Eqn. (5.3), which can be rewritten as

$$\rho = \frac{E[\mathbf{v}_a^T \mathbf{a}' \mathbf{b}'^T \mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T \mathbf{a}' \mathbf{a}'^T \mathbf{v}_a] E[\mathbf{v}_b^T \mathbf{b}' \mathbf{b}'^T \mathbf{v}_b]}} \tag{5.4}$$

where $\mathbf{a}' = \mathbf{A}\mathbf{w}_a$ and $\mathbf{b}' = \mathbf{B}\mathbf{w}_b$. Hence, given $\mathbf{w}_a$ and $\mathbf{w}_b$, the maximum of Eqn. (5.4) is achieved by solving canonical correlation analysis on the variables $\mathbf{a}'$ and $\mathbf{b}'$ (by the definition of CCA in Eqn. (5.1)). So $\mathbf{v}_a$ and $\mathbf{v}_b$ can be obtained as canonical factors of $\mathbf{a}'$ and $\mathbf{b}'$.

(2) Similarly, Eqn. (5.3) can also be rewritten as

$$\rho = \frac{E[\mathbf{w}_a^T \mathbf{a}'' \mathbf{b}''^T \mathbf{w}_b]}{\sqrt{E[\mathbf{w}_a^T \mathbf{a}'' \mathbf{a}''^T \mathbf{w}_a] E[\mathbf{w}_b^T \mathbf{b}'' \mathbf{b}''^T \mathbf{w}_b]}} \tag{5.5}$$

where $\mathbf{a}'' = \mathbf{A}^T \mathbf{v}_a$ and $\mathbf{b}'' = \mathbf{B}^T \mathbf{v}_b$. Hence, given $\mathbf{v}_a$ and $\mathbf{v}_b$, the maximum of Eqn. (5.5) is achieved by solving canonical correlation analysis on the variables $\mathbf{a}''$ and $\mathbf{b}''$. So $\mathbf{w}_a$ and $\mathbf{w}_b$ can be obtained as canonical factors of $\mathbf{a}''$ and $\mathbf{b}''$. This completes the proof of the lemma.

By the above Lemma, we present an iterative procedure for computing $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ as follows: given the initial choice of $\mathbf{w}_a$ and $\mathbf{w}_b$, we can compute $\mathbf{v}_a$ and $\mathbf{v}_b$ by computing canonical factors of $\mathbf{a}'$ and $\mathbf{b}'$; with the computed $\mathbf{v}_a$ and $\mathbf{v}_b$ (corresponding to the largest canonical correlation), we can then compute $\mathbf{w}_a$ and $\mathbf{w}_b$ by computing canonical factors of $\mathbf{a}''$ and $\mathbf{b}''$, and $\mathbf{w}_a$ and $\mathbf{w}_b$ (corresponding to the largest canonical correlation) will be used in next iteration. The procedure can be repeated until convergence. In this way, a number of at most $q = \min(m, j)$ left-side canonical factor pairs $\langle \mathbf{v}_a^1, \mathbf{v}_b^1 \rangle, \ldots, \langle \mathbf{v}_a^q, \mathbf{v}_b^q \rangle$ and a number of at most $p = \min(n, k)$ right-side canonical factor pairs $\langle \mathbf{w}_a^1, \mathbf{w}_b^1 \rangle, \ldots, \langle \mathbf{w}_a^p, \mathbf{w}_b^p \rangle$ can be obtained. The pseudo-code of the above iterative procedure is given in Algorithm 3, where `CCA(a, b)` computes the canonical factors and canonical correlations of the variables $\mathbf{a}$ and $\mathbf{b}$.

5.2.1 Proof of Convergence

Since correlation coefficient ρ is bounded between -1 and 1 from its definition, we prove the convergence of MCCA by the following theorem:

Algorithm 3: MCCA

1 Obtain initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$ for $\mathbf{w}_a$ and $\mathbf{w}_b$, and set $\rho^{(0)} \leftarrow -1$ and $i \leftarrow 0$;

2 repeat

3 $i \leftarrow i + 1$;

4 $(\mathbf{v}_a^s, \mathbf{v}_b^s, \rho^s) \leftarrow \texttt{CCA}(\mathbf{A} * \mathbf{w}_a^{(i-1)}, \mathbf{B} * \mathbf{w}_b^{(i-1)})$;

     /* $s = 1, \ldots, q$ */

5 $\mathbf{v}_a^{(i)} \leftarrow \mathbf{v}_a^1$, $\mathbf{v}_b^{(i)} \leftarrow \mathbf{v}_b^1$, $\rho^{(i)} \leftarrow \rho^1$;

6 $(\mathbf{w}_a^t, \mathbf{w}_b^t, \rho^t) \leftarrow \texttt{CCA}(\mathbf{A}^T * \mathbf{v}_a^{(i)}, \mathbf{B}^T * \mathbf{v}_b^{(i)})$;

     /* $t = 1, \ldots, p$ */

7 $\mathbf{w}_a^{(i)} \leftarrow \mathbf{w}_a^1$, $\mathbf{w}_b^{(i)} \leftarrow \mathbf{w}_b^1$, $\rho^{(i)} \leftarrow \rho^1$;

8 until $\rho^{(i)} - \rho^{(i-1)} < \epsilon$;

9 $\mathbf{V}_a \leftarrow [\mathbf{v}_a^1, \ldots, \mathbf{v}_a^q]$, $\mathbf{V}_b \leftarrow [\mathbf{v}_b^1, \ldots, \mathbf{v}_b^q]$;

10 $\mathbf{W}_a \leftarrow [\mathbf{w}_a^1, \ldots, \mathbf{w}_a^p]$, $\mathbf{W}_b \leftarrow [\mathbf{w}_b^1, \ldots, \mathbf{w}_b^p]$;

Theorem 2 *The MCCA algorithm monotonically non-decreases the value of correlation coefficient ρ, hence it converges in the limit.*

Proof Given $\mathbf{w}_a^{(i-1)}$, $\mathbf{w}_b^{(i-1)}$ and $\rho^{(i-1)}$ obtained in Line 7, $\texttt{CCA}(\mathbf{A} * \mathbf{w}_a^{(i-1)}, \mathbf{B} * \mathbf{w}_b^{(i-1)})$ in Line 4 finds optimal $\mathbf{v}_a$ and $\mathbf{v}_b$ that maximize

$$
\begin{aligned}
\rho &= \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} \\
&= \frac{E[\mathbf{v}_a^T \mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T \mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_a^{(i-1)^T}\mathbf{A}^T\mathbf{v}_a]E[\mathbf{v}_b^T \mathbf{B}\mathbf{w}_b^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b]}}
\end{aligned}
\tag{5.6}
$$

The value of $\rho^{(i-1)}$ is derived as

$$
\begin{aligned}
\rho^{(i-1)} &= \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} \\
&= \frac{E[\mathbf{v}_a^{(i-1)^T} \mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i-1)}]}{\sqrt{E[\mathbf{v}_a^{(i-1)^T} \mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_a^{(i-1)^T}\mathbf{A}^T\mathbf{v}_a^{(i-1)}]E[\mathbf{v}_b^{(i-1)^T} \mathbf{B}\mathbf{w}_b^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i-1)}]}}
\end{aligned}
\tag{5.7}
$$

which is less or equal to the maximized canonical correlation that $\texttt{CCA}(\mathbf{A} * \mathbf{w}_a^{(i-1)}, \mathbf{B} * \mathbf{w}_b^{(i-1)})$ finds. So the derived $\rho^{(i)}$ in Line 5 is no less than $\rho^{(i-1)}$. With regard to the first iteration, given any initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$, the canonical correlation $\rho^{(1)}$ derived by $\texttt{CCA}(\mathbf{A} * \mathbf{w}_a^{(0)}, \mathbf{B} * \mathbf{w}_b^{(0)})$ is no less than -1 ($\rho^{(0)}$). Therefore, the update of ρ in Line 5 does not

decrease its value, since the computed ρ is locally optimal.

Similarly, given $\mathbf{v}_a^{(i)}$, $\mathbf{v}_b^{(i)}$ and $\rho^{(i)}$ obtained in Line 5, $\mathtt{CCA}(\mathbf{A}^T * \mathbf{v}_a^{(i)}, \mathbf{B}^T * \mathbf{v}_b^{(i)})$ in Line 6 finds optimal $\mathbf{w}_a$ and $\mathbf{w}_b$ that maximize

$$\rho = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b^{(i)}]}{\sqrt{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a\mathbf{w}_a^T\mathbf{A}^T\mathbf{v}_a^{(i)}]E[\mathbf{v}_b^{(i)^T}\mathbf{B}\mathbf{w}_b\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b^{(i)}]}} \tag{5.8}$$

The value of $\rho^{(i)}$ in Line 5 is derived as

$$\begin{aligned} \rho^{(i)} &= \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} \\ &= \frac{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i)}]}{\sqrt{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_a^{(i-1)^T}\mathbf{A}^T\mathbf{v}_a^{(i)}]E[\mathbf{v}_b^{(i)^T}\mathbf{B}\mathbf{w}_b^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i)}]}} \end{aligned} \tag{5.9}$$

which is less or equal to the maximized canonical correlation that $\mathtt{CCA}(\mathbf{A}^T * \mathbf{v}_a^{(i)}, \mathbf{B}^T * \mathbf{v}_b^{(i)})$ finds. So the update of ρ in Line 7 do not decrease its value too. Therefore, the MCCA optimization process monotonically non-decreases the ρ value, and converges in the limit. This completes the proof of the theorem.

The convergence of the MCCA algorithm can also be verified experimentally. We show some examples of iterative learning in Figure 5.1 and Figure 5.2, where each example demonstrates the learning on a different training set. We can observe that the value of ρ becomes stable after at most 20-30 iterations. We also found that any variation on the initial choice of $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$ has almost no effect on convergence (as observed in Figure 5.2). The fast and stable convergence keeps the training cost low.

5.2.2 Effect of the Initial Choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$.

Theoretically, our solution to MCCA is only locally optimal. This solution depends on the initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$. However, in practice this does not have any ill-effect. We conducted extensive experiments using different choices for $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$, and found that, for image datasets, MCCA always converges to a similar (if not identical) solution regardless of the initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$.
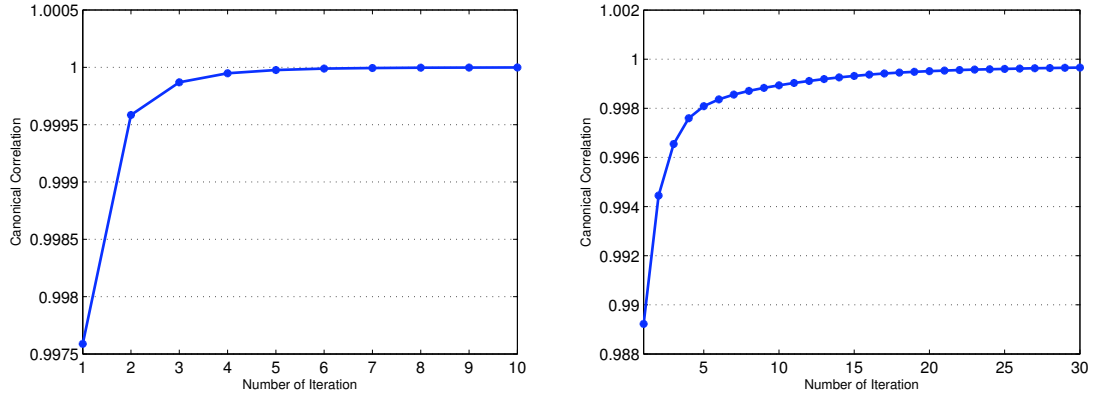
Figure 5.1: Convergence property of MCCA.
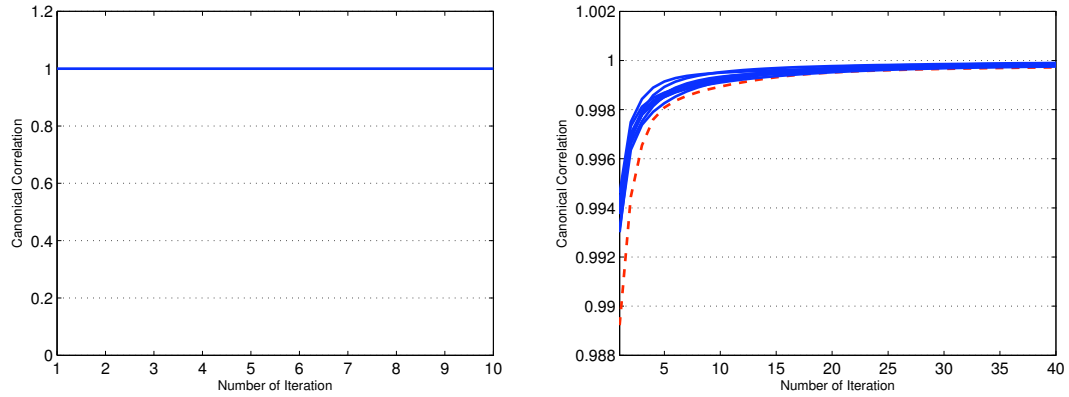
Figure 5.2: Sensitivity of MCCA to the initial choice $\mathbf{w}_a^0$ and $\mathbf{w}_b^0$: the ten solid curves correspond to the ten runs with random initializations, and the dash curve corresponds to $\mathbf{w}_a^0 = \mathbf{w}_b^0 = (1, 0, \dots, 0)^T$ (the dash curve in the *left* side is identical with solid curves, so is not visible).
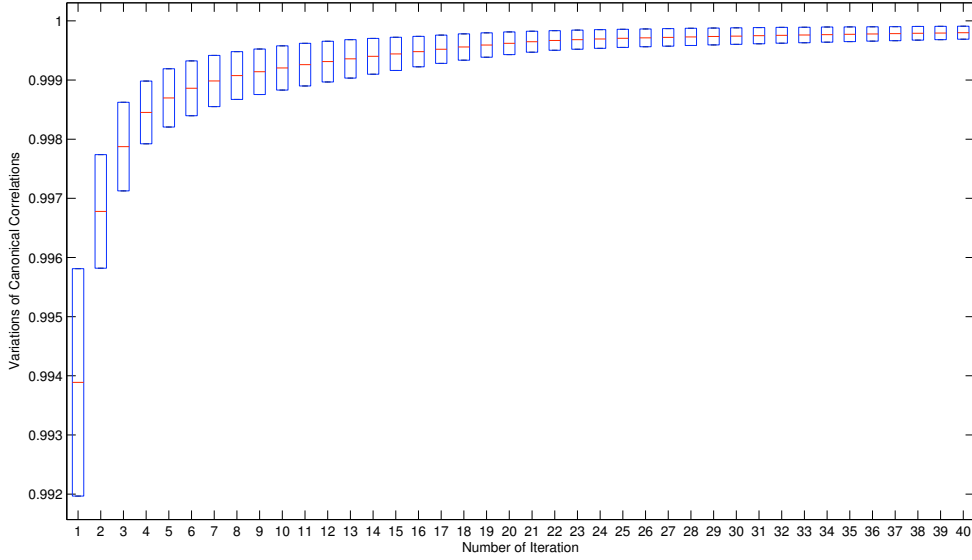
Figure 5.3: Variations of canonical correlation ρ when running MCCA with 100 randomly generated $\mathbf{w}_a^0$ and $\mathbf{w}_b^0$'s.

We show two typical results in Figure 5.2, where the horizontal axis is the number of iterations and the vertical axis is the value of ρ. Each sub-figure is the results on a different training set. We run MCCA with 10 randomly generated $\mathbf{w}_a^{(0)}$'s and $\mathbf{w}_b^{(0)}$'s, and another initialization $\mathbf{w}_a^{(0)} = \mathbf{w}_b^{(0)} = (1, 0, \ldots, 0)^T$. For the left sub-figure of Figure 5.2, we can observe that MCCA converges within two iterations for all eleven initial choices with the specified threshold ($\epsilon = 10^{-5}$), and also converges to the same solution. In the right sub-figure of Figure 5.2, MCCA converges slower. For all different initial choices, MCCA converges within 20-30 iterations with the threshold $\epsilon = 10^{-5}$, and converges to very similar solutions. The difference between the values of final ρ is very small ($< 1.6 \times 10^{-4}$). We further run MCCA with 100 randomly generated $\mathbf{w}_a^{(0)}$'s and $\mathbf{w}_b^{(0)}$'s on this training set, and plot in Figure 5.3 the variations of canonical correlation ρ at each iteration. The variation of the converged final ρ is less than 2×10^{-4}. These experiments demonstrate that, for image datasets, MCCA always converges to a similar (if not identical) solution regardless of the initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$. We used the initial choice $\mathbf{w}_a^{(0)} = \mathbf{w}_b^{(0)} = (1, 0, \ldots, 0)^T$ in our experiments (Section 5.3).

5.2.3 Generalization to High-Order Tensor Data

Although it is introduced for 2D matrix (second-order tensor) data, the proposed MCCA can be generalized for high-order tensor data in general. Given two nth-order tensor variables $\mathcal{A} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_n}$ and $\mathcal{B} \in \mathbb{R}^{k_1 \times k_2 \times \cdots \times k_n}$, Tensor-based CCA finds pairs of directions $\mathbf{w}_a^1 \in \mathbb{R}^{m_1}, \mathbf{w}_a^2 \in \mathbb{R}^{m_2}, \ldots, \mathbf{w}_a^n \in \mathbb{R}^{m_n}$ and $\mathbf{w}_b^1 \in \mathbb{R}^{k_1}, \mathbf{w}_b^2 \in \mathbb{R}^{k_2}, \ldots, \mathbf{w}_b^n \in \mathbb{R}^{k_n}$ that maximize the correlation between the projections $a = \mathcal{A} \times_1 \mathbf{w}_a^1 \times_2 \mathbf{w}_a^2 \cdots \times_n \mathbf{w}_a^n$ and $b = \mathcal{B} \times_1 \mathbf{w}_b^1 \times_2 \mathbf{w}_b^2 \cdots \times_n \mathbf{w}_b^n$, where $\times_n$ represents the n-mode product of a tensor by a matrix, and the n-mode product of a tensor $\mathcal{A} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ by a matrix $\mathbf{U} \in R^{J_n \times I_n}$ is defined as $(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{i_n} a_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} u_{j_n i_n}$ (see [88] for details). Similarly, it is difficult to find a close-from solution to this maximization problem, but, alternatively, we can solve it by an iterative algorithm. Specifically, in each iteration, $\mathbf{w}_a^1, \ldots, \mathbf{w}_a^{k-1}, \mathbf{w}_a^{k+1}, \ldots, \mathbf{w}_a^n$ and $\mathbf{w}_b^1, \ldots, \mathbf{w}_b^{k-1}, \mathbf{w}_b^{k+1}, \ldots, \mathbf{w}_b^n$ are assumed known, then the problem is reduced to CCA between $\mathcal{A} \times_1 \mathbf{w}_a^1 \cdots \times_{k-1} \mathbf{w}_a^{k-1} \times_{k+1} \mathbf{w}_a^{k+1} \cdots \times_n \mathbf{w}_a^n$ and $\mathcal{B} \times_1 \mathbf{w}_b^1 \cdots \times_{k-1} \mathbf{w}_b^{k-1} \times_{k+1} \mathbf{w}_b^{k+1} \cdots \times_n \mathbf{w}_b^n$. The canonical factors in all n dimensions can be iteratively optimized.

5.3 Experiments

As a case study, we investigate correlations between the mouth part (Mouth) and the right eye part (Eye) (as shown in Figure 5.4). These two parts have strong and a range of correlations corresponding to facial expressions. We conducted experiments on the Cohn-Kanade database [78] and face expression image sequences we captured. We manually normalized the faces based on three feature points, centers of the two eyes and the mouth, using affine transformation. In the normalized facial images (110×150 pixels), the mouth part is 53×68 pixels, and the eye part is 45×51 pixels.

5.3.1 Facial Parts Synthesis

We wish to reconstruct (synthesize) Mouth from Eye or vice versa using MCCA based regression. Specifically, to reconstruct image $\mathbf{B}$ from image $\mathbf{A}$, we first employ MCCA to establish

Figure 5.4: A case study on correlations between the mouth and the right eye facial parts.

their relationship, finding optimal projection directions in the sense of correlation, and then map A to the leading canonical variates by discarding directions with low canonical correlation. Finally we perform regression of $\mathbf{B}$ by taking these leading canonical variates of $\mathbf{A}$. Our procedure for synthesis is as follows.

1. Compute the leading factor pairs $\mathbf{V}_a, \mathbf{W}_a, \mathbf{V}_b, \mathbf{W}_b$ from N pairs of samples $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N\}$ and $\mathcal{B} = \{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_N\}$.

2. Map $\mathbf{A}_i$ $(i = 1, \dots, N)$ to the reduced correlation space $\widetilde{\mathbf{A}}_i = \mathbf{V}_a^T \mathbf{A}_i \mathbf{W}_a$.

3. Reshape 2D matrices $\widetilde{\mathbf{A}}_i$ and $\mathbf{B}_i$ to 1D vectors $\tilde{\mathbf{a}}_i$ and $\mathbf{b}_i$, and form data matrices $\widetilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_N]$ and $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_N]$; then compute the regression matrix $\mathbf{R} = (\widetilde{\mathbf{A}}^T)^{-1} \mathbf{B}^T$.

4. Given a new input $\mathbf{A}_{new}$, the corresponding $\mathbf{B}_{new}$ is reconstructed by:

$$\widetilde{\mathbf{A}}_{new} = \mathbf{V}_a^T \mathbf{A}_{new} \mathbf{W}_a, \quad \widetilde{\mathbf{A}}_{new} \to \tilde{\mathbf{a}}_{new} \tag{5.10}$$

$$\mathbf{b}_{new} = \mathbf{R}^T \tilde{\mathbf{a}}_{new}, \quad \mathbf{b}_{new} \to \mathbf{B}_{new} \tag{5.11}$$

Here $\widetilde{\mathbf{A}}_{new} \to \tilde{\mathbf{a}}_{new}$ represents reshaping 2D matrix $\widetilde{\mathbf{A}}_{new}$ to 1D vector $\tilde{\mathbf{a}}_{new}$, and $\mathbf{b}_{new} \to \mathbf{B}_{new}$ is reshaping 1D vector $\mathbf{b}_{new}$ to 2D matrix $\mathbf{B}_{new}$. This reconstruction procedure is not limited to facial parts but can also be generally applied to other types of image correlation analysis and synthesis.

We selected 10 subjects from the Cohn-Kanade database, each of which has around 70∼300 images of different facial expressions, in addition to the image sequences we captured. For the image set of each subject, we randomly sampled one tenth of the images as the testing set, and the remaining images as the training set. We applied MCCA, CCA, and the

standard linear least-squares regression (SR) approach to synthesize Mouth from Eye and vice versa on the testing set. We used 10 randomly selected training/testing combinations for reporting reconstruction errors. We observe that MCCA performs better than CCA and SR in reconstructing one facial part from another. Moreover, MCCA requires much fewer canonical factors to obtain better reconstruction results. We show the reconstruction results for six random selected subjects in Table 5.1, where the optimal average pixel errors (with standard deviation) and the corresponding dimensions of canonical factors used are reported. To clearly compare the three methods, we plot bar graphs of average pixel errors and the dimensions of the canonical factors used in MCCA/CCA in Figure 5.5. Some reconstruction examples are shown in Figure 5.6[1].

Subject	Algorithm	Eye → Mouth		Mouth → Eye	
		Pixel Errors	Dims	Pixel Errors	Dims
(1)	MCCA	11.2±2.0	11*6	8.8±1.2	2*23
	CCA	16.7±4.4	139	13.1±4.0	139
	SR	17.3±3.5	-	14.7±4.8	-
(2)	MCCA	8.5±2.5	9*6	8.4±1.8	5*10
	CCA	13.0±6.0	119	10.7±3.6	119
	SR	12.4±5.3	-	10.7±3.2	-
(3)	MCCA	13.4±5.5	15*3	10.0±3.3	28*1
	CCA	16.2±8.8	96	11.6±5.9	96
	SR	16.1±8.8	-	12.0±6.5	-
(4)	MCCA	16.3±5.0	39*1	19.5±6.0	17*3
	CCA	24.5±9.8	96	25.4±18.3	96
	SR	23.7±7.6	-	26.1±18.8	-
(5)	MCCA	9.9±1.8	14*2	10.5±2.5	22*2
	CCA	12.9±4.4	85	11.0±3.1	85
	SR	14.2±4.3	-	11.0±2.9	-
(6)	MCCA	13.8±2.4	28*1	12.6±2.9	18*2
	CCA	17.2±8.5	77	15.7±6.2	77
	SR	15.1±4.9	-	13.9±6.1	-

Table 5.1: The results of 6 subjects: the optimal average pixel errors (with standard deviation) of the three algorithms, and the corresponding dimensions of canonical factors used in MCCA and CCA.

It is compelling that MCCA outperforms CCA and SR consistently in facial parts synthesis. Crucially, as observed in Figure 5.5, the dimension of canonical factors needed in MCCA is always less than 50% of that of CCA. So MCCA can describe correlations among facial parts with better accuracy using much less canonical factors. The superior performance of MCCA

---

[1]A video demonstration is available at *http://www.dcs.qmul.ac.uk/∼cfshan/research/cca.html.*
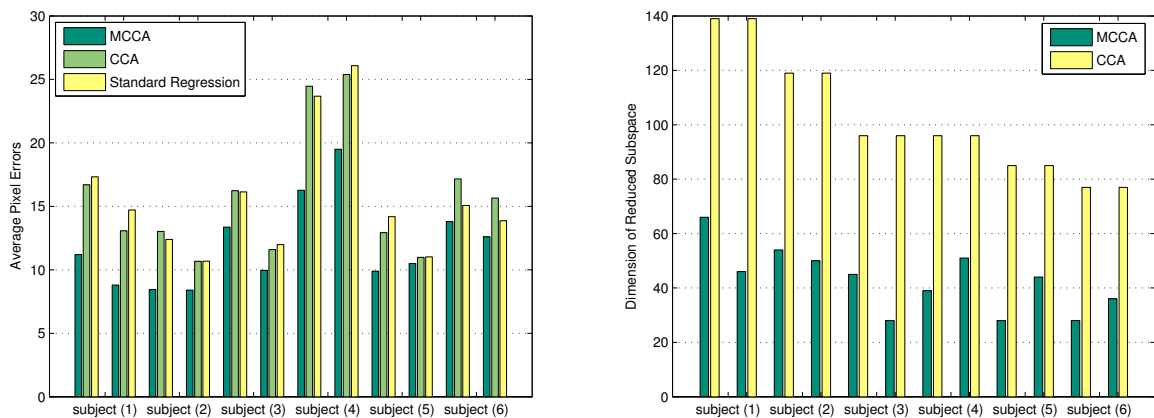
Figure 5.5: (*left*) Reconstruction errors of the three algorithms; (*right*) Dimensions of canonical factors used in MCCA and CCA. (Two groups bars for each subject: the left is 'Eye → Mouth' and the right is 'Mouth → Eye'.)

Figure 5.6: Some examples of facial parts synthesis using MCCA, CCA, and SR.

credits to its ability to preserve the intrinsic 2D spatial structure and capture the correlation store therein, and its robustness with limited number of training data. The strength of MCCA is also reflected by the average standard deviation. As shown in Table 5.1, MCCA always produces the smallest deviation, which suggests that MCCA is much more robust. Compared to SR, where the full-rank regression matrix has to be estimated from a limited

number of noisy training images, the MCCA based reduced-rank regression provides more reliable parameter estimates by taking advantage of correlations between the image sets, leading to better accuracy and robustness.

5.3.2 Facial Expression Recognition

We also conducted facial expression recognition experiments based on correlations between Mouth and Eye. The basic idea is that these two parts have distinctive correlations for different expressions, so the correlations modeled by MCCA should provide discriminant information for expression classification. Given image sets of different facial expressions $\mathcal{I}_1, \ldots, \mathcal{I}_c$ (c is the number of classes), we derive the leading factor pairs $(\mathbf{V}_a^i, \mathbf{W}_a^i, \mathbf{V}_b^i, \mathbf{W}_b^i), i = 1 \ldots c$ of parts Mouth (denoted by $\mathbf{B}$) and Eye (denoted by $\mathbf{A}$) for each class using MCCA. We then compute the regression parameters for reconstructing $\mathbf{B}$ from $\mathbf{A}$ in the reduced correlation space in the training set. Given a test image $\mathbf{I}_{new}$ of an unknown class, we map its Eye $\mathbf{A}_{new}$ and Mouth $\mathbf{B}_{new}$ to the reduced correlation space of class i as $\widetilde{\mathbf{A}}_i = (\mathbf{V}_a^i)^T \mathbf{A}_{new} \mathbf{W}_a^i$ and $\widetilde{\mathbf{B}}_i = (\mathbf{V}_b^i)^T \mathbf{B}_{new} \mathbf{W}_b^i$, and then calculate the error $err(i)$ of reconstructing $\widetilde{\mathbf{B}}_i$ from $\widetilde{\mathbf{A}}_i$ with the regression parameters of this correlation space. After computing the reconstruction error of each class $err(i), i = 1 \ldots c$, we classify the test image as the class having the smallest reconstruction error

$$\hat{i} = \arg \min_i err(i) \tag{5.12}$$

For our experiments, we selected 244 image sequences of basic emotions (Anger, Disgust, Joy, and Surprise) from the Cohn-Kanade database. The sequences come from 96 subjects, with 1 to 4 emotions per subject. For each sequence, the three frames showing the peak of facial expression were used, which lead to a total of 732 images of four expression classes. We first considered a 2-class (Joy and Surprise) recognition problem, then included Anger for a 3-class problem, and finally considered four expressions for classification (incrementally making the recognition task harder). To evaluate generalization performance, a 10-fold Cross-Validation testing scheme was adopted. The recognition results using MCCA and CCA are reported in Table 5.2. We can observe that expressions can be better classified using MCCA,

demonstrating again that MCCA outperform CCA in capturing correlations in facial parts. It is also evident that by modeling correlations between only two facial parts, the recognition accuracy degrades quickly for multi-class recognition. By considering correlations of multiple facial parts, we should be able to improve these recognition results.

	2-Class	3-Class	4-Class
MCCA	96.1±3.6	80.8±6.4	67.9±4.8
CCA	63.2±10.5	55.6±7.8	48.7±6.7

Table 5.2: Facial expression recognition based on correlations of Mouth and Eye modeled by MCCA and CCA.

5.4 Summary

In this chapter, we employ Canonical Correlation Analysis to model the correlations among facial parts. To overcome the inherent limitations of classical CCA for image data, we introduce a Matrix-based Canonical Correlation Analysis (MCCA) for better correlation analysis among image data. Experimental results have shown this technique can provide superior performance in regression and recognition tasks, whilst requiring significantly fewer canonical factors. The underlying reason is that MCCA is able to preserve and utilize the intrinsic 2D spatial structure in image data.

All the above work focuses on facial expressions. However, the face is usually perceived not as an isolated object but as an integrated part of the whole body, and the visual channel combining facial and bodily expressions is most informative [3]. Therefore, in the following chapter, we investigate two understudied problems in bodily expression analysis. By deriving a semantic joint feature space, we further combine face and body cues at the feature level for improved performance.

6 Multimodal Facial and Body Language Analysis

Beyond facial expression analysis, we study in this chapter body language analysis in videos. Specifically, we investigate two relatively understudied problems: gait-based gender discrimination and affective body gesture recognition. A large number of studies have investigated gender classification by human faces [104]. However, face information is not always available or reliable in real-world scenarios. On the contrary, as a unique biometric that can be recognized at a distance or at low resolution, human gaits provide important alternative cues for gender classification in unconstraint situations. With regard to affect analysis, little attention has been placed on emotional body gesture and posture analysis, although bodily expression is an important channel for humans to convey their emotional states.

Each modality, the face or the body, in isolation has its inherent weaknesses and limitations. Integrating face and body cues provides a potential way to accomplish improved gender discrimination or affect analysis. We further exploit CCA to fuse the two modalities at the feature level. CCA establishes the relationship between the modalities, and derives a semantic "gender" or "affect" space, in which the face and body features are compatible and can be effectively fused. We plot in Figure 6.1 the flow chart of our multimodal gender recognition system.

6.1 Learning Gender from Human Gaits

Gender classification is an important visual task for human beings, as many social interactions critically depend on the correct gender perception. As visual surveillance and human-
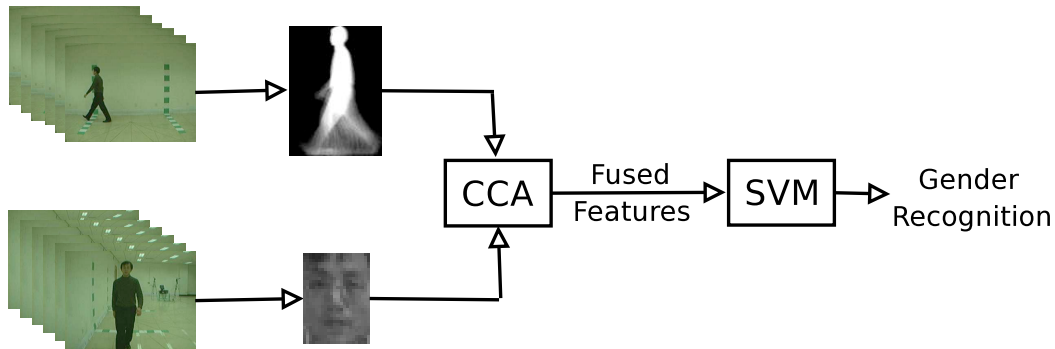
Figure 6.1: The flow chart of our multimodal gender recognition system.

computer interaction technologies evolve, computer vision systems for gender classification will play an increasingly important role in our lives, e.g. collecting valuable demographic information in a social environment. Human gaits contain subtle, yet informative, stylistic variations, providing complementary and alternative cues to faces for person identification and more fundamental gender discrimination.

In our work, we adopt Gait Energy Image (GEI) [64], a spatio-temporal compact representation of gaits, which has been demonstrated to be effective for representing gaits in the human identification problem [64, 184]. Using background substraction techniques, walking subjects can be extracted from original image sequences to derive binary silhouette image sequences. To make the gait representation insensitive to the distance between the camera and the subject, we perform silhouette preprocessing including size normalization and horizontal alignment [64]. Some examples of normalized and aligned silhouette images are shown in Figure 6.2. The entire human gait sequence can be divided into cycles as human walking repeats at a stable frequency. We decide the gait cycles by counting the number of foreground pixels in the bottom half of the silhouette [134], and the two consecutive strides in the variation of the number constitute a gait cycle.

Given the preprocessed binary silhouette image $B_t(x, y)$ at time t in a sequence, the GEI is defined as follows:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y) \tag{6.1}$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the
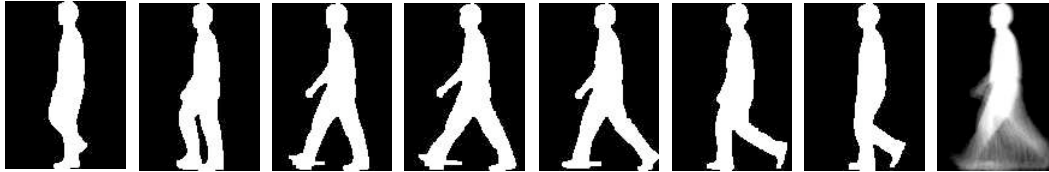
Figure 6.2: Examples of normalized and aligned silhouette images. The rightmost image is the corresponding GEI.

frame number of the sequence, and x and y are values in the 2D image coordinate (see Figure 6.2 for an example of GEI). GEI reflects shapes of silhouette and their changes over the gait cycle, and it is not sensitive to incidental silhouette errors in individual frames.

A successful technique for gender classification is SVM [104, 90, 177]. SVM is an optimal discriminant method based on the Bayesian learning theory. For the cases where it is difficult to estimate the density model in high-dimensional space, the discriminant approach is preferable to the generative approach (see Section 3.2.3 for details).

6.2 Affective Body Gesture Recognition

Recently spatial-temporal features have been investigated for event detection and behaviour recognition in videos [47, 87, 81, 43, 108]. Efros *et al.* [47] introduced a motion descriptor based on optical flow measurements in a spatio-temporal volume, which was applied to recognize human action at a distance.By extending 2D rectangle features into the spatio-temporal domain, Ke *et al.* [81] presented volumetric features for event detection in videos. They constructed a real-time event detector by learning a cascade of filters based on volumetric features that scan video sequences in space and time. Laptev and Lindeberg [87] extended spatial interest points into the spatio-temporal domain, and presented a method to detect local structures in space-time where the image values have significant local variation in both space and time. Dollár *et al.* [43] proposed an alternative approach to detect sparse space-time interest points based on separable linear filters, and utilized cuboids of spatio-temporal windowed data surrounding each feature point for human behaviour recognition. Based on their work, Niebles *et al.* [108] more recently presented an unsupervised learning method

for action categorization. Spatial-temporal features have been proven useful to provide a compact abstract representation of video patterns.

Here we adopt spatial-temporal features to represent body gesture in videos. Different from previous studies [7, 60], where the tracked motion of hands (body parts) was applied to gesture recognition, relying too much on human supervision and robust hand tracking and segmentation, our approach makes few assumptions about the observed data, such as background, occlusion and appearance. The underlining motivation is that although two instances of the same body gesture may vary in terms of overall appearance and motion, due to variations across subjects or within each individual, many of the spatial-temporal features detected are similar.

We extract spatial-temporal features by detecting space-time interest points in videos. Following [43, 108], we calculate the response function by application of separable linear filters. Assuming a stationary camera or a process that can account for camera motion, the response function has the form:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{6.2}$$

where $I(x, y, t)$ denotes images in the video, $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions (x, y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. In all cases we use $\omega = 4/\tau$ [43]. The two parameters σ and τ correspond roughly to the spatial and temporal scales of the detector. Each interest point is extracted as a local maxima of the response function. As pointed out in [43], any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response, while region undergoing pure translational motion, or areas without spatially distinguishing features, will not induce a strong response.

At each detected interest point, a cuboid is extracted which contains the spatio-temporally windowed pixel values. See Figure 6.3 for examples of cuboids extracted. The side length of cuboids is set as approximately six times the scales along each dimension, so containing
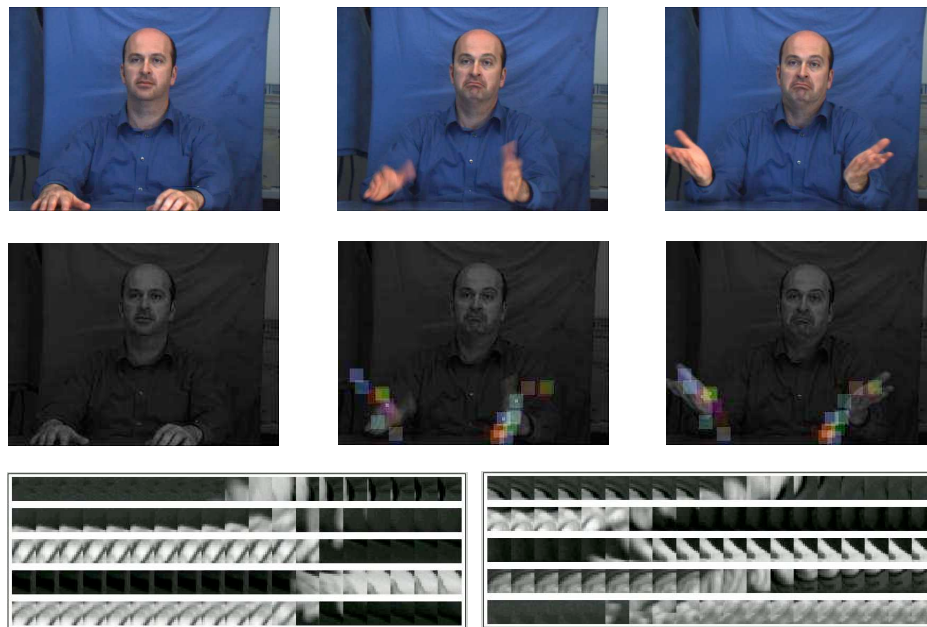
Figure 6.3: (Best viewed in color) Examples of spatial-temporal features extracted from videos: the first row is the original input video; the second row visualizes the cuboids extracted, where each cuboid is labeled with a different color; the third row shows some cuboids, which are flattened with respect to time.

most of the volume of data that contribute to the response function at each interest point. After extracting the cuboids, the original video is discarded, which is represented as a collection of the cuboids. To compare two cuboids, different descriptors for cuboids have been evaluated in [43], including normalized pixel values, brightness gradient and windowed optical flow, followed by a conversion into a vector by flattening, global histogramming, and local histogramming. As suggested, we adopt the flattened brightness gradient as the cuboid descriptor. To reduce the dimensionality, the descriptor is projected to a lower dimensional PCA space [43]. By clustering a large number of cuboids extracted from the training data using the K-Means algorithm, we derive a library of cuboid prototypes. So each cuboid is assigned a type by mapping it to the closest prototype vector. Following [43], we use the histogram of the cuboid types to describe the video. With regard to the classification technique, SVM is adopted to recognize body gesture expressing different emotions.

6.3 Fusing Face and Body Cues for Recognition

Each modality, the face or the body, in isolation has its inherent weakness and limitations. With regard to affect analysis, a single body gesture can be ambiguous. For instance, the videos shown in the second and fourth row in Figure 2.12 have much similar body gesture, but the affective state they express are quite different, as shown by their facial expressions. Integrating face and body cues is potentially beneficial to accomplish more effective gender discrimination or affect analysis.

Recently several attempts [138, 76, 184] have been made to integrate face and gait cues for the human identification problem. All these existing studies have focused on the decision-level fusion of face and gait, while the feature-level fusion is understudied. This is mainly because the two modalities may have incompatible feature sets and the relationship between the different feature spaces is unknown. On affective analysis, the psychological study [100] suggest the integration of facial expression and body gesture is a mandatory process occurring early in the human processing stream. Therefore, different modalities cannot be considered mutually independently and combined at the end of the intended analysis but, on the contrary, the input data should be processed in a joint feature space [73].

Here we propose to fuse face and body cues at the feature level using CCA. Our motivation is that, as the face and the body (face and gait, or facial expression and body gesture) are two sets of measurements for human gender or affective states, conceptually the two modalities are correlated, although their correlations may not be obvious on the original measurements. CCA can establish their relationship by finding the maximum correlation on transformed space. CCA derives a semantic "gender" or "affect" space, in which the face and body features are compatible and can be effectively fused.

Given $B = \{\mathbf{x}|\mathbf{x} \in \mathbb{R}^m\}$ and $F = \{\mathbf{y}|\mathbf{y} \in \mathbb{R}^n\}$, where $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors extracted from the body and the face respectively, we apply CCA to establish the relationship between $\mathbf{x}$ and $\mathbf{y}$. Suppose $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ are the canonical factors pairs obtained, we can use d $(1 \leq d \leq k)$ factor pairs to represent the correlation information. With $\mathbf{W}_x = [\mathbf{w}_x^1, \ldots, \mathbf{w}_x^d]$ and $\mathbf{W}_y = [\mathbf{w}_y^1, \ldots, \mathbf{w}_y^d]$, we project the original feature vectors as $\mathbf{x}' =$

$\mathbf{W}_x^T \mathbf{x} = [x_1, \ldots, x_d]^T$ and $\mathbf{y}' = \mathbf{W}_y^T \mathbf{y} = [y_1, \ldots, y_d]^T$ in the lower dimensional correlation space. The canonical variates x_i and y_i (corresponding to $\mathbf{w}_x^i$ and $\mathbf{w}_y^i$) are uncorrelated with the previous pairs x_j and $y_j, j = 1, \ldots, i-1$ (see Section 5.1). We then combine the projected feature vector $\mathbf{x}'$ and $\mathbf{y}'$ to form the new feature vector as

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x^T \mathbf{x} \\ \mathbf{W}_y^T \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x & 0 \\ 0 & \mathbf{W}_y \end{pmatrix}^T \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{6.3}$$

This fused feature vector effectively represents the multimodal information in a joint feature space.

6.4 Experiments

6.4.1 Learning Gender from Human Gaits and Faces

Data

We carried out experiments on the CASIA Gait Database (Dataset B) [178], currently one of the largest gait databases in the gait-research community. The database consists of 124 subjects aged between 20 and 30 years, of which 93 were male and 31 were female, and 123 were Asian and 1 was European. Each subject first walked naturally along a straight line six times, then put on his/her coat and walked twice, and finally walked twice carrying a bag (knapsack, satchel, or handbag). Each subject walked a total of ten times in the scene (6 normal + 2 with a coat + 2 with a bag). 11 cameras were uniformly set on the left hand side, with view angle interval of 18°, so 11 video sequences from different views were captured simultaneously for every walking scenario (see Figure 6.4). There are a total of 13,640 (124×10×11) video sequences in the database, with 2 to 3 gait cycles in each sequence. The frame size is 320-by-240 pixel, and the frame rate is 25 fps.

In our experiments we used video sequences from two views for gender recognition: frontal view for face cues and side view for gait cues. We selected video sequences of 119 subjects (88 Male and 31 Female) that are suitable for gait and face analysis. In total 2,380 (119×10×2) video sequences were used in our experiments. Compared to the small dataset (24 subjects)

Figure 6.4: The walking sequences captured from 11 different views.

used in the previous work [90], our study was performed on a much larger dataset.

As the database was collected for human gait analysis, there was no specific consideration of face data collection. Human faces were captured in an unconstrained environment like a real-world surveillance scenario. The sequences contain facial expression changes, head pose variations, hair and glasses presented in the low-resolution faces. We first adopted an AdaBoost based face detector to detect face regions in each video sequence. Then, for simplicity, we manually labeled the three points (two eyes and the mouth) of the detected face with the best resolution in a sequence, and normalized the face as a 30-by-22 pixel thumbnail to represent the video sequence. That is, we extracted a face image for each video sequence. To derive gait data, we computed the GEI for each video sequence. We show the processed face images and GEIs of 20 subjects (10 female + 10 male) in Figure 6.5, where the first row of GEIs are normal walking, and the second row is carrying a bag, while the bottom row is with wearing his/her coat.

Gender Recognition from Gaits

To evaluate the algorithms' generalization ability, we adopted a 5-fold cross-validation test scheme in all recognition experiments. That is, we divided the data set randomly into five
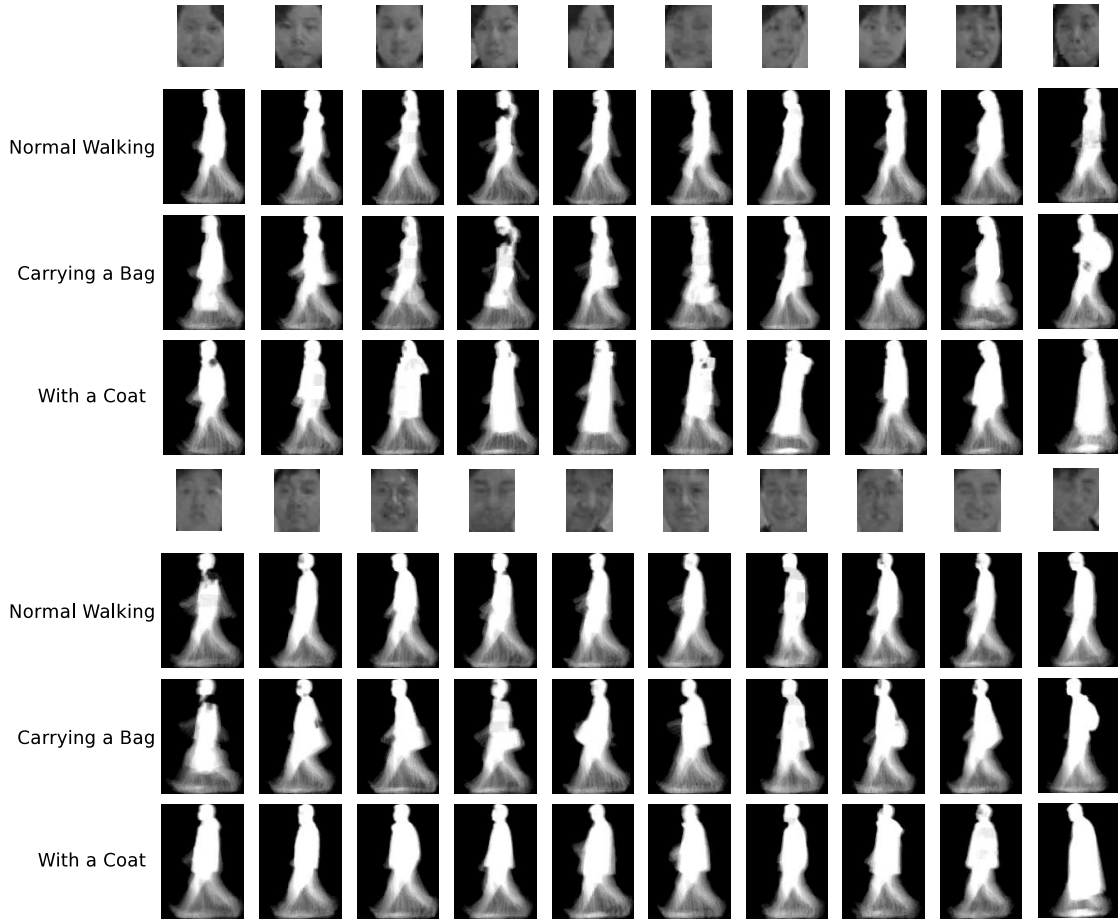
Figure 6.5: The extracted face images and GEIs of 20 subjects. (*Top*) Female; (*Bottom*) Male.

groups with roughly equal (female and male) subjects, and then used the data from four groups for training and the left group for testing; the process was repeated five times for each group in turn to be tested. We show the average recognition rates (with the standard deviation) here. In all experiments, we set the soft margin C value of SVMs to infinity so that no training error was allowed. Meanwhile, each training and testing vector was scaled to be between -1 and 1. With regard to the hyper-parameter selection of Polynomial and RBF kernels, as suggested in [71], we carried out grid-search on the kernel parameters in the 5-fold cross-validation. The parameter setting producing the best cross-validation accuracy was picked. We also used the SVM implementation in the machine learning library SPIDER.

We report the results of gait-based gender recognition in Table 6.1. It is observed that GEIs

Classifier	Recognition Rates		
	Overall	Male	Female
SVM (Linear/Polynomial)	94.2±2.1%	97.5±3.2%	84.7±10.4%
SVM (RBF)	93.6±2.3%	96.8±3.9%	84.4±10.7%
PCA+LDA	94.5±1.9%	98.0±2.4%	84.6±9.6%

Table 6.1: Experimental results of gait-based gender recognition.

based SVMs produce high overall recognition rates (93-94%), and the linear kernel and the (1st degree) polynomial kernel provide the same performance, slightly better than the RBF kernel. The number of support vectors of SVMs with different kernels were 13-16 percent of the total number of training samples. It is indicated that, for the GEI based gait representation, the linear decision surface is able to effectively classify gender, although there are many variations in GEIs due to wearing a coat or carrying a bag (as shown in Figure 6.5). To verify this, we further performed experiments with the linear subspace method PCA+LDA, which has frequently been used for the appearance-based object recognition. PCA reduces the dimension of feature space, and LDA identifies the most discriminant features. A nearest-neighbor classifier was used in our experiments. The experimental results summarized in Table 6.1 show that PCA+LDA achieves similar performance to the linear/polynomial kernels. Therefore, GEI is an effective gait representation for gender recognition, based on which the linear decision surface can discriminate gender with high confidence. The performance of GEI is also much better than that of dynamic features (84.5%) used in [90].

Gender Recognition from Faces

Before fusing gait and face modalities, we first performed gender recognition with faces, and show the results in Table 6.2. By comparing Table 6.1 and Table 6.2, we can see that recognition results based on faces alone were consistently inferior to that based on gaits, which indicates that it is hard to learn human gender from low-resolution faces captured in unconstrained environments. For face-based gender recognition, SVMs have a clear margin of superiority over the linear subspace method PCA+LDA. The polynomial kernel also achieved the same performance with the linear kernel, but RBF kernel was found to perform best. The results we obtained reinforce the findings reported in [104]. This indicates that the face

data can be better gender classified by nonlinear decision surfaces. The number of support vectors of the linear/polynomial kernels were 23-24 percent of the total number of training samples, while the RBF kernel employed 25-39 percent. The SVMs' performance of 87-90% we obtained is inferior to that reported in [104]. This is because our face data was captured in an unconstrained real-world scenario, with the presence of facial expression changes, head pose variations, various hair styles and glasses. Therefore it is more complex than the face images of FERET database used in [104].

Classifier	Recognition Rates		
	Overall	Male	Female
SVM (Linear/Polynomial)	87.5±1.8%	92.3±2.1%	74.3±10.3%
SVM (RBF)	90.4±1.8%	96.0±2.1%	74.6±9.7%
PCA+LDA	76.2±1.8%	79.6±3.5%	66.2±7.7%

Table 6.2: Experimental results of face-based gender recognition.

Gender Recognition from Gaits and Faces

We then fused gait and face cues at the feature level using CCA. Different numbers of CCA factor pairs can be used to project the original gait and face feature vectors to a lower dimensional CCA feature space, and the recognition performance varies with the dimensionality of the projected CCA features. We first tested SVM (Linear) with the CCA features of different dimensions. We plot in Figure 6.6 the average recognition rates of SVM (Linear) versus CCA dimensionality reduction. It is observed that the projected CCA features of gaits and faces with 90-dimensions provide the best performance. Hence we carried out subsequent experiments with CCA features of 90-dimensions.

To verify its effectiveness, we compared the presented CCA feature fusion with another three feature fusion methods: (1) Direct feature fusion, that is, concatenating the original gait and face feature vectors to derive a single feature vector; (2) PCA feature fusion: the original gait and face feature vectors are first projected to the PCA space respectively, and then the PCA features are concatenated to form the single feature vector. In our experiments, all principle components were kept. (3) PCA+LDA feature fusion: for each modality, the derived PCA features are further projected to the discriminant LDA space; the LDA features
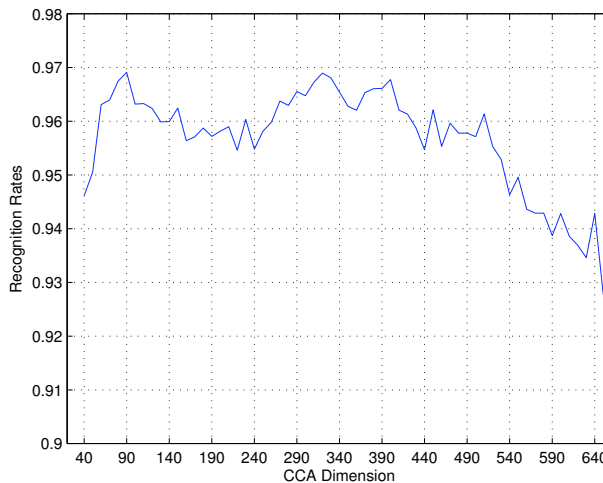
Figure 6.6: Recognition rates of SVM (Linear) versus dimensionality reduction of CCA.

are then combined to derive the single feature vector. We show the experimental results of different feature fusion schemes in Table 6.3, where it shows the linear kernel also achieves the same performance as the polynomial kernel. We also plot bar graphs of the recognition performance in Figure 6.7. We can see that the direct feature fusion and PCA+LDA feature fusion outperform slightly the single modality, while the PCA feature fusion provides the performance that is better than that of face cues but inferior to that of gait cues. In contrast, our proposed CCA feature fusion consistently achieves the best recognition results, producing considerable performance improvement over the single modality. This is because CCA captures the relationship between the feature sets in different modalities, and the fused CCA features effectively represent information in each modality, removing noisy and redundant data. More crucially, the CCA feature fusion brings significant time and space benefit. Compared to the high dimensionality (4,160) in the direct feature fusion. the compact 180-dimension CCA features reduce the memory space by a order of 23. Another strength of the CCA feature fusion is that it always produces the smallest standard deviation of cross-validation, which demonstrate it is more robust than each single modality and other feature fusion schemes. The performance 97.2% that the CCA feature fusion based SVM (RBF) obtained is better than 96.6% reported in [104], and, to our best knowledge, is the best gender

135

		Recognition Rates			Feature
		Overall	Male	Female	Dimension
Direct Fusion	SVM (Linear/Polynomial)	95.6±1.7%	98.3±2.4%	88.0±8.6%	4,160
	SVM (RBF)	94.5±1.8%	97.4±3.1%	86.3±8.5%	
CCA Fusion	SVM (Linear/Polynomial)	96.9±1.1%	99.0±1.1%	91.0±5.2%	180
	SVM (RBF)	**97.2**±0.8%	**99.0**±1.3%	**92.0**±4.6%	
PCA Fusion	SVM (Linear/Polynomial)	92.3±0.9%	94.6±1.9%	85.6±6.9%	1,600
	SVM (RBF)	92.5±1.3%	95.8±1.1%	83.1±7.1%	
PCA+LDA Fusion	SVM (Linear/Polynomial)	95.6±1.9%	98.3±1.7%	87.9±8.0%	2
	SVM (RBF)	95.6±1.9%	98.3±1.7%	87.9±8.0%	

Table 6.3: Experimental results of gender recognition by fusing gaits and faces.

recognition performance reported so far in the published literature[1].

We note that, in Tables 6.1 - 6.3, all the female recognition rates are poorer than the male (with larger variance). In previous studies [104, 139], different classifiers also had higher error rates in classifying females. This phenomenon is possibly because the female gaits and faces have less prominent and distinct features. For example, the female has much variation in their hair styles and clothing. Another possible reason is the unbalanced data set (88 Male and 31 Female) in our experiments. An encouraging observation is the female recognition performance based on each single modality is improved much by the CCA feature fusion (from 74-84% to 91-92%) which is significant.

In the above experiments, the face images were manually aligned. To investigate the effect of the misalignment of faces on final fusion results, we further carried out experiments by taking the face images directly from the face detector. Due to the unaligned face, the recognition performance of different fusion methods degrades to 85-91%, although the CCA feature fusion still provides the best performance.

6.4.2 Fusing Facial and Bodily Expressions for Emotion Recognition

Data

There are several facial expression databases in affective-computing community, but few databases containing affective body gestures. Gunes and Piccardi [61] recently collected a bimodal face and body gesture database (FABO), which consists of facial expression and

---

[1]A video demonstration is available at *http://www.dcs.qmul.ac.uk/∼cfshan/research/gender.html.*
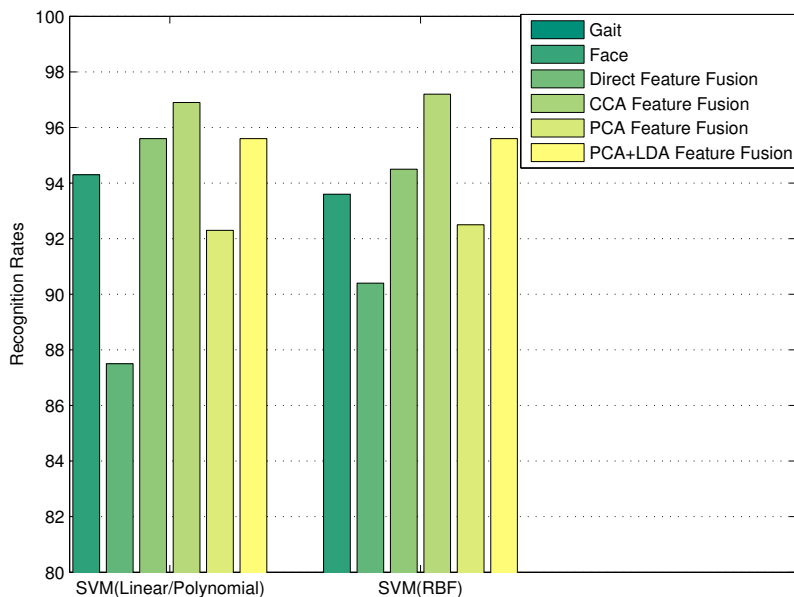
Figure 6.7: Gender recognition using different features.

body gesture recorded simultaneously. The database includes 23 subjects in age from 18 to 50 years, of which 12 were female, 23 were from Europe, 2 from Middle East, 3 from Latin America, 7 from Asia, and 1 from Australia. In total there are around 1900 videos. Examples of the video sequences are shown in Figure 2.12. In our experiments, we selected 262 videos of seven emotions (Anger, Anxiety, Boredom, Disgust, Joy, Puzzle, and Surprise) from 23 subjects. Gunes and Piccardi [60] reported some preliminary results on this database, but they only used 54 videos from 4 subjects.

Affective Body Gesture Recognition

To evaluate the algorithms' generalization ability, we adopted a 5-fold cross-validation test scheme in all recognition experiments. That is, we divided the data set randomly into five groups with roughly equal number of videos, and then used the data from four groups for training and the left group for testing; the process was repeated five times for each group in turn to be tested. We report the average recognition rates here. In all experiments, we set the

	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.63	.00	.07	.04	.13	.09	.04
anxiety	.04	.83	.00	.00	.00	.13	.00
boredom	.00	.03	.83	.03	.03	.07	.00
disgust	.08	.00	.04	.68	.04	.12	.04
joy	.08	.00	.15	.05	.65	.08	.00
puzzle	.05	.07	.05	.11	.07	.59	.07
surprise	.00	.00	.00	.11	.11	.05	.74

	anger	anxiety	boredom	disgust	joy	puzzle	surprise
anger	.67	.00	.07	.04	.11	.07	.04
anxiety	.00	.87	.00	.00	.00	.13	.00
boredom	.00	.03	.86	.00	.03	.07	.00
disgust	.08	.00	.04	.80	.00	.04	.04
joy	.08	.00	.15	.05	.65	.08	.00
puzzle	.05	.07	.02	.11	.07	.61	.07
surprise	.00	.00	.00	.06	.12	.06	.76

Figure 6.8: Confusion matrices of affective body gesture recognition with the 1-nearest neighbor classifier (*left*) and the SVM classifier (*right*).

soft margin C value of SVMs to infinity so that no training error was allowed. Meanwhile, each training and testing vector was scaled to be between -1 and 1. In our experiments, the RBF kernel always provided the best performance, so we report the performance of the RBF kernel. With regard to the hyper-parameter selection of RBF kernels, as suggested in [71], we carried out grid-search on the kernel parameters in the 5-fold cross-validation. The parameter setting producing the best cross-validation accuracy was picked.

We compare the SVM classifier with the 1-nearest neighbor classifier used in [43] for affective body gesture recognition. The average recognition rates of SVM and 1-nearest neighbor classifier are 72.6% and 68.6% respectively. We plot the confusion matrices of the two classifier in Figure 6.8. It can observed that the SVM classifier slightly outperforms the 1-nearest neighbor classifier.

Emotion Recognition by Fusing Face and Body Cues

In the FABO database, video sequences were recorded simultaneously using two video cameras, one is for capturing the facial expression only and the other for capturing upper-body movements. We extracted the spatial-temporal features from the face video and the body video, and then fuse the two modalities at the feature level using CCA. We first show the classification performance based on facial cues only. The confusion matrices of the two classifiers are shown in Figure 6.9, and the recognition rates of SVM and 1-nearest neighbor
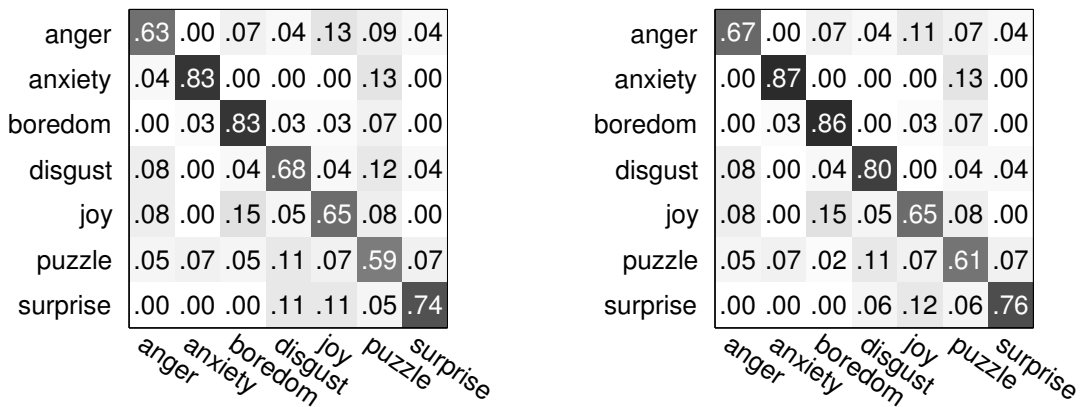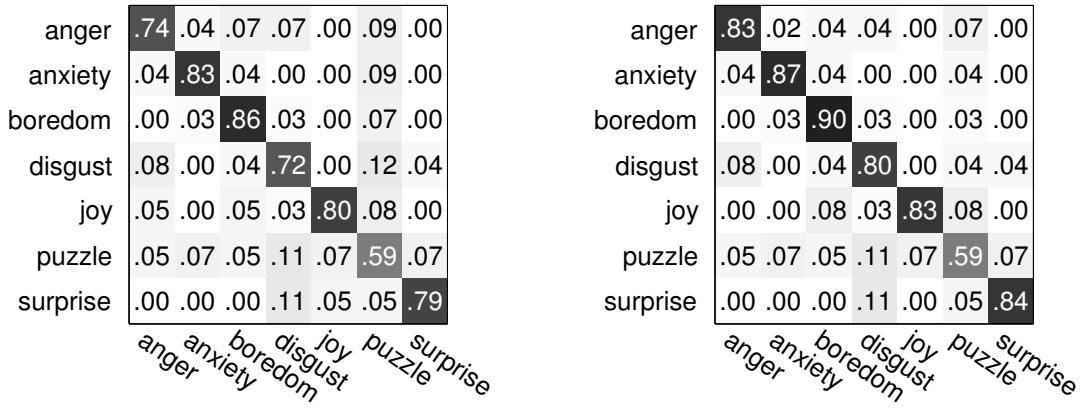
Figure 6.9: Confusion matrices of facial expression recognition with the 1-nearest neighbor classifier (*left*) and the SVM classifier (*right*).

classifier are 79.2% and 74.8% respectively. We can see that emotion classification based on facial expressions is better than that of body gesture. This is possibly because there are much variation in affective body gestures.

We then fused facial expression and body gesture at the feature level using CCA. Different numbers of CCA factor pairs can be used to project the original face and body feature vectors to a lower dimensional CCA feature space, and the recognition performance varies with the dimensionality of the projected CCA features. We report the best result obtained here. We compared the CCA feature fusion with another three feature fusion methods: (1) Direct feature fusion, that is, concatenating the original body and face features to derive a single feature vector; (2) PCA feature fusion: the original body and face features are first projected to the PCA space respectively, and then the PCA features are concatenated to form the single feature vector. In our experiments, all principal components were kept. (3) PCA+LDA feature fusion: for each modality, the derived PCA features are further projected to the discriminant LDA space; the LDA features are then combined to derive the single feature vector. We show the experimental results of different feature fusion schemes in Table 6.4. The confusion matrices of the CCA feature fusion and the direct feature fusion are shown in Figure 6.10. We can see that the presented CCA feature fusion provides best recognition performance. This is because CCA captures the relationship between the feature sets in different modalities, and the fused CCA features effectively represent information from each

Feature Fusion	CCA	Direct	PCA	PCA+LDA
Recognition Rate	88.5%	81.9%	82.3%	87.8%

Table 6.4: Experimental results of affect recognition by fusing body and face cues.
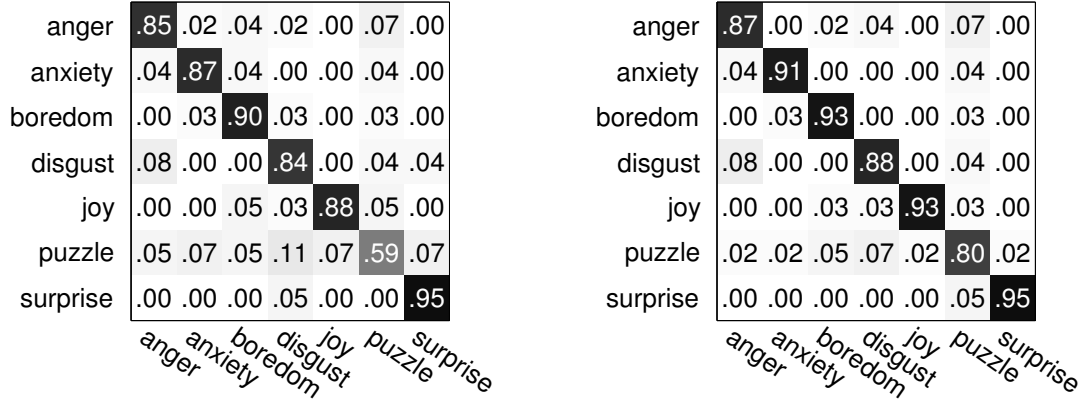
Figure 6.10: Confusion matrices of affect recognition by fusing facial expression and body gesture. (*left*) Direct feature fusion; (*right*) CCA feature fusion.

modality.

6.5 Summary

In this chapter, we investigated two important but understudied problems, gender classification by human gaits and affective body gesture analysis, which have important applications in intelligent visual surveillance and human-computer interaction. With the GEI based gait representation, our experiments illustrate that visual gender recognition from human gaits is very effective. Spatial-temporal features are exploited for representing body gestures in videos. Considering each modality in isolation has its limitations, we present to fuse face and body cues at the feature level using CCA for improved performance. Experiments on large datasets demonstrate that our multimodal systems can achieve the superior recognition performance in gender discrimination and affect analysis.

7 Conclusions

7.1 Conclusions

Human facial and body language play an important and fundamental role in social interpersonal communication, and reveal a large variety of information such as identity, gender, and age. Machine analysis of facial and body language has been emerging as an active research area over the past decades with a number of important applications such as human-computer interaction, visual surveillance, security, computer animation, and medical diagnose, and law enforcement.

In this thesis, we have presented research building towards computational frameworks capable of automatically understanding facial expressions and behavioural body language. In particular, we emphasize the following issues.

Facial Feature Selection and Representation

We first present a thorough examination in issues surrounding facial representation based on statistical local features. The method of Local Binary Patterns is empirically investigated for person-independent facial expression recognition. Different machine learning methods, including template matching, SVM, and linear programming, are systematically examined on several public databases. Extensive experiments illustrate that LBP features are effective and efficient for facial expression recognition.

In real-world environments, the input face images are often in lower resolution. So we also investigate LBP features for low-resolution facial expression recognition. Besides the evaluation on different image resolutions, we performed experiments on real-world compressed

low-resolution video sequences. It is observed that LBP features perform stably and robustly over a useful range of low resolutions of face images, yielding promising performance in compressed video sequences captured in real-world environments.

In order to derive the most effective LBP features from face images for better facial representation, in addition to utilizing AdaBoost, we further present a Conditional Mutual Information based learning procedure. Extensive experiments show the CMI based method enables much faster training, and the best recognition performance is obtained by using SVM classifiers with the selected LBP features.

Manifold based Facial Expression Analysis

Subsequently we present a method to capture and represent the expression dynamics by discovering the underlying low-dimensional manifold. Locality Preserving Projections is exploited to learn the expression manifold in the LBP based dense appearance feature space. One challenging problem in expression manifold learning is to obtain a generalized representation for facial expressions from different subjects. By deriving a universal discriminant expression subspace using the supervised LPP, we effectively align manifolds of different subjects on a generalized expression manifold.

We comprehensively evaluate linear subspace methods, including PCA, LDA, LPP, SLPP, ONPP, and LSDA, using different facial representation on several public databases, Extensive comparative experiments demonstrate that SLPP perform best in expression subspace learning.

We further formulate a Bayesian framework to examine both the temporal and appearance characteristics for dynamic facial expression recognition employing the derived manifold representation. Our method provides superior performance to both the static frame-based methods and the state-of-the-art dynamic models in the person-dependent recognition experiments. We also show that the expression intensity can be easily estimated on the expression manifold using the Fuzzy K-Means method.

Capturing Correlations Among Facial Parts

To have a closer look at the correlations among facial parts, we employ Canonical Correlation Analysis to model correlations of facial parts. To overcome the inherent limitations of the traditional CCA for image data, we introduce a Matrix-based Canonical Correlation Analysis for better correlation analysis of 2D image or matrix data in general. We evaluate the proposed MCCA in capturing correlations of facial parts for facial expression analysis. Experimental results demonstrate that MCCA can better measure correlations in 2D image data, providing superior performance in regression and recognition tasks, whilst requiring much fewer canonical factors.

Multimodal Facial and Body Language Analysis

We investigate two understudied problems in body language analysis, gait-based gender discrimination and affective body gesture recognition. With the GEI based gait representation, it is observed that visual gender recognition from human gaits is very effective. We investigate affective body gesture analysis by exploiting spatial-temporal features for representing body gestures in videos.

We further exploit CCA to fuse the two modalities at the feature level. CCA derives a semantic "gender" or "affect" space, in which the face and body features are compatible and can be effectively fused. Experiments on large dataset demonstrate that our multimodal recognition system achieves the superior recognition performance in gender discrimination and affect analysis.

7.2 Future Work

So far the above issues in facial and body language understanding have been intensively discussed. Although much progress has been made in this thesis, our work still have some limitations. Here we list these limitations and possible directions should be addressed in future work as follow.

- The feasibility of LBP features in real-life situations where free head motion and oc-

clusion exist is still unknown. Moreover,our experiments were carried out on precisely aligned face images. So LBP features should be studied in more realistic setting.

- Our work have been mainly done on the posed facial and body data collected in lab environments. More elaborated studies are necessary for spontaneous facial expression and body behaviour analysis.

- The presented Matrix-based CCA is still a linear technique, so it cannot effectively deal with higher-order statistics among image data. Nonlinear MCCA should be further studied.

- It seems probabilistic graphical models, such as BN and DBNs are well suited for correlating and fusing of different sources of information. An improvement may be achieved if the statistical and probabilistic models are integrated.

- Facial and body language are context-dependent. For better understanding, one must know the context in which the observed signal was displayed. More work should be conducted for context-sensitive facial and body language analysis.

Bibliography

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.

[2] Z. Ambadar, J. W. Schooler, and J. F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005.

[3] N. Ambady and R. Rosenthal. Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, February 1992.

[4] M. A. Amin, N. V. Afzulpurkar, M. N. Dailey, V. E. Esichaikul, and D. N. Batanov. Fuzzy-c-mean determines the principle component pairs to estimate the degree of emotion from facial expressions. In *International Conference on Natural Computation / Fuzzy Systems and Knowledge Discovery*, pages 484–493, 2005.

[5] K. Anderson and P. W. McOwan. A real-time automated system for the recognition of human facial expressions. *IEEE Trans. Systems, Man, and Cybernetics*, 36(1):96–105, February 2006.

[6] M. Argyle. *Bodily Communication (2nd ed.)*. Methuen & Co. Ltd., New York, 1988.

[7] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in man-machine interaction systems. In *Machine Learning for Multimodal Interaction, LNCS 3361*, pages 318–328, 2005.

[8] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and spatial actors in gait perception that influence gender recognition. *Perception & Psychophysics*, 23(2):145–152, 1978.

[9] M. Bartlett, G. Littlewort, M. Frank, C. Lainscseck, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, September 2006.

[10] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spotaneous behavior. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 568–573, 2005.

[11] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Trans. Neural Networks*, 13(6):1450–1464, 2002.

[12] M.S. Bartlett, G. Littlewort, I. Fasel, and R. Movellan. Real time face detection and facial expression recognition: Development and application to human computer interaction. In *IEEE Computer Vision and Pattern Recognition Workshop*, pages 53–53, 2003.

[13] J. N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower area of the face. *Journal of Personality and Social Psychology*, 37(11):2049–2058, 1979.

[14] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Networks*, 5(4):537–550, 1994.

[15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[16] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding

and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591, 2001.

[17] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *British Machine Vision Conference*, pages 797–806, 2002.

[18] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York, 1981.

[19] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linkoping University, SE-581 83 Linkoping, Sweden, 1998. Dissertation No 531.

[20] B. Braathen, M.S. Bartlett, G. Littlewort, E. Smith, and J. R. Movellan. An approach to automatic recognition of spontaneous facial actions. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 231–235, 2002.

[21] R. Brunelli and T. Poggio. Hyperbf networks for gender classification. In *DRAPA Image Understanding Workshop*, pages 311–314, 1992.

[22] J. K. Burgoon, D. B. Buller, and W. G. Woodall. *Nonverbal Communication: The Unspoken Dialogue*. McGraw-Hill, New York, 1996.

[23] J. K. Burgoon, M. L. Jensen, T. O. Meservy, J. Kruse, and J. F. Nunamaker. Augmenting human identification of emotional states in video. In *International Conference on Intelligent Data Analysis*, 2005.

[24] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. In *International Joint Conference on Artificial Intelligence*, pages 708–713, 2007.

[25] Y. Chang, C. Hu, and M. Turk. Mainfold of facial expression. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 28–35, 2003.

[26] Y. Chang, C. Hu, and M. Turk. Probabilistic expression analysis on manifolds. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 520–527, 2004.

[27] I. Cohen, F. Cozman, N Sebe, M. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(12):1553–1566, Dec 2004.

[28] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 2003.

[29] J. F. Cohn. Foundations of human centered computing: Facial expression and emotion. In *International Joint Conference on Artificial Intelligence, Workshop on AI for Human Computing*, pages 5–12, 2007.

[30] J. F. Cohn, L. I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions in spontaneous facial behavior. In *IEEE International Conference on Systems, Man, and Cybernetics*, page 610, 2004.

[31] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1–12, 2004.

[32] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498, 1998.

[33] M. Costa, W. Dinsbach, A. S. R. Manstead, and P. E. R. Bitti. Social presence, embarrassment, and nonverbal behavior. *Journal of Nonverbal Behavior*, 25(4):225–240, 2001.

[34] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004.

[35] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[36] J. E. Cutting, D. R. Proffitt, and L. T. Kozlowski. A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3):357–372, 1978.

[37] G. Dai and D.-Y. Yeung. Tensor embedding methods. In *National Conference on Artificial Intelligence*, pages 330–335, 2006.

[38] C. Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, London, 1872.

[39] J. G. Daugman. Complete discrete 2D Gabor transform by neural networks for image analysis and compression. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, July 1988.

[40] J. W. Davis and H. Gao. An expressive three-mode principal components model for gender recogniton. *Journal of Vision*, 4(5):362–377, 2004.

[41] J. W. Davis and H. Gao. Gender recognition from walking movements using adaptive three-mode pca. In *IEEE Workshop on Articulated and Nonrigid Motion*, pages 9–9, 2004.

[42] B. de Gelder. Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7:242–249, 2006.

[43] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[44] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[45] R. Donner, M. Reiter, G. Langs, P. Peloscheck, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, October 2006.

[46] E. Douglas-Cowie, R. Cowie, and M. Schroder. A new emotion database: Considerations, sources and scope. In *Proceeding of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 39–44, 2000.

[47] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, 2003.

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

[49] P. Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *J. Personality Social Psychol.*, 17(2):124–129, 1971.

[50] P. Ekman and E. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. New York: Oxford Univ. Press, 1997.

[51] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.

[52] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36:259–275, 2003.

[53] X. Feng, M. Pietikäinen, and T. Hadid. Facial expression recognition with local binary patterns and linear programming. *Pattern Recognition and Image Analysis*, 15(2):546–548, 2005.

[54] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, December 2004.

[55] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[56] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems*, pages 572–577, 1991.

[57] G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. Technical report, Stanford, CA, USA, 1992.

[58] H. Gu and Q. Ji. Facial event classification with task oriented dynamic bayesian network. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 870–875, 2004.

[59] H. Gu and Q. Ji. Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, 16:105–115, 2005.

[60] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications (In press)*, 2006.

[61] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *International Conference on Pattern Recognition*, volume 1, pages 1148–1153, 2006.

[62] G. Guo and C. R. Dyer. Simultaneous feature selection and classifier training via linear programming: A case study for face expression recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 346–352, 2003.

[63] A. Hadid, M. Pietikäinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 797–804, 2004.

[64] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(2):316–322, February 2006.

[65] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[66] X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, 2003.

[67] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using LaplacianFaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3):328–340, Mar 2005.

[68] U. Hess, S. Blairy, and R. E. Kleck. The intensity of emotional facial expression and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4):241–257, 1997.

[69] J. Hoey and J. J. Little. Value directed learning of gestures and facial displays. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1026–1033, 2004.

[70] H. Hotelling. Relations between two sets of variates. *Biometrika*, 8:321–377, 1936.

[71] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Taipei, 2003.

[72] C. Hu, Y. Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *International Workshop on Face Processing in Video*, pages 81–81, 2004.

[73] A. Jaimes and N. Sebe. Multimodal human computer interaction: A survey. *Computer Vision and Image Understanding*, in press.

[74] A. K. Jain and S. Z. Li, editors. *Handbook of Face Recognition.* Springer, 2005.

[75] M. Jones and P. Viola. Face recognition using boosted local features. Technical report, MERL, 2003.

[76] A. Kale, A. K. R. Chowdhury, and R. Chellappa. Fusion of gait and face for human identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 901–904, 2004.

[77] R. E. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *IEEE Computer Vision and Pattern Recognition Workshop*, pages 154–154, 2004.

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

[79] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *ACM International Conference on Multimedia*, pages 677–682, 2005.

[80] A. Kapoor, Y. Qi, and R. Picard. Fully automatic upper facial action recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 195–202, 2003.

[81] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *IEEE International Conference on Computer Vision*, pages 166–173, 2005.

[82] T-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *European Conference on Computer Vision*, pages 251–262, 2006.

[83] S. Kimura and M. Yachida. Facial expression recognition and its degree estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 295–300, 1997.

[84] E. Kokiopoulou and Y. Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence (In press)*, 2007.

[85] L. T. Kozlowski and J. E. Cutting. Recognizing the sex of a walker from dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

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

[87] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision*, pages 432–439, 2003.

[88] L. Lathauwer, B. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[89] C. S. Lee and A. Elgammal. Facial expression analysis using nonlinear decomposable generative models. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 17–31, 2005.

[90] L. Lee and W. E. L. Grimson. Gait analysis for recognition and classification. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 155–162, 2002.

[91] S. Z. Li and Z. Zhang. Floatboost learning and statistical face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(9):1–12, 2004.

[92] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surfaces for recognition. *International Journal of Computer Vision*, 53(1):71–92, 2003.

[93] S. Liao, W. Fan, C. S. Chung, and D.-Y. Yeung. Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features. In *IEEE International Conference on Image Processing*, pages 665–668, 2006.

[94] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li. Subtly different facial expression recognition and expression intensity estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 853–859, 1998.

[95] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, June 2006.

[96] G. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Joint Symposium on Neural Computation*, pages 1–1, 2006.

[97] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial

images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[98] D. P. Mandic, D. Obradovic, A. Kuh, T. Adali, U. Trutschell, M. Golz, P. D. Wilde, J. Barria, A. Constantinides, and J. Chambers. Data fusion for modern engineering applications: An overview. In *International Conference on Artificial Neural Networks*, pages 715–721, 2005.

[99] G. Mather and L. Murdoch. Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society: Biological Sciences*, 258(1353):273–279, 1994.

[100] H. Meeren, C. Heijnsbergen, and B. Gelder. Rapid perceptual integration of facial expression and emotional body language. *Procedings of the National Academy of Sciences of USA*, 102(45):16518–16523, November 2005.

[101] A. Mehrabian. Communication without words. *Psychology Today*, 2(4):53–56, 1968.

[102] T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 39(9):1961–1973, 2003.

[103] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90–126, 2006.

[104] B. Moghaddam and M. Yang. Learning gender with support faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(5):707–711, May 2002.

[105] S. Mota and R. W. Picard. Automated posture analysis for detecting learner's interest level. In *IEEE Computer Vision and Pattern Recognition Workshop*, pages 49–49, 2003.

[106] V. Nayak and M. Turk. Emotional expression in virtual agents through body language. In *International Symposium on Visual Computing*, pages 313–320, 2005.

[107] R. J. Neagle, K. Ng, and R. A. Ruddle. Studying the fidelity requirements for a virtual ballet dancer. In *Vision, Video and Graphics*, pages 181–188, 2003.

[108] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference*, pages 1249–1258, 2006.

[109] M. S. Nixon and J. N. Carter. Automatic recognition by gait. *Proceedings of The IEEE*, 94(11):2013–2024, November 2006.

[110] M. S. Nixon, T. Tan, and R. Chellappa. *Human Identification based on Gait*. Springer, 2005.

[111] T. Ojala, M Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition*, 29(1):51–59, 1996.

[112] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[113] N. Oliver, A. Pentland, and F. Berard. Lafter: a real-time face and lips tracker with facial expression recognition. *Pattern Recognition*, 33:1369–1382, 2000.

[114] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):812–816, May 2005.

[115] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. In K. Kurihara, editor, *Face Recognition*, pages 377–416. Advanced Robotics Systems, Vienna, Austria, 2007.

[116] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Systems, Man, and Cybernetics*, 36(2):433–449, April 2006.

[117] M. Pantic, A. Pentland, A. Nijholt, and T.S. Huang. Human computing and machine understanding of human behavior: A survey. In T.S. Huang, A. Nijholt, M. Pantic, and A. Pentland, editors, *Artificial Intelligence for Human Computing*, volume 4451, pages 47–71. Springer, 2007.

[118] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: the state of art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[119] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. *Image and Vision Computing*, 18(11):881–905, 2000.

[120] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceeding of the IEEE*, volume 91, pages 1370–1390, 2003.

[121] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Systems, Man, and Cybernetics*, 34(3):1449–1461, June 2004.

[122] M. Pantic, N. Sebe, J. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *ACM International Conference on Multimedia*, pages 669–676, 2005.

[123] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*, pages 317–321, 2005.

[124] L. K. Paragas, J. Varin, and R. Berenter. Use of gait patterns to reveal possible disorders in the geriatric patient. *Journal of American Podiatric Medical Association*, 90(4):183–93, April 2000.

[125] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, USA.

[126] F. E. Pollick, J. W. Kay, K. Heim, and R. Stringer. Gender recognition from point-light walkers. *Journal of Experimental Psychology: Human Perception and Performance*, 31(6):1247–1265, 2005.

[127] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceeding of the IEEE*, 77(2):257–285, 1989.

[128] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Trans. Image Processing*, 15:3349–3361, 2006.

[129] P. Ravindra De Silva and N. Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterizatino of posture features. *Computer Animation and Virtual Worlds*, 15:169–276, 2004.

[130] P. Ravindra De Silva, M. Osano, and A. Marasinghe. Towards recognizing emotion with affective dimensions through body gestures. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 269–274, 2006.

[131] M. Reiter, R. Donner, G. Langs, and H. Bischof. 3D and infrared face reconstruction from RGB data using canonical correlation analysis. In *International Conference on Pattern Recognition*, pages 425–428, 2006.

[132] D. Ridder, O. Kouropteva, O. Okun, M. Pietikainen, and R. P. W. Duin. Supervised locally linear embedding. In *International Conference on Artificial Neural Networks and Neural Information Processing*, pages 333–341, 2003.

[133] J. A. Russell. Is there univeral recognition of emotion from facial expression. *Psychological Bulletin*, 115(1):102–141, 1994.

[134] S. Sarkar, Phillips P. J., Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The HumanID gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(2):162–177, February 2005.

[135] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[136] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Maching Learning*, 37(3):297–336, 1999.

[137] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang. Authentic facial expression analysis. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 517–522, 2004.

[138] G. Shakhnarovich and T. Darrell. On probabilistic combination of face and gait cues for identification. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 169–174, 2002.

[139] G. Shakhnarovich, P. A. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 14–21, 2002.

[140] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. In N. Sebe, M. S. Lew, and T. S. Huang, editors, *Computer Vision in Human-Computer Interaction*, LNCS 3723, pages 221–230. Springer, 2005.

[141] C. Shan, S. Gong, and P. W. McOwan. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference*, volume 1, pages 399–408, Oxford, UK, September 2005.

[142] C. Shan, S. Gong, and P. W. McOwan. Recognizing facial expressions at low resolution. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 330–335, Como, Italy, September 2005.

[143] C. Shan, S. Gong, and P. W. McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing*, volume 2, pages 370–373, Genoa, Italy, September 2005.

[144] C. Shan, S. Gong, and P. W. McOwan. A comprehensive empirical study on linear subspace methods for facial expression analysis. In *IEEE Computer Vision and Pattern Recognition Workshop*, pages 153–158, New York, USA, June 2006.

[145] C. Shan, S. Gong, and P. W. McOwan. Dynamic facial expression recognition using a

bayesian temporal manifold model. In *British Machine Vision Conference*, volume 1, pages 297–306, Edinburgh, UK, September 2006.

[146] C. Shan, S. Gong, and P. W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. In *British Machine Vision Conference*, Warwick, UK, September 2007.

[147] C. Shan, S. Gong, and P. W. McOwan. Capturing correlations among facial parts for facial expression analysis. In *British Machine Vision Conference*, Warwick, UK, September 2007.

[148] C. Shan, S. Gong, and P. W. McOwan. Learning gender from human gaits and faces. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, London, UK, September 2007.

[149] J. Skelley, R. Fischer, A. Sarma, and B. Heisele. Recognizing expressions in a new database containing played and natural expressions. In *International Conference on Pattern Recognition*, pages 1220–1225, 2006.

[150] M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *International Joint Conference on Pattern Recognition*, pages 408–410, 1978.

[151] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, Dec 2000.

[152] Y. Tian. Evaluation of face resolution for expression analysis. In *International Workshop on Face Processing in Video*, pages 82–82, 2004.

[153] Y. Tian, L. Brown, A. Hampapur, S. Pankanti, A. Senior, and R. Bolle. Real world real-time automatic recognition of facial expression. In *IEEE workshop on performance evaluation of tracking and surveillance*, Australia, March 2003.

[154] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[155] Y. Tian, T. Kanade, and J. Cohn. *Handbook of Face Recognition*, chapter 11. Facial Expression Analysis. Springer, 2005.

[156] Y. Tong, W. Liao, and Q. Ji. Inferring facial action units with causal relations. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1623–1630, 2006.

[157] N. F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):371–387, 2002.

[158] M. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1991.

[159] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Computer Vision and Pattern Recognition Workshop*, page 149, 2006.

[160] M. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *International Conference on Multimodal Interfaces*, pages 162–170, 2006.

[161] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection from face video. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, pages 635–640, 2004.

[162] M. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *IEEE Computer Vision and Pattern Recognition Workshop*, pages 76–84, 2005.

[163] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[164] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *IEEE International Conference on Computer Vision*, pages 281–288, 2003.

[165] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[166] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 1692–1698, 2005.

[167] J. J. Wang and S. Singh. Video analysis of human dynamics - a survey. *Real-Time Imaging*, 9(5):321–346, 2003.

[168] Y. Wang, H. Ai, B. Wu, and C. Huang. Real time facial expression recognition with adaboost. In *International Conference on Pattern Recognition*, pages 926–929, 2004.

[169] M. Weiser. The computer for the twenty-first century. *Scientific American*, 265(3):94–104, 1991.

[170] J. Whitehill and C. Omlin. Harr features for facs AU recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 97–101, 2006.

[171] Y. Wu and T. Huang. Human hand modeling, analysis and animation in the context of human computer interaction. *IEEE Signal Processing Magazine*, 18(3):51–60, 2001.

[172] Y. Yacoob and L. S. Davis. Recognizing human facial expression from long image sequences using optical flow. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[173] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang. Multilinear discriminant analysis for face recognition. *IEEE Trans. Image Processing*, 16(1):212–220, 2007.

[174] J. Yang, A. F. Zhang, D. Frangi, and J. Yang. Two dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(1):131–137, 2004.

[175] J. Ye. Generalized low rank approximations of matrices. *Machine Learning*, 61:167–191, November 2005.

[176] M. Yeasin, B. Bullot, and R. Sharma. From facial expression to level of interests: A spatio-temporal approach. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 922–927, 2004.

[177] J.-H. Yoo, D. Hwang, and M. S. Nixon. Gender classification in human gait using support vector machine. In *Proceedings of Advanced Concepts for Intelligent Vision Systems*, pages 138–145, 2005.

[178] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition*, pages 441–444, 2006.

[179] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, August 2006.

[180] G. Zhang, X. Huang, S. Z. Li, Y. Wang, and X. Wu. Boosting local binary pattern (lbp)-based face recognition. In *Chinese Conference on Biometric Recognition*, pages 179–186, 2004.

[181] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):1–16, May 2005.

[182] Z. Zhang, M. J. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 454–461, 1998.

[183] X. Zhou and B. Bhanu. Human recognition at a distance in video by integrating face profile and gait. In *International Conference on Audio and Video-Based Person Authentication*, pages 533–543, 2005.

[184] X. Zhou and B. Bhanu. Integrating face and gait for human recognition. In *IEEE Computer Vision and Pattern Recognition Workshop*, pages 55–55, 2006.

[185] C. Zou, N. Sun, Z. Ji, and L. Zhao. 2DCCA: A novel method for small sample size face recognition. In *IEEE Workshop on Application of Computer Vision*, pages 43–43, 2007.