SPATIO-TEMPORAL FRAMEWORK AND ALGORITHMS FOR VIDEO-BASED FACE RECOGNITION

JOHN SEE SU YANG

DOCTOR OF PHILOSOPHY
MULTIMEDIA UNIVERSITY
MAY 2014
SPATIO-TEMPORAL FRAMEWORK
AND ALGORITHMS FOR
VIDEO-BASED FACE RECOGNITION

BY

JOHN SEE SU YANG

B.Eng. (Hons), M.EngSc., Multimedia University, Malaysia

THESIS SUBMITTED IN FULFILMENT OF THE
REQUIREMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
(by Research)
in the
Faculty of Computing and Informatics

MULTIMEDIA UNIVERSITY
MALAYSIA

May 2014
The copyright of this thesis belongs to the author under the terms of the Copyright Act 1987 as qualified by Regulation 4(1) of the Multimedia University Intellectual Property Regulations. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this thesis.

© John See Su Yang, 2014
All rights reserved
DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

________________________

John See Su Yang
ACKNOWLEDGEMENTS

I am indebted to the following people who have journeyed with me through the years.

My most heartful thanks goes to my wife, Mei Sin, for her constant love, encouragement and patience. Many achievements in this work, including getting across the line would not been possible without her sacrifice and support. Thanks also to little Joshua, who did a big favour in expediting the long overdue process of getting this thesis finished. I am also grateful to my parents for being the utmost pillar of support and providence in times of need.

A word of thanks to the following people whose guidance and scholarship are a spring of inspiration to this work. Thanks to Prof. Eswaran for being a steadfast supervisor, an advisor and an inexhaustible source of clarity, thought and wisdom. His careful reading of this thesis resulted in quite a bit of red ink. Thanks to Faizal for being very dependable and helpful in many ways, and for always asking the right question. Thanks to Andrew Teoh for embracing my thoughts in research and lighting the path to this work.

A quick word of gratitude to many others who have made my Ph.D study a little more bearable. Thanks to fellow pursuers of the same destination for the company: To Peter Ho for quick navigation through administrative potholes, to Ian Tan for after-five tea conversations, and to Albert for necessary enthusiasm in almost anything. Thanks to Raphael for believing in what I can do then and now, to Kuan Hoong for good stories, to Lai Kuan for asking all the good questions. Thanks also to Koinonia cellgroup for all the prayers and encouragement through the years.

Above all, I thank God for His blessings and faithfulness at all times.

"It is good to have an end to journey towards; but it is the journey that matters, in the end" —Ursula K. Le Guin
To my parents, my wife, and my son.
ABSTRACT

Face recognition is one of the most active research areas in computer vision and pattern recognition, having drawn tremendous attention in the last few decades. A variety of practical applications such as biometrics, visual surveillance, multimedia retrieval and consumer electronics has greatly benefited from it. However, under unconstrained environments and complex facial variations, conventional approaches seem to falter. These challenges, together with the emergence of video media, have given rise to a rapidly growing area of video-based face recognition (VFR). Despite the abundance and ubiquity of video data, existing VFR approaches are not without flaws, often degenerating into conventional image-based recognition or under-utilizing the richness of features and properties found in videos. This severely limits its potential, both theoretically and performance-wise.

The objective of this research is to devise algorithms for video-based face recognition based on spatio-temporal manifolds. The primary focus is to better represent nonlinear face patterns on a data manifold, while incorporating spatio-temporal information that is inherent in videos. In addition, a systematic approach to evaluate sampling is also crucial as current experimental setups are inadequate to provide balanced evaluation. These goals are inspired by motivations from computer vision, psychology and cognitive neuroscience.

The research contributions of this thesis are five-fold. Firstly, a novel supervised manifold learning algorithm, Neighborhood Discriminative Manifold Projection (NDMP) is proposed for feature extraction step to better characterize complex nonlinear facial variations across identities in a video manifold. In the prior clustering step, a new Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC) approach is designed to exploit the inherent ordering of video frames by means of interweaving the distances in both spatial and temporal dimensions, resulting in meaningful clusters and exemplars. Next, classification of faces in videos is performed using two newly proposed maximum-a-posteriori (MAP) classifiers: Exemplar-Driven Bayes (EDB) clas-
sifier that encodes the causal relationships within an exemplar set, and Dual-Feature Bayes (DFB) classifier that captures relevant dependencies between features of the exemplar set and image set sub-manifolds. Fourthly, two frameworks are introduced for VFR— a more conventional exemplar-based framework and a feature-rich cluster-centric framework, which is an entirely novel proposition. Both frameworks successfully integrate spatio-temporal characteristics and manifold representations into various stages of their pipelines. Finally, on the experimental setup, an augmented test set generation protocol is formalized to introduce balanced evaluation by obtaining a variety of query subsequences with different starting frame positions and frame lengths.

From the experimental evaluation conducted, the proposed algorithms demonstrate superior performance in terms of recognition accuracy, in comparison to existing approaches in literature. The NDMP algorithm is clearly an effective feature extraction method as shown by the good recognition rates achieved, surpassing other linear and nonlinear manifold representation techniques with an almost perfect recognition rate on one of the datasets. In the clustering of video data, the STHAC algorithm is capable of producing meaningful clusters, resulting in a vast improvement over spatial-only methods. Significantly, this result is also consistent across different feature representations. The DFB classifier under a cluster-centric framework further extends the recognition performance, as evident from a comprehensive set of experiments which also investigates the effect of subsequence length and computational cost. Although the DFB easily outperforms its predecessor, the EDB classifier, there are minor concerns over the longer feature training time. Overall, these results show that there is promising potential in harnessing the strengths of spatio-temporal information and the elegant manifold representation that models the complex facial variations in videos. Without further doubt, the principles expounded in this thesis can be extended to a range of problems such as object recognition in general, video surveillance and video retrieval.
# TABLE OF CONTENTS

COPYRIGHT PAGE ii  
DECLARATION iii  
ACKNOWLEDGEMENTS iv  
DEDICATION v  
ABSTRACT vi  
TABLE OF CONTENTS viii  
LIST OF TABLES xi  
LIST OF FIGURES xii  
List of Symbols and Abbreviations xvii  

## CHAPTER 1: INTRODUCTION 1

1.1 Overview 1  
1.1.1 Video-based Face Recognition 3  
1.2 Motivations and Problem Statements 5  
1.2.1 Machine Vision Perspective 5  
1.2.2 Psychological and Neural Perspective 9  
1.2.3 Experimental Perspective 12  
1.2.4 Research Questions 14  
1.3 Research Objectives 14  
1.4 Scope of Thesis 15  
1.5 Contributions 15  
1.6 Organization of Thesis 17  

## CHAPTER 2: LITERATURE REVIEW 19

2.1 Non-temporally Motivated Approaches 22  
2.1.1 Assembly-based Approaches 23  
2.1.2 Exemplar-based Approaches 26  
2.1.3 Image Set-based Approaches 31  
2.2 Temporally Motivated Approaches 40
LIST OF TABLES

Table 4.1 Average recognition rates (%) and confidence intervals of various manifold learning methods on the evaluated datasets 89

Table 5.1 Optimal values for parameters used in STHAC-GF and STHAC-LP 108

Table 5.2 Average recognition rates (%) of various exemplar extraction methods with different feature representations 115

Table 5.3 Average recognition rates (%) of various exemplar extraction methods on different datasets (with variable-length subsequences) 118

Table 6.1 Parameter settings and subsequence lengths used in the generation of augmented test set for Experiment I 148

Table 6.2 Average recognition rates (%) and standard deviation of the evaluated methods across all three datasets for Experiment I 149

Table 6.3 Parameter settings and subsequence lengths used in the generation of augmented test set for Experiment II 151

Table 6.4 Average recognition rates (%) of selected evaluated methods for Experiment II, with larger variety of subsequence lengths (larger $T$ values) 152

Table 6.5 Computational cost of selected methods, showing modeling and classification time averaged across all subsequences with Honda/UCSD (10, 5, 20) 154
LIST OF FIGURES

Figure 1.1 Comparison of different biometric modalities with respect to its compatibility for machine readable travel document (MRTD) system, reproduced from (Heitmeyer, 2000) 2

Figure 1.2 Quality (universality and performance) vs availability (acceptability and collectability) of different image-based biometric modalities, reproduced from (Gorodnichy, 2005) 4

Figure 1.3 Three stages of a generic video-based face recognition (VFR) pipeline. The clustering stage (first block) is typically used in the training phase if required. 6

Figure 1.4 VFR configurations denoting image-to-image (left), video-to-video (middle) and image-to-video (left) matching tasks for face recognition 7

Figure 1.5 A collection of face images mapped onto a low-dimensional manifold space, as described by the first two coordinates of Locally Linear Embedding (LLE) algorithm. Figure reproduced from (Roweis & Saul, 2000). 9

Figure 1.6 A model for mapping recognition of moving faces onto neural systems by O’Toole, Roark, and Abdi (2002), following the distributed neural system for face perception originally proposed by Haxby, Hoffman, and Gobbini (2000) as a framework for understanding psychological findings regarding the effects of facial motion on memory for faces. Figure reproduced from (O’Toole et al., 2002). 12

Figure 2.1 Taxonomy of video-based face recognition (VFR) approaches reviewed in this chapter 21

Figure 2.2 Distribution of face data from a video sequence of a single person in LLE embedding space, reproduced from (Fan, Wang, & Tan, 2005) 28

Figure 2.3 Mutual subspace method (MSM), reproduced from (Yamaguchi, Fukui, & Maeda, 1998) 36

Figure 2.4 A complex and nonlinear face appearance manifold can be approximated as a union of several simpler pose manifolds, reproduced from (K. Lee, Ho, Yang, & Kriegman, 2005) 44

Figure 2.5 Identity surfaces defining face patterns of two subjects, reproduced from (Y. Li, Gong, & Liddell, 2003) 45

Figure 2.6 Temporal HMM modeling of a sample face sequence, reproduced from (X. Liu & Chen, 2003) 46

Figure 2.7 A face sequence depicted as a rectangular prism (left) and an example of 3D neighborhood of a pixel in Volume LBP (right), reproduced from (Hadid & Pietikäinen, 2009a) 48
Figure 3.1 Exemplar-based framework for video-based face recognition. The red stars indicate tasks where novel approaches are proposed in this thesis. 53

Figure 3.2 Cluster-centric framework for video-based face recognition. Tasks involving both point and cluster features are shown on the left and right branches respectively. The red stars indicate tasks where novel approaches are proposed in this thesis. 55

Figure 3.3 Illustration of cluster-centric image set and exemplar features constructed from the training data 56

Figure 3.4 Illustration of cluster-centric image set and exemplar features constructed from the test data 57

Figure 3.5 Sample video frame shots of three near-frontal views from the CMU MoBo dataset 58

Figure 3.6 Sample video frame shots from the Honda/UCSD dataset 59

Figure 3.7 Sample video frame shots from the NICTA ChokePoint dataset, showing subjects captured with near-frontal view with Camera 1 (left images), entering Portal 1 (top right) and leaving Portal 1 (bottom right). 60

Figure 3.8 Sequentially ordered image frames of a sample video subsequence from the evaluated datasets. These face images were detected, cropped and pre-processed according to the described procedure. 62

(a) CMU MoBo 62
(b) Honda/UCSD 62
(c) NICTA ChokePoint 62

Figure 3.9 Illustration of how the augmented test set is generated from the original video sequences. The example above is based on the protocol parameters \( \{W, T, C\} = \{3, 3, 20\} \). Due to limitation in display, note that all \( W \) subsequences of a certain same length and class are randomly sampled from the original test video, and thus are not the same. 63

Figure 3.10 Incremental experimentation methodology 65

Figure 4.1 Spectral plots of data points from a sample video sequence in different embedding spaces, described by their respective first three coordinates. 78

(a) PCA 78
(b) MDS 78
(c) Isomap 78
(d) LLE 78
Figure 4.2 A data point (black) is embedded in NDMP-space using weights from intra-class (blue) and inter-class (red) neighborhoods. Solving the constrained minimization problem leads to distinction between intra-class and inter-class structures. 81

Figure 4.3 Comparison of cumulative match characteristic (CMC) curves of various manifold learning methods on the CMU MoBo dataset. 90

Figure 4.4 Comparison of cumulative match characteristic (CMC) curves of various manifold learning methods on the Honda/UCSD dataset. 91

Figure 4.5 Comparison of different values of $\beta$ on the Honda/UCSD dataset. 93

Figure 5.1 Linear cluster patches (in different colors) in LLE space after applying HAC for clustering 101

Figure 5.2 A sample dendrogram of $M = 11$ clusters. To select $M$ clusters, the dendrogram collapses the lower branches of the tree to show no more than $M$ leaf nodes representing the clusters. The height of each partition represents the distance between each two connected clusters. 103

Figure 5.3 Illustration of how data points might be grouped 104  
(a) Spatial 104  
(b) Temporal 104  
(c) Spatial-temporal 104

Figure 5.4 Graphical example of building common spatio-temporal neighbor (CSTN) and ¬CSTN sets based on spatial neighborhood size of $k = 2$ and temporal neighborhood size of $w = 2$ 106

Figure 5.5 Residual error plot of three different training videos (drawn with different colors) partitioned with different number of clusters. The "elbow" of the curve is approximately at values 5 to 9. 109

Figure 5.6 Selected exemplars of CMU MoBo (top three rows) and Honda/UCSD (bottom three rows) datasets extracted using the STHAC-GF method, showing the most representative faces of each person. 114

Figure 5.7 Cumulative Match Characteristic (CMC) curve of various exemplar extraction methods. Probabilistic voting strategy is applied for classification with NDMP features. 116

Figure 5.8 Performance of STHAC-GF (Honda/UCSD, variable-length subsequence experiment) using different $\alpha$ values with $M$ number of exemplars extracted. 118

Figure 5.9 Performance of STHAC-LP (Honda/UCSD, fixed-length subsequence experiment) with different $k$ and $w$ parameters, in comparison with HAC. 119
Figure 5.10 Cluster spectral plot of the 'hide' video (from Honda/UCSD dataset) in LLE manifold space. Clusters are color-coded with exemplars for each cluster shown. Adjacent video frames are connected by lines.

(a) STHAC-GF
(b) STHAC-LP
(c) HAC

Figure 6.1 Cluster-centric scheme for VFR illustrating the extraction of subspace and point sub-manifold features from a training video sequence. In terms of feature representation, cluster image sets are characterized by linear subspaces while the exemplar set consists of exemplar images from across all classes projected onto NDMP-space.

Figure 6.2 Graphical models depicting the dependencies of the proposed Exemplar-Driven Bayes (EDB) and Dual-Feature Bayes (DFB) classifiers

(a) EDB
(b) DFB

Figure 6.3 Graphical illustration of similarity metrics involving exemplar points (circles) and image set subspaces (colored regions) within local class-specific clusters. Two sample classes are shown, each with only two extracted clusters.

Figure 6.4 Selected exemplars of two subjects from CMU Mobo (top two rows), Honda/UCSD (bottom two rows) and ChokePoint (middle two rows) datasets.

Figure 6.5 Recognition rates of various classification methods on different datasets for Experiment I. The last group of bars indicate the average recognition rates across all three datasets.

Figure 6.6 Performance of selected methods in Experiment II based on different test subsequence lengths

(a) Honda/UCSD
(b) ChokePoint-LL

Figure 6.7 Posterior plot of a subsequence sampled from the 'rakesh' video of the Honda/UCSD dataset. Posterior probabilities of the seven most probable subjects are shown in different colors. The subject (in blue line) is correctly identified at the end of the subsequence.

Figure 6.8 Comparisons between the posterior plots of the EDB (blue line) and NB (red line) approaches on a subsequence sampled from the (a) 'chia' video and (b) 'louis' video of the Honda/UCSD dataset.

(a) Subsequence from 'chia' video
(b) Subsequence from 'louis' video
Figure 6.9  Comparison of cumulative match characteristic (CMC) curves of various classification methods evaluated in Experiment II on the three datasets.

(a) Honda/UCSD dataset
(b) CMU MoBo dataset
(c) ChokePoint-LL dataset
LIST OF SYMBOLS AND ABBREVIATIONS

\( C \) \hspace{1cm} \text{Number of classes/subjects}

\( M \) \hspace{1cm} \text{Number of exemplars, or number of clusters}

\( N \) \hspace{1cm} \text{Number of samples (face images)}

\( \Theta \) \hspace{1cm} \text{A set of temporally ordered face image segments (from test data)}

\( E \) \hspace{1cm} \text{A set of exemplar face images}

\( X \) \hspace{1cm} \text{A sequence of face images from video}

\( Z \) \hspace{1cm} \text{A set of face image clusters (from training data)}

CMC \hspace{1cm} \text{Cumulative Match Characteristic}

CSTN \hspace{1cm} \text{Common Spatio-Temporal Neighbor}

DCC \hspace{1cm} \text{Discriminative Canonical Correlations}

DFB \hspace{1cm} \text{Dual-Feature Bayes}

EDB \hspace{1cm} \text{Exemplar-Driven Bayes}

HAC \hspace{1cm} \text{Hierarchical Agglomerative Clustering}

LDA \hspace{1cm} \text{Linear Discriminant Analysis}

LLE \hspace{1cm} \text{Locally Linear Embedding}

MAP \hspace{1cm} \text{maximum-a-posteriori}

MMD \hspace{1cm} \text{Manifold-Manifold Distance}

MSM \hspace{1cm} \text{Mutual Subspace Method}

NB \hspace{1cm} \text{Naive Bayes}

NDMP \hspace{1cm} \text{Neighborhood Discriminative Manifold Projection}
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPE</td>
<td>Neighborhood Preserving Embedding</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>STHAC</td>
<td>Spatio-Temporal Hierarchical Agglomerative Clustering</td>
</tr>
<tr>
<td>STHAC-GF</td>
<td>STHAC-Global Fusion</td>
</tr>
<tr>
<td>STHAC-LP</td>
<td>STHAC-Local Perturbation</td>
</tr>
<tr>
<td>VFR</td>
<td>Video-based Face Recognition</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Face recognition is one of the most active research areas in computer vision and pattern recognition, generating tremendous interest and development in the last few decades. Practically, its usefulness is very much far-reaching than its most obvious and general use for person identification and verification. A host of other technological applications such as visual surveillance, human-computer interaction, multimedia retrieval and annotation, law enforcement, consumer electronics and computer entertainment, have all benefited from the use of face recognition (T. Huang, Xiong, & Zhang, 2011). More recently, popular online social networking services such as Facebook (2011) and Google (2013), with their large crowd-sourced collection of images, have begun incorporating name-tagging features that apply face recognition techniques to their respective photo archives.

1.1 Overview

The act of identifying people from faces is almost an effortless task for humans. Nonetheless, the computer identifies people from faces by means of learning and interpreting patterns such as shapes, color or other characteristics of a person’s face captured digitally in image or video form. A face is regarded as a biometric (National Institute of Standards and Technologies (NIST), 2010), or a "biological measure" that uniquely identifies a person through various physical or behavioral patterns or traits. In comparison to other biometric modalities such as fingerprint, iris, hand geometry, the face enables identification to be performed naturally and unobtrusively, with very little user cooperation required. In a work by Heitmeyer (2000) that assessed the compatibility of different biometrics used in a machine readable travel document (MRTD) system, face biometric ranks first among a list of six modalities: face, finger, hand, voice, eye and signature (see Figure 1.1). Six criteria were reportedly used in this
evaluation: enrollment, renewal, machine-assisted identity verification requirements, redundancy, public perception, and storage requirements and performance. This lends credence to the advantages of using face biometric in a wide range of applications.

The development and advancement of face recognition in still images have been comprehensively documented in a few notable surveys (Chellappa, Wilson, & Sirohey, 1995; W. Zhao, Chellappa, Phillips, & Rosenfeld, 2003) in the past decades. A wide variety of established methods proposed through the years have become core algorithms in the area of face recognition today, and they have been proven successful in achieving good recognition rates primarily in still image-based scenarios. These methods include Eigenfaces (Turk & Pentland, 1991), Fisherfaces (Belhumeur, Hespanha, & Kriegman, 1997), elastic bunch graph matching (EBGM) (Wiskott, Fellous, Kuiger, & Von Der Malsburg, 1997) and active appearance models (AAM) (Cootes, Walker, & Taylor, 2000). These well-known methods for image-based face recognition have stood the test of time, many with over a thousand academic citations in literature. These landmark techniques have since also evolved into further kernel and bayesian

![Figure 1.1: Comparison of different biometric modalities with respect to its compatibility for machine readable travel document (MRTD) system, reproduced from (Heitmeyer, 2000)](image-url)
variants (M.-H. Yang, 2002; Moghaddam, Jebara, & Pentland, 2000) that are present state-of-the-art algorithms that have been proven successful in both academic research and commercial applications.

After much progress in almost three decades of research, current state-of-the-art approaches (for still images) are able to obtain high recognition accuracies, with significantly good results under controlled environments. Typical controlled settings include the use of images taken from near-perfect passport/ID photos, face images captured with limited pose angles with relatively well-balanced lighting, and also manual handpicking of suitable images for system enrollment and training. Several problems arise when the acquired image is less than ideal. Under unconstrained environments, there may be significant facial variability, i.e. complex face poses, 3-D head orientations, unexpected expressions, unbalance illumination, and also external circumstances, i.e. occlusion by objects or other faces, poor quality image, or an imposter’s attempt to use a replicated face photo instead of his/her live face. In such cases, a single image may be severely limiting in terms of providing sufficient information for reliable and accurate recognition.

1.1.1 Video-based Face Recognition

The emergence of video has enabled a huge amount of face information to be captured, stored and analyzed. Spurred on by availability of cheap video cameras and ever-increasing processing power, video data has become an easily attainable source of biometric information – perfectly suited to address the limitations of using still images for face recognition. Hence, the ubiquitous nature of video data has presented a new fast-growing area of research in video-based face recognition, or VFR in short.

One of the most important features of a biometric is its ability to collect information from non-cooperating persons (Heitmeyer, 2000), or more concisely, its collectability property. In the work by Gorodnichy (2005), the author proposed for a distinct differentiation to be made between photographic (image-based) facial data and video-acquired facial data in terms of modalities for biometric recognition. It was
argued that image-based facial data is considered a hard biometric trait due to its nature of acquisition in very controlled conditions, resulting in very informative data, but at the expense of an increased intrusiveness and difficulty in collection. On the contrary, video-acquired facial data are very accessible and non-intrusive; similar to how collecting information such as eye, skin colour, height and walking style are easy to accomplish from a distance. However, the quality of information collected is much worse, although the abundance of it makes up for the loss. Figure 1.2 shows how the different biometric modalities stack up against each other and how the two observed factors (quality and availability) are mutually compensating. This proposition sets apart traditional facial data obtained from still images and facial data from video as two different modalities; the latter a result of today’s increasing availability of video media.

The Face Recognition Grand Challenge (FRGC) (Phillips et al., 2005), a research initiative designed to spur researchers towards new directions, stipulated an experiment to investigate the effect of multiple still images on recognition performance under controlled environment. Although the results compiled indicate that solutions to practical face recognition applications can be improved with multiple facial images only, a small sample of four images were captured in a controlled recording session to construct the target and query image sets. This setup does not facilitate the collection of unconstrained continuous video data, therefore it does not seem to constitute "face

Figure 1.2: Quality (universality and performance) vs availability (acceptability and collectability) of different image-based biometric modalities, reproduced from (Gorodnichy, 2005)
recognition in video” intrinsically. Fundamentally, image data in the form of videos or video sequences, can be simply regarded as a temporally ordered set of images (although it appears to be widely acceptable in literature to regard the use of videos in a non-temporally-ordered fashion, i.e. sets of images, as VFR. Chapter 2 discusses this at length).

The notion of having more information is not always beneficial or advantageous. In this age where videos are becoming increasingly available and accessible, it may become computationally intractable to store and search a significant amount of video data. In high frame rate videos, sequentially-ordered faces may appear very similar from frame to frame, and methods of filtering and grouping the sheer amount of data becomes vital in reducing the redundancy of data (Mau, Chen, Sanderson, & Lovell, 2010). Thus, this has also presented new challenges and opened up new directions for research in the area of VFR.

1.2 Motivations and Problem Statements

In this section, the motivations for delving into this area of work are discussed from three distinct viewpoints—namely from the machine vision perspective (computer), psychological and neural perspective (human), and experimental perspective (practical).

1.2.1 Machine Vision Perspective

The work by Zhou (2004) brought new insights into the process of face recognition through the use of more than one still image (multiple still images or video sequences) to three distinct properties: Multiple observations, temporal continuity or dynamics, and 3-D model. The author noted that the first and third properties are shared by multiple still images and video sequences, but the second property solely belongs to video sequences.

As patterns in computers, a video sequence is simply a collection of still images, thus existing traditional face recognition approaches can be applied directly to
each individual image in a video, in an assembly manner. In this scheme, each still image is processed separately using single-image-based algorithms, and the output results are combined using various combining rules such as additive, multiplicative, and so on. With the inclusion of the temporal dimension, there are many computational approaches that seek to model the temporal patterns encoded among the spatial patterns.

A generic VFR framework pipeline consists of three stages: clustering, feature extraction and representation, and classification (Figure 1.3). The first stage, clustering, is normally performed only in the training phase and is dependent on requirements of the approach. Typically, the redundancy in video data can be reduced by selecting a smaller set of representative face images by grouping them into clusters. The second stage, feature extraction and representation, focuses on how face patterns are best extracted and represented. Here, the representation of features may possibly result in a further reduction in data dimensionality. The final stage, classification, deals with the task of estimating the best or closest possible identity match for a new test/query video. A one-shot recognition process is also possible by combining stages two and three into a single pass process.

The abundance of images in video poses a methodological challenge. While the traditional still image-based face recognition is a straightforward matching of a test image to a gallery of training images, i.e. an image-to-image recognition task, it is an ill-posed problem for video sequences: which image from the training video is to be matched with images from the test video?

---

1In the abbreviated recognition settings, the first and third words denote the type of data (image or video) used in the training/gallery set and test/probe set respectively.

Figure 1.3: Three stages of a generic video-based face recognition (VFR) pipeline. The clustering stage (first block) is typically used in the training phase if required.
There are two common configurations used to accomplish a complete video-to-video setting; both are feasible but not without their drawbacks. One approach is to further simplify the problem to an image-to-video recognition task, whereby each training video is represented by a set of exemplars (Krüeger & Zhou, 2002; Hadid & Peitikäinen, 2004; W. Liu, Li, & Tang, 2006). This opens up further questions as to how these exemplars are to be selected, and how much loss of information can be afforded from the summarization of video data. Also, the use of exemplars may potentially oversimplify the complex variations present in videos. Secondly, from the purist point of view, video-to-video matching can be intrinsically achieved as it is. This involves characterizing full length videos (every frame of a moving face) using various state-space and probabilistic models (X. Liu & Chen, 2003; Aggarwal, Chowdhury, & Chellappa, 2004; M. Kim, Kumar, Pavlovic, & Rowley, 2008). Due to issues relating to overfitting and underfitting an enormous amount of data, a good deal of works (Fan & Yeung, 2006a; R. Wang, Shan, Chen, & Gao, 2008; Harandi, Sanderson, Shirazi, & Lovell, 2011) cast it as a learning problem over image sets, whereby matching can be performed much quicker at the expense of temporal dynamics between frames. Figure 1.4 shows a pictorial description of the three configurations discussed here.

There are clear advantages and drawbacks of both exemplar-based and image set-based methods. The challenge here is, how can we harness the advantages of both exemplar (point) and image set (manifold) features within a VFR framework pipeline to improve recognition of faces?

On the representation of features, there is recent increased of attention to-

---

**Figure 1.4:** VFR configurations denoting image-to-image (left), video-to-video (middle) and image-to-video (left) matching tasks for face recognition
wards manifold-based representations for VFR as opposed to traditional linear subspace representations. This is regardless of the manner in which the recognition task is configured, and the usage (or non-usage) of temporal information in its process (Arandjelović, Shakhnarovich, Fisher, Cipolla, & Darrell, 2005; K. Lee et al., 2005; R. Wang et al., 2008). Generally speaking, a manifold is a topological space or surface that is locally Euclidean (i.e. for every point in the space, there is a neighborhood that is homeomorphic to Euclidean space). Images can be visualized as points embedded in a high-dimensional Euclidean space that can be sufficiently drawn on a low-dimensional manifold (Brand, 2003). As such, a face video contains a collection of face images, or sample points, that lie on a certain appearance manifold, a concept first introduced by Murase and Nayar (1995) to characterize the "appearance" of images (parameterized by pose and illumination) by projecting the data samples to a lower-dimensional eigenspace. In video data which possesses a large range of complex facial variations, proper mapping of a nonlinear face manifold to a low-dimensional spatial embedding is essential for subsequent recognition tasks. Figure 1.5 shows a collection of face images mapped to a low-dimensional manifold space using the Locally Linear Embedding (LLE) algorithm proposed by Roweis and Saul (2000).

On using an appropriate manifold representation, this point is reflective of the first factor highlighted in an analysis on video-to-video recognition by Poh et al. (2010). The analysis suggested that the performance of VFR by using entire video sequence data for recognition depends on several crucial factors: an appropriate manifold representation, a distance metric between two manifolds, and the length of video frames. While it may be computationally more efficient than image-to-image matching, the difference in performance is not found to be significant unless these factors are further addressed with improved techniques and schemes. The second factor on good selection of a distance metric between two manifolds is also important. Although significant work has been done in the area of distance metric learning (L. Yang & Jin, 2006), it was pointed out that a range of practical issues have not been well addressed, while theoretically, there appears to be no clear runaway choice. Feasible solutions to the choice of distance metric depend heavily on the task it is applied to (dimensionality
reduction, clustering, classification, retrieval etc.), availability of class labels (supervised, unsupervised etc.), structuring of data (global, local etc.) and the importance of computational efficiency.

Based on these premises, the primary attention is to address this: How can we better represent challenging face data in video by means of a good manifold representation? The essentiality of the first factor far outweighs the second one, since the choice of distance metric is dependent on the way feature matching (classification) is done within the VFR framework used eventually.

1.2.2 Psychological and Neural Perspective

In this work on machine recognition of faces in video, it is unavoidable to discuss on the motivations derived from psychology and cognitive neuroscience, the psychophysical elements of human recognition.
In a recent summarization of findings in neuroscience on the way humans recognize faces (Sinha, Balas, Ostrovsky, & Russell, 2006), computational research is encouraged to look for insights and draw inspiration from the biological process of recognizing faces as a step towards eventually translating them into machine algorithms (thus, giving rise to what is now commonly known as biologically-motivated models or methods). Among the nineteen important results compiled in this comprehensive paper, two points were relevant to the direction of this research work: (i) view-generalization appears to be mediated by temporal association (Result 13), and (ii) motion of faces appears to facilitate subsequent recognition (Result 14).

Firstly, a work by (Wallis & Bülthoff, 2001) seems to suggest that temporal association serves as a "perceptual glue" that binds different images of the same object into a useful representation on a whole. Humans appear to continuously associate views of objects to support later recognition, and this is accomplished not only by physical similarity (note: spatial image features), but also their correlated appearance time (note: temporal video features). Also, close temporal association of new images viewed in sequence is sufficient to induce infero-temporal (IT) neurons to respond similarly to arbitrary image pairs (Miyashita, 1993). These findings seemed to suggest that temporal proximity of images provides crucial information for establishing object representations. Hence, the recognition performance suffers when representations build a view-invariant memory of an object without considering the temporal context.

Secondly, information for identifying a human face can be found in both the invariant structure of features and in idiosyncratic movements and facial gestures (O’Toole et al., 2002). Rigid motion (faces arbitrarily captured from multiple viewpoints) can facilitate recognition of familiar faces, but there is no significant improvement of using these multiple views during the learning or memorization phase. Meanwhile, non-rigid motion (where facial expressions, emotions or speech), or "dynamic facial signatures" as what is visually observed in a short frame of time, seems to play a greater role in facilitating recognition of faces (Knight & Johnston, 1997), with increased significance when the motion is distinctive (Lander & Chuang, 2005). Biologically, the media tem-
poral cortex of a human brain performs motion processing, which aids the recognition of dynamic facial signatures. In principle, the work by (O’Toole et al., 2002) concluded that facial dynamic information is found to contribute greatly to recognition under degraded viewing conditions and also when a viewer’s experience with the same face increases. When both static and dynamic facial information are available, psychological evidence points to these principles in balancing the contributions of both kinds of input. The authors went on to propose an enhanced model for mapping recognition of moving faces onto neural systems (see Figure 1.6). The fusiform face area (FFA) contains invariant facial information useful for identifying faces while the posterior superior temporal sulcus (pSTS) processes changeable aspects of movements of faces. The ventral stream (bottom half, comprising of ‘Static-based information’ and ‘FFA’ blocks) processes static structures of the face while the dorsal stream (top half, comprising of ‘Facial motion’, ‘MT’, ‘STS’ blocks) takes care of social processing including facial motion. This generally suggests that the two distinct parts of the human neural system have their respective roles in processing visual information from faces.

Inspired by these findings, researchers in computer vision and pattern recognition have attempted to improve machine recognition of faces by utilizing video sequences, where temporal dynamics is an inherent property. Current temporally-motivated approaches have laboured to propose meaningful schemes. This include approaches that directly model temporal transitions between face appearances, and approaches that utilize mutually-benefitting simultaneous tracking and recognition. (A more detailed review of approaches can be found in Chapter 2.)

**How can we better exploit the spatio-temporal information in video data to improve recognition of faces?** This generic question is further unpacked into more precise queries, based on the motivations culminating from this perspective:

i) How can we learn view-generalization for faces in video by promoting close temporal association? The way existing algorithms build view-based appearance manifolds only consider the spatial information (physical structure) present, and not the temporal information (correlated appearance time).
Figure 1.6: A model for mapping recognition of moving faces onto neural systems by O’Toole et al. (2002), following the distributed neural system for face perception originally proposed by Haxby et al. (2000) as a framework for understanding psychological findings regarding the effects of facial motion on memory for faces. Figure reproduced from (O’Toole et al., 2002).

ii) How can we model "recognition over time", or how classification of face identities can be more accurate when more stimuli (of a familiar face) are presented?

1.2.3 Experimental Perspective

The emergence of face video datasets in the last decade also contributed to a growing interest in VFR. Several key datasets such as the CMU Motion of Body (MoBo) (Gross & Shi, 2002) and Honda/UCSD (K. Lee et al., 2005) have since been established as benchmark databases widely used by the research community. More contemporary datasets such as the recently collected NICTA ChokePoint (Wong, Chen, Mau, Sanderson, & Lovell, 2011) dataset for person identification under real-world surveillance conditions have also been found its way into VFR experiments.
From the many VFR works surveyed in this thesis, there are two distinct methods of how videos are evaluated in experiments. Frame-based evaluation considers a video in a frame-by-frame manner, while video-based evaluation considers the whole video sequence as a single observation, or as a face manifold defined by a set of face images. Frame-based evaluation has numerous drawbacks which are undesirable. Experimentally, frame by frame processing reduces video data into an assembly of individual image classification tasks. In other words, each face image will be classified independently of other images in the same video, thus removing any possibility of utilizing temporal relationships to assist recognition. Performance-wise, some works (Zhou, Krüeger, & Chellappa, 2003; Stallkamp, Ekenel, & Stiefelhagen, 2007) have also reported that frame-based evaluation is consistently outperformed by video-based evaluation in recognition accuracy since the increase of data helps to resolve some ambiguities.

While a majority of video-based approaches claimed to be superior in terms of computation and result, it is observed that video-based evaluation procedures are often reported with little clarity, unbalanced in terms of video sequence lengths (skewed towards long sequences only due to evaluation on entire videos), or involving manual pre-selection (or omission) of training or test face images. Some works attempted to employ practical schemes that allow for test sequences to be randomly chosen (X. Liu & Chen, 2003; Hadid & Peitikäinen, 2004) or training images that are augmented with virtual samples (Stallkamp et al., 2007), but they somewhat lack the extensiveness and consistency in providing a balanced evaluation.

In the analysis of VFR performance by Poh et al. (2010), the third factor highlighted—the length of video data, remains largely an issue of experimental setup. There is scant attention towards the study of the influence of video length with respect to recognition. It is commonly assumed that recognition accuracy improves with longer video sequences, but this is not necessarily true when videos contain a large amount of complex facial variations. How can we formulate a systematic approach to experimental sampling and evaluation for video-based face recognition?
1.2.4 Research Questions

There are multiple problems that are to be addressed in this thesis. The following points are a recapitulation of research questions that arose from the motivations from the machine vision, experimental, psychological, and neural perspectives:

1. How can we harness the advantages of both exemplar (point) and image set (manifold) features within a VFR framework pipeline to improve recognition of faces?

2. How can we better represent challenging face data in video by means of a good manifold representation?

3. How can we better exploit the spatio-temporal information in video data to improve recognition of faces?

4. How can we formulate a systematic approach to experimental sampling and evaluation for video-based face recognition?

1.3 Research Objectives

The principal goal of this thesis work is to devise algorithms for video-based face recognition based on spatio-temporal manifolds. By that, a VFR framework that adopts manifold representations, strengthened by spatio-temporal characteristics of video data is to be proposed. Based on the four research questions presented previously, this thesis work sets out to accomplish the following research objectives:

1. To leverage on both exemplar and image set features in video data to improve face recognition performance

2. To design a good manifold representation that can better characterize complex nonlinear facial variations in video

3. To exploit the advantages of spatio-temporal information that is inherent in video data and incorporating it into various steps of the recognition process

4. To formulate a systematic approach to experimental sampling and evaluation for video-based face recognition
1.4 Scope of Thesis

This work done in this thesis covers a range of tasks in a generic VFR pipeline *i.e.* clustering, feature extraction and classification, as illustrated earlier in Fig. 1.3. Throughout this work, it is assumed that the face(s) (or more precisely, the face regions) that are present in each video sequence have been duly extracted prior to the recognition task, which remains the focal point of this work. Hence, video scenes containing multiple faces are assumed to have all faces individually detected, tracked and pieced into a sequence of frames corresponding to different persons. Investigation of methods involving face detection and other preparatory steps leading to the recognition task (such as pre-processing, illumination normalisation and resampling) lies outside the scope of this work.

The work in this thesis is also constrained to video sequences in a third-person surveillance environment, captured from a moderate-distance field-of-view (FOV), and not "in the wild" settings such as in the Labeled Faces in the Wild (LFW) dataset (G. B. Huang, Ramesh, Berg, & Learned-Miller, 2007), or from a mobile device with a close-up FOV such as the Mobile Biometry (MOBIO) dataset (McCool et al., 2012). Consequently, this influences the choice of video data used in this work (see Section 3.3 for the datasets used). Moreover, the investigation of facial dynamics is also limited to the facial transitions between frames, and not detailed facial region movements found on the face.

1.5 Contributions

This thesis encompasses a collection of algorithms, frameworks and protocol, that are designed on the notion of capturing spatio-temporal patterns within a face video manifold. The intrinsic spatio-temporal nature of videos allows for the proposal of these techniques to further enrich the way how patterns in video data can be better manipulated and exploited. The contributions of this work are summarized as follows, in no particular order of significance as they are to be taken together complementarily.

1. Two frameworks for video-based face recognition—*exemplar-based and cluster-*
centric frameworks, are proposed, integrating spatio-temporal characteristics and manifold representations into various steps in their pipelines. The usage of spatio-temporal information leverages both the clustering and classification steps. Also, manifold embedding is applied to extracted face images while exemplar features are projected to a discriminative manifold space. While the exemplar-based framework is a conventional pipeline, the cluster-centric framework is an entirely novel proposition, utilizing the rich set of sub-manifold features (which are extracted from clusters) comprising both point (exemplar image) features and subspace (image set) features.

2. A novel supervised manifold learning algorithm, Neighborhood Discriminative Manifold Projection (NDMP), is proposed for feature extraction and representation in VFR. The NDMP constructs a discriminative eigenspace projection of the video face manifold based on the intrinsic structure of both intra-class and inter-class neighborhood exemplar points. NDMP is neighborhood-preserving and discriminative in nature; thus enabling it to maximize separation between classes while retaining the topology of the original data manifold.

3. A new spatio-temporal clustering algorithm, Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC), is devised to exploit the inherent ordering of video frames by means of the distances in both spatial and temporal dimensions. Two variants are introduced: A global fusion variant (STHAC-GF) that blends spatial and temporal distances between points in space, and a local perturbation variant (STHAC-LP) that perturbs spatial and temporal distances based on a local neighborhood criterion. Good selection of clusters are essential to build meaningful view-based sub-manifolds for subsequent tasks.

4. Two novel probabilistic classification approaches based on Bayesian maximum-a-posteriori estimation is proposed. Firstly, the Exemplar-Driven Bayes (EDB) classifier is proposed to model the causal relationship between an exemplar and its associated class-specific exemplar set. Secondly, the Dual-Feature Bayes (DFB) classifier leverages features of the exemplar set (point features) and image set (subspace features) sub-manifolds, which are derived from the appearance-
based clusters extracted from the face video manifold. In both approaches, a joint probability function is designed to elegantly capture relevant dependencies among the features using similarity metrics.

5. An augmented test set generation protocol is formalized to introduce thorough experimentation by systematic sampling of a variety of query subsequences with different starting frame positions and frame lengths. This reduces biases and better mimics realistic scenarios with arbitrary sets of facial views. Majority of works in VFR literature do not apply random sampling of subsequences from original test sequences. Also, it is common for evaluation procedures to be rather unorganized and not explicitly described, leading to much ambiguity over the cogency and consistency of results reported.

1.6 Organization of Thesis

The rest of the thesis is organized as follows: Chapter 2 presents a comprehensive taxonomical review on various existing strategies and approaches for VFR. The surveyed works are grouped based on the usage or non-usage of the temporal property in video sequences, and then further categorized into subgroups that have more distinct characteristics. Chapter 3 provides a concise overview of the proposed frameworks and video datasets utilized in this work, and also defines the general mathematical notation used for the algorithms. The augmented test set generation protocol for sampling test video subsequences is also elaborated in detail.

The next three chapters (4-6) discuss the core novelties proposed for different stages of the VFR frameworks: Chapter 4 presents the NDMP, a novel supervised manifold learning algorithm for feature extraction and representation in VFR. Some related literature and motivations that led to the idea are provided as well. Experiments are conducted to evaluate the performance through comparative evaluation and rank-based identification. Chapter 5 introduces STHAC, a new spatio-temporal clustering algorithm that incorporates temporal information into the partitioning of meaningful clusters. Two variants of the STHAC are discussed— a global fusion variant and a local
perturbation variant. Experiments are conducted on fixed-length and variable-length subsequences, followed by further discussions on the empirical choice of parameters. **Chapter 6** presents a new dual-feature approach to probabilistic classification of faces in video, whereby image set and exemplar features derived from cluster sub-manifolds are used to capture both variational and appearance-based information respectively. Comprehensive experiments are carried out to demonstrate the effectiveness of recognition while an assessment of computational cost is also thoroughly deliberated.

Finally, **Chapter 7** provides the concluding remarks to this thesis, together with a brief summary of the limitations of this work and some future directions worth pursuing.
CHAPTER 2

LITERATURE REVIEW

Since the advent of the widely popular face recognition surveys by (Chellappa et al., 1995) and W. Zhao et al. (2003), which exclusively focused only on image-based face recognition, there is no single authoritative survey on the specific area of video-based face recognition (VFR) despite its far-reaching advances in recent years.

There was a period of interest in 3-D\(^1\) approaches to face recognition. Two surveys (Bowyer, Chang, and Flynn (2006); Abate, Nappi, Riccio, and Sabatino (2007)) comprehensively reviewed approaches and challenges in 3-D face recognition, including approaches that combine both modalities (2D and 3D) in an attempt to improve performance baseline. There were other similar surveys tailored to emphasize on specific type of approaches such as, face recognition from a single image per person (Tan, Chen, Zhou, & Zhang, 2006), face recognition using local binary pattern approaches (Ahonen, Hadid, & Pietikäinen, 2006), and illumination invariant face recognition (Zou, Kittler, & Messer, 2007).

A very recent attempt to survey VFR literature by H. Wang, Wang, and Cao (2009) covers some significant ground in terms of different stages within a face recognition system (including face detection and tracking approaches), but it appears to be broad-based and does not seem to taxonomically categorize the approaches in an orderly manner. For instance, there are a number of "statistical model-based" approaches that could very well be "spatio-temporal information" approaches as well. Three other brief surveys (Zhou, 2004; Matta & Dugelay, 2009; Z. Zhang, Wang, & Wang, 2011) presented a better organized taxonomical viewpoint of the approaches in literature, but they do not encompass the entire rich collection of works available in literature. The

\(^1\)The third dimension of the ‘3-D’ here refers to depth or range data that is added to constitute a face model, not temporal information as found in videos.
latest compilation of VFR literature by Barr, Bowyer, Flynn, and Biswas (2012) is the most detailed and comprehensive attempt so far, dichotomizing all approaches into two significant groups—set-based methods and sequence-based methods, which are then further divided into smaller sub-groups. They also took the effort to summarise the results collectively (with no regard to the databases used) in tables based on their respective categories.

Most of these surveys view from the perspective of how video information is utilized, i.e. multiple frame observations (without temporal information), temporal dynamics across frames, creation of 3-D models. In addition to this categorization, there is a notable lack of attention for a proper categorization of "exemplar-based" VFR methods, which involves the usage of exemplar features. On the other hand, many of these surveys also do not specifically distinguish between approaches that utilize an entire image set as a single feature, and approaches that utilize image exemplars as features. In (Barr et al., 2012), "image set-based" VFR methods are viewed in a rather generic way (where exemplar-based methods are also included but scattered across various sub-categories), by focusing on the way the multitude of observations are exploited without temporal consideration. This group of methods is undeniably popular in recent years of VFR research. Nevertheless, approaches that utilize features from these two categories can fall within the scopes of both non-temporally or temporally motivated groups, a majority of which tend to be associated with the former due to the assumption that temporal information has been stripped off.

For the purpose of clarity, categorization based on the properties of multiple still images or video sequences suggested by Zhou (2004) (and to a lesser extent, (Barr et al., 2012)) is adopted in this review. The first two properties ("Multiple observations" and "Temporal continuity") can be simply denoted as non-temporally motivated and temporally motivated methods, respectively. In addition to that, the non-temporally motivated category is further expanded into feature-oriented sub-categories such as assembly of images, exemplar and image set approaches. On the other hand, the temporally motivated category is divided into two main sub-categories that are of immedi-
ate interest to this research—state-space and appearance-based manifold approaches, while a large assortment of other temporally motivated approaches are lumped together in a separate sub-category. The third property ("3D model") presented in (Zhou, 2004) falls outside the scope of this research and will not be reviewed in this thesis. Figure 2.1 shows the flowchart depicting the categorization of various reviewed VFR approaches.

The extensiveness of this work in covering various stages of the VFR pipeline necessitates further review on approaches that are most relevant to each problem scope covered in the three main chapters on proposed algorithms (i.e. Chapters 4, 5 and 6). Hence, this chapter provides a broad view of the VFR task as a whole, with brief description and remarks on selected notable and landmark works in literature. Overall, this is neither an attempt at an exhaustive review of the fast-expanding VFR area of research, nor the only authoritative way of categorization, as can be seen from the different styles meted out by previous surveys.

In this chapter, the two main branches in the taxonomy of methods – non-temporally motivated approaches and temporally motivated approaches, are reviewed in Sections 2.1 and 2.2 respectively. Approaches are further categorized into sub-branches and discussed within their respective sections. At the end, a summary of this review is presented in Section 2.3 together with important observations that provide key directions of this research.

![Figure 2.1: Taxonomy of video-based face recognition (VFR) approaches reviewed in this chapter](image)

21
2.1 Non-temporally Motivated Approaches

The extrinsic nature of video sequences or multiple still images (or image sets) allows it to be deconstructively regarded as a single still image in a degenerate, fragmented manner. Thus, it is not difficult to apply conventional single image-based face recognition algorithms by means of constructing an *assembly-based* approach (Zhou, 2004). This can be easily accomplished by merely re-iterating the algorithm through all still images in the entire video sequence, while performing recognition on each input single image. With that, a reductionistic strategy of frame selection or fusion of multiple frames is required to achieve recognition of persons in a video sequence. The assembly-based approach is the most simplistic arrangement necessary to perform a complete recognition task on a video sequence, albeit having clearly stripped away the intrinsic temporal information. In some works, the exploration of various feature representations become the focus of attention since the combination of recognition results across frames can be trivially determined ad-hocly.

Also, various methods of summarizing video sequences or multiple still images have also been proposed in literature, many of which are able to produce reasonably good recognition performance. One popular scheme involved the selection of exemplars or representative face images that closely characterizes the subject in a training video sequence, in what is categorized as *exemplar-based* approaches. Another group of works rely on the direct use of multiple still images or image sets (not necessarily in an ordered manner), which can be directly extracted from video sequences in entirety or by further selection techniques. In this review, these works are referred to as *image set-based* approaches.

Thus, the non-temporally motivated approaches can be loosely divided into three distinct categories, as assembly-based, exemplar-based and image set-based approaches.
2.1.1 Assembly-based Approaches

Assembly-based approaches were mainly introduced at the turn of the millennium, at the advent of the establishment of video-based face recognition (VFR) as a new promising area of research. This ubiquity and abundance of video data led to the adaptation of conventional still image-based face recognition techniques for the recognition of faces in video. By assuming video data as an "assembly" of single image frames, the crossing from image domain to video domain is almost effortless and easily implementable. A key factor that determines the effectiveness of assembly-based approaches is the question of how multi-frame information (in video data) can be combined or manipulated to improve recognition of faces in video.

The most straightforward solution is to consider all frames while applying an appropriate matching evaluation scheme for recognition. In an early work, Satoh (2000) proposed a direct extension of the traditional Eigenface (Turk & Pentland, 1991) and Fisherface (Belhumeur et al., 1997) methods, by introducing a simple distance measure for matching video sequences. The "face sequence distance" between videos is computed by considering the smallest distance between frame pairs (one from each video) in the reduced feature space. Moreover, this scheme also turns out to be inefficient when classifying long video sequences. Generally, the reported recognition accuracy was unsatisfactory (at levels of 35-55%) as the proposed distance measure appeared to be prone to classification errors. In another similar extension scheme (Torres, Lorente, & Vila, 2000), faces were first manually selected from training sequences to construct person-wise PCA spaces called "self-eigenfaces" with the objective of recognizing specific person faces that are found within a test video sequence. Then, test images from the video sequences are projected and reconstructed using each set of the different self-eigenfaces. Verification is performed on a frame-by-frame basis by matching with the face (subject) that minimizes the reconstruction error. Generally, commonly used subspace projection methods include Principal Component Analysis (PCA) (or Eigenface (Turk & Pentland, 1991)), Linear Discriminant Analysis (LDA) (or Fisherface (Belhumeur et al., 1997; Etemad & Chellappa, 1997)), Independent Component Analysis (ICA) (Bartlett, Lades, & Sejnowski, 1998), and Local Feature
Alternatively, a combined matching evaluation can be achieved by multi-frame fusion. The recognition results (or similarity scores) across all evaluated frames are aggregated to yield a final recognition result for the whole video sequence. This frame-by-frame fusion procedure can take on different combination rules such as sum, product and majority vote rules, as mooted in the seminal discourse on classifier combination by Kittler, Hatef, Duin, and Matas (1998). One early example is a template-based approach to video-based face recognition proposed by Price and Gee (2001), where a parallel system of classifiers, called "observers" is used to characterize and classify different facial regions. A simple sum rule is used to fuse the classification scores produced from comparisons made in different eigenfeature observer spaces. The underlying mechanism still operates a frame-by-frame matching of each query image with a set of randomly selected training face images. Their experiments were carried out using a few standard face image databases in an attempt to construct image sets of faces. This may not necessarily mimic the distribution of face views found in a realistic video sequence. However, there are some positive motivations that can be drawn from this work – the use of multiple fused classifiers can greatly improve baseline classifiers, and that the fusing by sum rule is both effective and resilient to estimation errors when appropriately applied.

Some later works employ various innovative schemes to deal with multiple frames for the improvement of recognition results. A real-time video-based system that estimates the identity of a video sequence by progressively combining the confidence scores from the classification of individual frames was introduced in (Stallkamp et al., 2007). While its classification is performed using typical appearance-based algorithms (a discriminative \(k\)-nearest neighbor approach and a generative Gaussian Mixture Model (GMM) approach), the innovation lies in the proposal of new distance measures that seek to encourage temporal fusion and reduce the impact of ambiguous frames by weighing the influence of frames with respect to the overall classification decision.
To offer more robustness when dealing with multiple frames, selection schemes can be useful to measure the contribution of each face in an assembly of images and assign appropriate weights to these face images based on their quality. A weighted probabilistic approach proposed in (Y. Zhang & Martínez, 2006) introduced an algorithm that selects and weights the images of a sequence based on its usefulness. Faces are partitioned into smaller sub-regions represented by linear subspaces and Gaussian mixtures, which can robustly mitigate occlusions and localization errors. Contributions of these sub-regions and the entire face are weighted, and the class conditional probabilities of all corresponding sub-regions in each probe-gallery pair of images determines the matched class. On a 50-subject database containing three neutral expression images per subject for training and two 40-frame test videos for every subject, their approach achieved a remarkable recognition rate of 99% using LDA features.

Mian (2008) proposed an unsupervised learning approach that first extracts local SIFT features from video face images. Similarities between pairs of faces are measured by taking the number of matching SIFT vectors and the angle between each set of vectors. A similarity score is then derived from a weighted average of these measures. In the subsequent hierarchical clustering process, which partitions the feature data into a pre-determined number of clusters containing faces of related appearances. Finally, a voting-based algorithm is applied to pick a representative set of features for the clusters. During recognition process, the test video frames are sequentially matched to the clusters belonging to all persons in the database and the best temporally cohesive cluster gives the matched identity.

In another interesting novel approach, a face from video is recognized by extracting "face patches" in a frame by frame manner, which are then matched to an overall face model and stitched together (Hu, Harguess, & Aggarwal, 2009). This approach uniquely treats the VFR problem by reconstructing the full face from the patches extracted from multiple frames before performing recognition. As a result, problems such as variation in pose and challenging occlusions can be handled through the accumulation of patches. Their experiment on video sequences yielded promis-
ing results even in the event of missing data in the reconstructed face due to severe occlusion or non-detection of face.

2.1.2 Exemplar-based Approaches

A video sequence, in its purest form, may contain many samples that are very similar in terms of facial appearances, particularly consecutive frames in order. Hence, the usage of all available frames in a video may sometimes be an overkill – a computational burden and potential over-generalization in terms of data representation.

Representative frames or a reduced subset of selected informative frames also known as exemplars, is a widely used feature extracted to summarize video sequences or sets of multiple frames for video-based face recognition. Typically, the large set of image frames available in the training videos warrants the selection of a set of informative frames to represent the distinctive facial appearances (pose, expression, illumination, etc.) of each subject class in the gallery set. Naturally, this reduction of data representation usually meant that classification or matching scores between the exemplars (from training data) and the test data requires a combination rule (as mentioned in (Kittler et al., 1998)) to realize a complete video-to-video recognition. There are a significant number of notable exemplar-based VFR works in recent literature, many of which, center upon the clustering technique used in the organization of data for the purpose of exemplar selection. Much of this section is dedicated to a review of techniques used in VFR approaches where exemplars are extracted as the main video feature.

In one of the early works that launched the development of this trend (Krüeger & Zhou, 2002), the authors outlined a major problem that persists in the case of video data (in the light of conventional image-to-image recognition), where a set of "mug-shots" per person is now available: How should one select the right mug-shots? The use of exemplars is proposed as a good solution to tackle this problem, as was previously applied in some earlier works on video sequence tracking (Toyama & Blake, 2001) and a variety of clustering applications (Frey & Jojic, 2000). Theoretically,
The selection of face exemplars should ideally be based on minimizing the expected distance between exemplars and the video frame samples, as noted by Krüeger and Zhou (2002). By using a radial basis function (RBF) neural network approach, an "online" frame-by-frame technique was proposed to learn the exemplars in video. The exemplars are selected by seeking the centers of the clusters produced through the online learning scheme. The exemplars are then used as centers of mixture models representing each trained person during the recognition task.

The $k$-means clustering (Duda, Hart, & Stork, 2000) remains a popular clustering technique in computer vision literature. Hadid and Peitikäinen (2004) proposed a view-based scheme which seeks to partition data in a projected embedding (space) using $k$-means clustering. A state-of-the-art nonlinear dimensionality reduction method, Locally Linear Embedding (LLE) (Roweis & Saul, 2000) is used to project the training data to a low-dimensional linear embedding before clustering is performed. Likewise, the cluster centers extracted from each training video are taken as a person-specific set of exemplar faces. A probabilistic voting strategy was employed to combine image-to-image matching task over all frames to arrive at a final recognition decision. The primary weakness of the $k$-means clustering method is that it produces sub-optimal clustering, where the assignment of initial seeds will greatly influence the formation of clusters, i.e. with random initialization of seeds, every round of clustering may produce different clusters. Hence, it is particularly not robust against outliers and the algorithm is likely to be trapped in the local optimum.

The notion of extracting view-based clusters (or in other terms in literature describing the similar item—*submanifolds* or *local models*) from a low-dimensional embedding of a complex nonlinear video face manifold began to gain increasing attention. Subsequent approaches (Fan et al., 2005; Fan & Yeung, 2006b) also made use of nonlinear dimensionality reduction methods such as LLE and Isometric Feature Mapping (Isomap) (Tenenbaum, de Silva, & Langford, 2000) to effectively uncover the intrinsic dimensionality of nonlinear data patterns in video. Figure 2.2 shows an example of the distribution of face data (points) from a video sequence in LLE embed-
Further motivated by the Isomap algorithm, Fan and Yeung (2006a, 2006b) estimated the geodesic distances between sample images on a face manifold. A hierarchical agglomerative clustering (HAC) method is used on the training set to extract face clusters for each subject. With the image features represented in the original image space or using traditional subspace methods such as PCA and LDA, majority voting is applied to combine the output of different frames in the evaluated test set. Experiments conducted on a 40-subject video dataset demonstrate a marked improvement in recognition capability by clustering with the HAC algorithm using geodesic distances. Unlike $k$-means clustering, the HAC algorithm produces optimal clustering while it also allows one to decide what level or scale of clustering is most appropriate to specific applications at hand. Unlike the $k$-means algorithm, the HAC does not require initial selection of seeds while hierarchical merging ensures that the resulting clusters are optimal.

Figure 2.2: Distribution of face data from a video sequence of a single person in LLE embedding space, reproduced from (Fan et al., 2005)
R. Wang et al. (2008) introduced clustering by Maximal Linear Patch (MLP), which are local linear patches on the face manifold. Inspired by geometric intuition, the MLP is spanned by a maximal linear subspace and its linear perturbation is reflected by a nonlinearity degree measured by the deviation between the Euclidean distances and geodesic distances in the patch. A potential drawback lies in the inconsistency of patch size, whereby the sequential one-shot manner of the algorithm may benefit the patches computed earlier while those computed later might be smaller in size. Later on, the authors extended their work to explore the use of Hierarchical Divisive Clustering (HDC) (R. Wang & Chen, 2009), which is purportedly more efficient than the bottom-up HAC as a complete construction of the hierarchy all the way down to individual samples is often not necessary. This is because the appropriate number of clusters is much smaller than the number of data samples in most cases. In both works mentioned, the use of exemplars (which are taken as the sample means of the local models) is pivotal to the computation of distance between two manifolds.

The usefulness of clustering for face exemplar selection also found its place in large scale face recognition in web videos (M. Zhao, Yagnik, Adam, & Bau, 2008). In this work, "key faces" or face exemplars are selected by clustering each "face track" which contains a sequence of faces derived earlier by facial feature tracking. The HAC algorithm is also utilized here for clustering. To mitigate the situation that one person could appear several times in a video, the face tracks are first clustered by HAC algorithm (where similarity between two tracks can be determined as the maximum similarity of the key faces), followed by the key faces belonging to all tracks within the same cluster. This robust scheme acts to obtain representative key faces while at the same time, remove duplicated key faces encountered earlier. Finally, the recognition process is completed by fusing both majority voting and probabilistic voting. An extensive evaluation on large-scale web videos (up to 1,500 hours of video containing 44,453 tracks) achieved an average top-5 precision of 80% on tested persons.

The criteria of selecting a subset of informative frames (or as similarly described earlier as representative frames or exemplars) does not rely only on its frame
locality in space or time dimension. Some other methods attempt to obtain the exemplar set by exploiting other forms of information.

Tang and Li (2004) proposed to use the audio signals in video to align video sequences belonging to different persons. In other words, frames containing similar facial expressions (from different persons) are synchronized corresponding to the associated speech. This was done manually by using the speech signals of the recited sentences to locate frames with distinctive expressions. This was possible due to the fact that all videos contained the same recited sentence. To match the spatio-temporally synchronized video sequences, a multi-level discriminant subspace analysis algorithm was developed where overall recognition is achieved through majority voting or sum rule on a frame-by-frame classification. Later on, W. Liu et al. (2006) extended this work to implement an additional synchronized frame clustering to partition video frames into relevant clusters, an approach that outputs aligned clusters across all video sequences in an incremental manner. To achieve this, a spatio-temporal objective function is plugged into a classical \( k \)-mean clustering algorithm to consider both spatial distance and temporal order. Then, homogenous "keyframes", or exemplars are learnt by a novel Bayesian keyframe learning algorithm. Overall, this work presents a spatio-temporal embedding of video sequences to recognize the person in video. Encouraging results were reported on the XM2VTS dataset (Messer, Matas, Kittler, Luettin, & Maitre, 1999), achieving the best possible recognition accuracy of 99.3%.

Topkaya and Bayazit (2008) applied simple geometrical analysis using facial features and their corresponding positions on the face as criteria for selecting the subset of representative frames. Three sets of training data were created – the first with all detected faces considered, the second containing representative faces with aligned features and the third containing representative faces with additional preprocessing and masking steps. The results reported better success rates in the recognition task, testifying the importance of using only a subset of representative frames while eliminating other frames with poor face information.
The usage of all face images (in a video), including images of poor quality can possibly degrade the face recognition performance as realistic videos can often be subjected to uncontrolled conditions such as head pose and illumination variations, shadowing, motion blur and focus change over the course of the sequence. Wong et al. (2011) proposed an efficient patch-based probabilistic image quality assessment for the selection of faces in video-based face recognition. Their algorithm computes the similarity of a face image to a probabilistic face model, which represents an "ideal" face, by outputting a single score without resorting to fusion. This work improves upon similarly focused works (Nasrollahi & Moeslund, 2008; Rúa, Castro, & Mateo, 2008) that attempted to introduce complex fusion models for simultaneous detection of multiple quality characteristics followed by a combination of their respective scores. An extensive experiment based on their newly created surveillance dataset called ChokePoint demonstrates a considerable improvement in verification accuracy when using a subset of top quality images as ranked by its score.

In a largely experimental work by (Thomas, Bowyer, & Flynn, 2008), it is acknowledged that the quality of frames combined with the difference between frames in video improves the performance better than either property alone. Some interesting conclusions can be drawn here. This means that both the quality of selected representative face images, and also the variational property in sets of face images are both crucial factors that can improve the recognition ability of algorithms.

### 2.1.3 Image Set-based Approaches

Another group of approaches that use image sets as their choice of feature, true to its non-temporal nature, does not assume temporal coherency between consecutive images. Basically, the VFR problem is simply formulated as a task of matching a test image set against all the training image sets, each of which represents a single class. The advantage of making this assumption lies in the flexibility in applying these approaches to image sets acquired through various manners or sources—temporally ordered image samples collected over time (e.g. real-time video), multiple still shots collected arbitrarily (e.g. photo collections) or long-term discontinuous observations.
(e.g. recorded video footages from closed circuit television). Pragmatically, even in video data, it remains an uphill task to fully utilize temporal information or facial dynamics due to limitations afforded by the face acquisition process, such as the need to detect and/or track faces continuously for each frame, or the sparsity of unordered image sets collected from different instances of a lengthy video footage.

The mini review of literature in a few notable works (T. Kim, Kittler, & Cipolla, 2007; R. Wang et al., 2008) appeared in consensus over the categorization of image set-based (or image set classification) approaches. R. Wang et al. (2008) went further to the extent of coining this branch of visual recognition work as a Face Recognition based on Image Set (FRIS) problem. Generally speaking, there are two broad categories within the scope of image set-based VFR approaches: parametric model-based approaches and non-parametric sample-based approaches. Both categories will be discussed in the following sub-sections.

2.1.3 (a) Parametric Model-based Approaches

Parametric model-based approaches function based on the assumption that images from the set are part of some density distributions on the face manifold, which can be typically modelled using statistical learning methods. These approaches are often described in terms of a two-stage process: Representation of each image set by a parametric distribution, followed by computation of similarity (or dissimilarity) between the two distributions. In this review, approaches that directly utilize distribution models to characterise image sets while involving some degree of model parameter estimation, are regarded as parametric model-based approaches. By way of the categorization here, some parametric model-based approaches that are spatio-temporal in character will be discussed in Section 2.2 instead.

The task of recognizing faces from image sets can be cast as a statistical hypothesis testing problem. In one of the preliminary attempts, Shakhnarovich, Fisher, and Darrell (2002) proposed a straightforward method of matching the probability distributions between the test and training sets. Images in the image set are assumed to be...
independently and identically distributed (i.i.d.) samples of a probability density function. In this work, the face sample distribution is represented with a multivariate Gaussian function, while similarity between distributions are measured using the Kullback-Leibler (KL) divergence metric. In their experiments, face images with mostly frontal views were used, and their approach achieved an improved performance in recognition as compared to various image assembly methods and the mutual subspace method. A case of over-generalization may occur by assuming that face appearances within an image set are normally distributed, which may possibly be too simplistic for characterizing complex manifolds.

Following that, the extensive work by researchers in University of Cambridge in this particular focus area has resulted in further advancements in parametric distribution modelling of face image sets. One of their early works (Arandjelović & Cipolla, 2004) proposed to discard the use of KL-divergence as measure between distributions, and instead suggested the use of the resistor-average distance (RAD). The RAD between two distributions can be taken in a closed-form expression, enabling it to be computationally tractable while the dissimilarities between nonlinear face manifolds can be more accurately expressed. A high recognition accuracy of 98% was achieved on a relatively large database collected under various imaging conditions. This is a vast improvement over the methods that used the KL-divergence metric and also the mutual subspace method.

Arandjelović et al. (2005) also proposed an alternative approach to the modelling of distribution for the face appearances in image sets. To better deal with face appearance variations caused by illumination and pose variations, they adopted Gaussian Mixture Models (GMMs) to represent each image set, instead of single Gaussian functions as in their previous work (Arandjelović & Cipolla, 2004). Using KL-divergence as similarity measure between the modelled distributions, the proposed approach yielded an average recognition performance of 94% on the same dataset of 100 persons, an improvement over the method of Shakhnarovich et al. (2002) on single Gaussians. Later, they proposed a more sophisticated approach
Cipolla, 2009) which uses three Gaussian pose clusters decomposed from each face manifold, to characterize the different poses that constitute its face motion. The corresponding pose clusters of two face manifolds are compared, and the similarity between manifolds are measured by combining the pair-wise comparisons between pose clusters. Meanwhile, illumination invariant recognition is achieved by considering coarse region-based Gamma correlation with fine illumination manifold-based normalization. On a dataset containing 60 persons, this recent approach was able to consistently obtain an impressive average recognition rate of 95.2% in comparison to the earlier described KL-divergence method, mutual subspace method and simple majority vote method using Eigenfaces.

The work by Turaga, Veeraraghavan, and Chellappa (2008) described appropriate distance metrics on the Stiefel and Grassmann manifolds, which are claimed to be more natural in certain vision applications such as spatio-temporal modeling, invariant shape analysis and image matching. Video-based face data exhibits natural characteristics in these manifolds. Also, the geometric structure of manifolds are also reflected in the parametric and non-parametric density functions introduced. The Procrustus distance metric proved to be best choice in terms of reliability and computational efficiency, while matrix Langevin and matrix Bingham distributions provide the statistical modeling for the manifold density distributions.

In short, the main weakness of this particular group of approaches lies in the enforcement of strong assumptions about the actual distributions of the sample data. Parameter estimation is often an inaccurate task which may result in an undesirable generalization of the underlying distributions when the training sets and test sets are weakly correlated. It is common for test sets to contain a wide range of face poses, illumination and expression variations which are previously unseen in the training stage. The difficulty of parameter estimation can be further aggravated under limited training data where the characterization of face distributions fail to capture a holistic representation in the subspace.
2.1.3 (b) Non-parametric Subspace-based Approaches

In contrary to the earlier group of approaches, approaches that utilize subspaces of image sets can also be non-parametric in nature. This is a somewhat attractive solution to the cumbersome operation of selecting and estimating parameters to model distributions. Generally, the observation samples of image sets are not directly processed in its original space due to its high data dimensionality (say, a three-dimensional matrix of $M$-by-$N$-by-$T$ for an image set containing $T$ images of size $M$-by-$N$). Thus, it is natural to construct feature subspaces from image sets\(^2\). It turns out that this procedure is also reductionistic, though in a lesser extent compared to that of exemplar-based methods (in Section 2.1.1 where image set clusters are reduced to a mere single representative image). This representation also encodes more intra-class variability compared to exemplars.

In brief, non-parametric subspace-based approaches apply the concept of pairwise matching of subspaces derived from samples in the image sets. Typically, the distribution of a face image set can be neatly represented by a lower dimensional linear or nonlinear subspaces (which also includes higher-dimensional kernel subspaces). Two image sets can then be recognized as belonging to the same class if they produce the highest measure of similarity (or closest measure of distance).

The concept of canonical correlations or principal angles have gradually established itself as a primary technique used in image set matching, following some early works (Hotelling, 1936; Björck & Golub, 1973). Each image set is represented by a linear subspace and the principal angles between the two subspaces are used to compute the degree of similarity between the two image sets. Intuitively, principal angles reflect the common variations found in two subspaces. More formally, they are defined as the minimal angles between vectors of two subspaces. Canonical correlations, on the other hand, are cosines of the principal angles, geometrically regarded as a measure of similarity between subspaces. This method is well known for its ben-

\(^2\)Theoretically speaking, samples or single data points can also construct a subspace of itself (of a single sample), but the subspaces in context here are matrix subspaces representing a set of images.
efits over sample-based and parametric-based approaches in terms of computational
efficiency, robustness in representation, and accuracy.

The seminal work by Yamaguchi et al. (1998) on the adoption of principal angles for face recognition with image sequences was proposed with the name Mutual Subspace Method (MSM), which later became synonymous with image set matching. In MSM, the principal components (PCA) of the images are used to span the linear subspace for each set, while the smallest principal angle is used to compute the similarity between sets. The model shown in Figure 2.3 depicts the angle \( \theta \) between two subspaces, \( D \) and \( G \), whereby their respective vectors \( d \) and \( g \) defines the angle \( \theta \):

\[
\cos^2 \theta = \sup_{d \in D, g \in G, ||d|| \neq 0, ||g|| \neq 0} \frac{|(d, g)|^2}{||d||^2 ||g||^2}
\]

\[
= \lambda_{\text{max}}
\]

where \( \cos^2 \theta \) is also the largest eigenvalue \( \lambda_{\text{max}} \) of the matrix \( PQP \), \( P \) and \( Q \) are respectively the orthogonal projection matrices onto the subspaces \( D \) and \( G \).

Further on, Fukui and Yamaguchi (2003) extended the MSM to the Constrained Mutual Subspace Method (CMSM), which increases the consideration to cope with variations such as illumination and pose changes. CMSM projects the test and reference subspaces onto a constraint subspace, which allows smaller variance within each

Figure 2.3: Mutual subspace method (MSM), reproduced from (Yamaguchi et al., 1998)
subspace and better separation between the two subspaces. The success of this approach was demonstrated in a real-time implementation on LSI chips suitable for systems including home security, robot vision and mobile phones (Kozakaya & Nakaia, 2004).

In (Nishiyama, Yamaguchi, & Fukui, 2005), the authors took the idea further by introducing Multiple Constrained Mutual Subspace Method (MCMSM), which uses ensemble learning algorithms such as bagging and boosting to build multiple constraint subspaces. The recognition process is realized by combining the similarities obtained on every constraint subspace. In two separate experiments (one on a database of 50 subjects, another on a large database of 500 subjects), the results achieved a marked improvement over the MSM and CMSM methods. However, this improvement in accuracy is offset by the cumbersome, iterative procedure of generating the multiple constraint subspaces.

MSM-based methods are generally wrought with some obvious drawbacks in the aspect of information representation. Firstly, the appearance variations of face patterns encountered in a typical video is distributed nonlinearly. Hence, the linear modelling of these patterns using MSM is inadequate to fully represent a complex face manifold. Secondly, the fusion of information contained in different principal angles is performed rather ad-hoc-ly. There seemed to be no authoritative scheme or manner of utilizing the similarities between different modes of variation.

Further extended methods that utilize nonlinear subspaces or manifolds were proposed to increase the capability in generalizing highly nonlinear distribution of samples in attempt to stretch the upper bound of recognition accuracy. Wolf and Shashua (2003) adopted the "kernel trick" to compute principal angles between nonlinear manifolds. Similar to most methods that utilize kernels, finding the optimal function (one that represents the true distribution of the samples) is a difficult task. Also, a two- to three-fold increase in computational complexity is expected compared to conventional QR factorization based algorithm (with a single-pass singular value
decomposition (SVD)) since this approach requires a three-pass SVD.

The use of principal angles is further extended to nonlinear manifolds in a work by T.-K. Kim, Arandjelović, and Cipolla (2007). This is accomplished by integrating both local and global variations of the manifold. Firstly, locally linear manifold patches, which are extracted using mixtures of probabilistic PCA, are extracted to capture local variations. Then, a weighted average of the similarity between global modes of data variation and the best matching local behaviour is used as the measure of similarity between manifolds. For optimality in the fusion of principal angles, the AdaBoost algorithm is adopted to learn a set of weights corresponding to the principal angles. On a large database of 100 subjects, experiments conducted highlighted the superiority of this modelling approach over the classic MSM approach.

In attempt to increase the upper bounds of classification accuracy, discriminative learning methods were proposed. To classify image sets, T. Kim et al. (2007) proposed an objective method that maximizes the canonical correlations of within-class sets and minimizes canonical correlations of between-class sets. The maximization of that objective function can be done iteratively to obtain a transformation matrix that represents the new discriminative subspace. The transformed image sets are then compared and matched by means of canonical correlations. Experiments on various object achieved superior recognition performance over typical parametric and non-parametric approaches. In (Geng, Shan, & Hao, 2009), a generalized square loss based regularized LDA (SLR-LDA) is proposed to add discriminative capabilities to the classification of face image sets using canonical correlations. A more recent method proposed by Harandi et al. (2011) applied discriminative analysis on Grassmannian manifolds based on a graph embedding framework for an improved image set matching.

Another method (J. Li, Wang, & Tan, 2005) opted for the Earth Mover’s Distance (EMD) in place of typical principal angles or canonical correlations as a measure of similarity between two image sets or videos. The EMD is purportedly more robust in the way it handles incomplete set matching, allowing image sets of different sizes
to be compared. This adds the capability of allowing unwanted faces to be discarded from a set. Fan and Yeung (2006b) proposed a probability measure that considers both intrapersonal and extrapersonal subspaces with respect to a local manifold where the face images reside on. To determine if test face image lies on any local manifold, the measure computes the largest canonical angles between the (cluster mean-shifted) difference image and the intrapersonal and extrapersonal subspaces, before taking the normalized difference of the two angles. Experiments showed significant improvements over the principal angle and KL-divergence metrics.

The work by R. Wang et al. (2008) on Manifold-Manifold Distance (MMD), which was categorized earlier as an exemplar-based approach, can also be regarded as an image set-based approach by virtue of utilizing image sets to measure distance between manifolds. As the face manifold is modeled by a set of local linear subspaces called maximal linear patches (MLP), the dissimilarity measure between manifolds are computed in terms of differences with respect to variational modes (by mutual subspace angles) and appearance modes (by distance between exemplars). A good recognition accuracy of 93.6% on the CMU MoBo dataset and 96.9% on a subset of the first Honda/UCSD dataset were reported, highlighting the importance of both variational and appearance modes between image sets. However, it is important to note that this work roughly extracts the faces from video to form image sets without consideration for difficult faces that were undetected during the extraction process. The authors further extended this method to a discriminative embedding learning method called Manifold Discriminant Analysis (MDA) (R. Wang & Chen, 2009) seeks to maximize the "manifold margin" for better classification. Local linear models, or MLPs are constructed from image sets using hierarchical divisive clustering. Good results were reported in an extensive experiment conducted on several benchmark video face and image set databases. Despite the impressive results obtained by both MMD and MDA methods, their experiments seemed tailored to capture global changes across entire sequences (as image sets) while its effectiveness remain untested in shorter video segments where changes are more abrupt and localized.
Hadid and Pietikäinen (2009b) proposed a manifold learning approach that uses a novel manifold distance measure for video to video matching. Similar to the authors’ previous works (Hadid & Peitikäinen, 2004; Hadid & Pietikäinen, 2005), locally linear embedding (LLE) is performed separately for each training video that is person-specific (only one subject per video). In the subsequent recognition step, a test video sequence is then projected to the low-dimensional embedding of each training video manifold, resulting in a set of test video manifolds. The closest manifold, in terms of a manifold distance measure gives the subject identity in the sequence. An impressive recognition rate of 99.8% was obtained from experiments performed on the VidTIMIT dataset (Sanderson & Paliwal, 2002). The proposed method was also shown to be superior in comparison to PCA, LDA, LBP, HMM and ARMA techniques.

2.2 Temporally Motivated Approaches

The primary characteristic of the approaches described in the previous section is that observation samples are taken *sine tempore*, or without considering the temporal dimension. Successive or consecutive frames in the video sequences are continuous in the temporal dimension. The continuity arising from dense temporal sampling is based on the proposition that both the facial movements and change in appearance are continuous in nature (Zhou, 2004). Hence, video data allows for faces to be represented using these intrinsic continuities, an additional dimension of information not found in still images.

Psychological and neural studies (Knight & Johnston, 1997; O’Toole et al., 2002; Sinha et al., 2006) have shown the importance of facial dynamics in the human process of recognizing faces of people. Among the findings that garnered general consensus among psychologists and neuroscientists is the point that the human visual system uses both static and dynamic facial information to recognize the identity of persons observed. Also, the researchers in (O’Toole et al., 2002) stressed on the vital role of facial dynamics under degraded viewing conditions such as poor illumination or visibility, blurred vision, and recognition from a distance. While some of these problems have been investigated in face recognition literature, the notion of us-
ing temporal dimension have been proven to address these challenges in VFR. For instance, Arandjelović and Cipolla (2006) attempted to solve the challenging issue of illumination for videos by introducing a shape-illumination manifold that combines a photometric model and statistical appearance model. In a different aspect, Gorodnichy (2003) showed that the problem of recognition under poor resolution in videos can be countered using the dynamic information encoded along the temporal dimension, a concept primarily motivated by findings from neurobiology (Itti & Koch, 2001) and cognitive science (O’Toole et al., 2002). More recently, a concise survey by Hadid, Dugelay, and Pietikäinen (2011) discussed some recent advances while laying down some interesting challenges concerning the usage of dynamic features for various aspects of face biometrics such as gender recognition, age estimation, and ethnicity classification. This provides an evident case for the importance of facial dynamics, not just for face recognition in video, but in many different related biometric tasks.

### 2.2.1 State-space Approaches

One of the earliest works that attempted to improve recognition of faces by using temporal evidence over a sequence of images was reported in (Edwards, Taylor, & Cootes, 1999). In comparison with a single-image model, a statistical face model can be trained to accumulate identity evidence over a series of multiple images, with increasing robustness towards noise and effectiveness in classification. Through this work, the motivation of utilizing state-space models to facilitate temporal integration of observations was established, leading to a few landmark works thereafter.

One of the seminal works using state-space models (Zhou, Krüeger, and Chellappa (2002); Zhou and Chellappa (2002)) proposed a scheme that simultaneously keeps track of the evolving kinematics (motion) and identity of a face in video using a state-space model parameterized by a tracking state vector (continuous) and an identity variable (discrete). Each frame at time $t$ is considered as an observation $z_t$ and it is assumed that a transformed observation is a noise-corrupted gallery template:

$$
\tau_{\theta_t}(z_t) = I_{n_t} + v_t
$$

(2.3)
where $v_t$ is the observation noise, $I_{nt}$ a gallery template and $n_t$ the person’s identity. By assuming statistical independence on $p(\theta_0, z_0)$ and $p(n_0, z_0)$, temporal information can be modelled by estimating

$$p(x_t|x_{t-1}) = p(n_t|n_{t-1})p(\theta_t|\theta_{t-1})$$ (2.4)

Eventually, the joint posterior probability (across all faces in the sequence) $p(n_t|z_{0..t})$ is determined using Sequential Importance Sampling (SIS) algorithm, while marginalizing the posterior over the distribution of the identity variable will result in an estimation of the person’s identity. A notable remark should be made here – due to the nature of this model, both tracking and recognition of faces in video can be performed simultaneously. However, the primary drawback of this elegant framework lies in how transformation $\tau_{\theta_t}(z_t)$ from Equation (2.3) can avoid simplistic choices of affine transformations in order to generalize better under illumination and pose variations. On the CMU MoBo database (Gross & Shi, 2002), still-to-video experiments conducted only yielded a poor 56% accuracy rate for the top matching face identity and 88% when top three matches are considered. The authors proceeded to adopt an exemplar learning method (Krüeger & Zhou, 2002) that selects representative face images from the training videos to accomplish a still-to-video matching task. This is a generalization of a full video-to-video matching setting, and their evaluation on the CMU MoBo database produced some promising results.

Another state-space approach towards modelling face dynamics was presented in (Aggarwal et al., 2004). In this method, the authors presented a discrete state-space system with both linear observation and transition dynamics, posing the VFR task as a dynamical system identification problem. A linear dynamical system is learned for each moving face, regarding each frame as the output of the system. Concisely, the system can be represented by an autoregressive and moving average (ARMA) model:

$$x(t+1) = Ax(t) + v(t)$$ (2.5)
$$y(t) = I(t) = Cx(t) + w(t)$$ (2.6)
where $v(t) \sim \mathcal{N}(0, Q)$ and $w(t) \sim \mathcal{N}(0, R)$ indicating zero-mean additive Gaussian noise. The model parameters $A, C, Q$ and $R$ can be estimated in closed-form, while the dissimilarity between the gallery and probe (video) system models $(M_1, M_2)$ can be computed using a principal angle-based metric as follows,

$$d_M(M_1, M_2)^2 = \ln \prod_{i=1}^{n} \frac{1}{\cos^2 \theta_i} \tag{2.7}$$

where $\theta_i$ is the $i$-th principal angle from a total of $n$ principal angles. Their method achieved over 90% recognition rate on a dataset of 16 persons and the Honda/UCSD dataset (K. Lee et al., 2005). Although the results seemed promising in figures and given the extent of difficulty of the datasets, each probe video sequence was only tested once (entire sequence used) based on the reported experiments.

### 2.2.2 Appearance Manifold Approaches

A work by K.-C. Lee, Ho, Yang, and Kriegman (2003) offers an interesting insight into how a nonlinear appearance manifold of a face video can be learnt by approximating piecewise linear subspaces (computed by PCA). An exemplar-based approach was adopted to first sample representative frames from training videos. These exemplars are then partitioned into groups of similar face views or poses using $k$-means clustering. To accomplish recognition, the authors introduced a probabilistic maximum-a-posteriori formulation that combines the pose manifold likelihoods and the transition probability from the previous frame (to the current pose manifold). An image from a single frame $I$ is then classified to the class $k$ with the closest appearance manifold $M_k$:

$$k^* = \arg \min_k \ d(I, M_k) \tag{2.8}$$

where $d(I, M_k)$ is the $L^2$ distance between an image and a manifold in image space. An analytic solution for Equation (2.8) was presented through a probabilistic framework that integrates the distance between the image and all pose sub-manifolds $C^{ki}$, and also the pose manifold likelihoods $p(C^{ki}|I)$ (which are computed recursively using
transition probabilities):

\[ d(I, M_k) = \sum_{i=1}^{m} p(C^{ki} | I) d(I, C^{ki}) \]  

\[ (2.9) \]

The transition probability between pose sub-manifolds are learned from the training videos to capture the temporal dynamics between the sub-manifolds. Figure 2.4 illustrates how a complex and nonlinear manifold can be approximated as a union of several simpler pose manifolds.

The authors later proposed a strategy to perform simultaneous tracking and recognition (K. Lee et al., 2005) to good effect, achieving a recognition accuracy of almost 99% on the challenging Honda-UCSD dataset. While the reported results seemed impressive, all their results were reported on per-frame basis (percentage of correctly recognized frames) which does not meaningfully determine the person’s identity in a sequence. One major disadvantage of their scheme so far lies in the batch learning processing, which is infeasible when video sequences are lengthier. To mitigate this problem, an online learning version (K.-C. Lee & Kriegman, 2005) was proposed to construct a person-specific appearance manifold in an incremental manner, using consecutively-ordered frames obtained from the subject video with a generic initial prior.

Figure 2.4: A complex and nonlinear face appearance manifold can be approximated as a union of several simpler pose manifolds, reproduced from (K. Lee et al., 2005)
2.2.3 Other Approaches

There are a variety of other methods that were proposed to model facial temporal dynamics in videos. Works by Y. Li (2001); Y. Li et al. (2003) attempted this task through the construction of *identity surfaces*, or facial identity structures across views and time. This creates a "face trajectory" that models the spatio-temporal dynamics of moving faces, which in turn, facilitates recognition over time by means of measuring the trajectory distance. A best possible recognition rate of 93.9% was reported on a dataset of 12 subjects. Figure 2.5 shows the identity surfaces of two different subjects and their respective trajectories.

In another popular work (X. Liu & Chen, 2003), Hidden Markov Models (HMM) were adopted for video-based face recognition. Firstly, individual temporal dynamics are learned in the training stage by building a single HMM for each individual based on the PCA features of the images in sequence. Figure 2.6 shows the modeling of temporal dynamics in a face sequence using HMM in the training process. In the recognition stage, the test sequence is evaluated over time by determining the model that provides the highest likelihood values. In a practical evaluation setup where random sampling of test sequences were performed, the proposed adaptive HMM yielded a best possible error rate of 1.2% on the CMU MoBo dataset (Gross & Shi, 2002).

A few other works proposed some innovative schemes based on HMM. In a particular work by Mitra, Savvides, and Kumar (2006), feature representation based on

![Figure 2.5: Identity surfaces defining face patterns of two subjects, reproduced from (Y. Li et al., 2003)]
Figure 2.6: Temporal HMM modeling of a sample face sequence, reproduced from (X. Liu & Chen, 2003)

face asymmetry cues from the frequency domain are combined with an HMM feature set, producing an error rate of 3.3% on a dataset of 55 subjects with three emotions in different clips. A two-tier multi-dimensional generalization of the HMM is proposed in (Tistarelli, Bicego, & Grosso, 2009) to model dynamic information in human faces. In this hierarchical approach, the lower-level appearance-based spatial HMMs are used to model the emission distributions of the hidden states in the higher-level HMM where temporal continuity is represented. By considering a more discriminative representation of the sequences using Linear Discriminant Analysis (LDA), M. Kim et al. (2008) fused together pose-discriminant and person-discriminant features within a HMM framework to perform both tracking and recognition tasks. They reported a perfect recognition score of 100% on the Honda/UCSD (but based on the similar frame-by-frame recognition evaluation used in (K. Lee et al., 2005)) and over 70% on another database of YouTube videos consisting of 35 celebrities in 1,500 different videos. Nonetheless, it is well known that HMM techniques are prone to scalability issues, particularly considering that its training process is tedious and requires a large amount of data to attain a good generative model.

Alternatively, neural networks also offers capability in learning state transitions in addition to the estimation of subject identity over time. In a prime work along this direction, Gorodnichy (2005) came up with a biologically-motivated idea based
on neuro-associative principle for face recognition in video. The author’s model attempts to mimic the accumulation of visual stimuli (in this case, faces) over time, such that the associative recall of nametags (class labels) associated with the stimuli can be efficiently performed. Both memorization and recognition (which are, two distinct stages of the associative process) are formally modelled to capture the neuro-biological properties of the receptor-effector stimuli pair. Also, temporal dependencies (of the current frame from the previous frames) can be encoded by adding extra neurons to the network to serve as neural transmitters of outcome obtained from previous frames. Although the IIT-NRC database (Gorodnichy, 2005) used in this work had only 11 subjects with almost no lighting variations, this interesting approach managed a respectable recognition accuracy of more than 95% on a rather severe 160x120 video image resolution. The authors acknowledge the potential real-time and memory limitations of memorizing a larger number of individuals using the proposed model for selected applications. In another work, Barry and Granger (2007) fused together results from a fuzzy ARTMAP (predictive adaptive resonance theory) neural network used for recognition, and an array of Kalman filters used for motion tracking, into a temporal accumulator. At each specific time step, the responses from the neural network are used to update the accumulation variables to estimate a probable subject identity.

In brief, HMMs and neural networks are severely disadvantaged by the way they implicitly represent the temporally changing geometric properties of a face, causing difficulties in directly estimating face pose and motion states in new videos without performing tedious parameter estimations, special training and mechanisms in place. While the HMM and neural networks work primarily with global holistic facial features across frames in a sequence, local information is neglected. Local feature-based approaches are able to tackle pose variations by encouraging flexibility in geometric configuration between features, and improving the robustness towards alignment variations (Heisele, Ho, Wu, & Poggio, 2003).

With the recent growing interest in Local Binary Patterns (LBP) for face recog-
nition (Ahonen et al., 2006), a texture-based representation that utilizes local volumetric spatio-temporal features was proposed in (Hadid & Pietikäinen, 2009a) for video-based face and gender recognition. In this volume LBP (or VLBP), each face sequence is considered as a rectangular volume (in three dimensions) where LBP features are extracted from a three-dimensional neighborhood. The main rigidity of the VLBP is its simplistic consideration for only three frames at a time (or just two neighboring frames at each instance), which does not provide sufficient temporal information from the point neighborhood. The extended VLBP (or EVLBP) was proposed in the same paper to overcome this problem, whereby a temporal window is used to include different number of neighboring points from different frames (see Figure 2.7). Learning of the best set of EVLBP features was conducted using the Adaboost algorithm, and recognition is performed by nearest neighbor matching of vectors containing local histograms of EVLBP patterns. Evaluation on the Honda/UCSD and CMU MoBo datasets yielded a good recognition performance of 96.0% and 97.9% respectively.

In the last five years, the development of new feature representation paradigms have taken a major leap, signalling a distinct move away from conventional handcrafted features such as LBP (Ojala, Pietikainen, & Maenpaa, 2002), SIFT (Lowe, 2004) and HOG (Dalal & Triggs, 2005). The introduction of sparse representations (or sparse coding which is a more commonly used term today) (Wright, Yang, Ganesh, Sastry, & Ma, 2009), particularly for face recognition, deals with the large variations in data by harnessing its sparsity by learning sets of over-complete bases (Olshausen & Field, 1997) in an unsupervised fashion to represent data more efficiently. Sparse
coding expounds the theory that the choice of features are no longer critical if the sparsity in the recognition problem can be properly addressed. Deep learning methods or architectures (Hinton, Osindero, & Teh, 2006; Bengio, 2009) are also becoming increasingly popular in recent years (despite being fundamentally related to neural networks) as researchers seek to better model high-level abstractions in data. A majority of deep learning algorithms are strongly motivated by unsupervised learning of feature representations in each layer, whereby each layer forms a hierarchy from low-level to high-level features. Two recently published works (Taigman, Yang, Ranzato, & Wolf, 2014; Sun, Wang, & Tang, 2014) have pushed the state-of-the-art performance in large-scale face verification to the topmost limit (even surpassing human performance), achieving impressive results of more than 97% accuracy on the LFW dataset, which amounts to an error rate reduction of >27% compared to the previous best results. Up till date, deep learning methods have not been applied to the VFR task, thus providing a plausible direction for future work.

2.3 Summary

In the last decade, the fast-growing area of video-based face recognition (VFR) have seen a marked increase in popularity and interest, particularly from the computer vision and machine learning research community. This is primarily motivated by various techniques derived from image feature descriptors, manifold learning, statistical methods, and biologically-inspired algorithms.

In this chapter, a taxonomical review of VFR approaches is presented in a concise manner to provide a broad perspective that leads towards the contributions of this research. To conclude this review, there are a few essential observations that are worth highlighting:

1. Exemplar-based and image set-based features have unique and opposing strengths that are equally important: Exemplars provide a discriminative characterization of facial appearance modes, while image sets offer a generative characterization of facial variational modes.
2. A majority of methods in literature discard information from the temporal dimension in favour of purely spatial (image/video features) methods. The strengths exemplified by exemplar-based and image set-based methods (non-temporally motivated methods) can be better harnessed under a temporally-motivated scheme.

3. Accumulation of evidence in a video sequence is a key ingredient towards better classification. This is also supported from the psychological and neurological viewpoint.

4. Methods that are highly dependent on parametric estimation (GMMs, HMMs) are susceptible to scalability issues and over-generalization of data.

5. The inclusion of temporal information is not limited to only data representations or transition between samples/models. Other tasks in the recognition pipeline (such as classification or data clustering) can draw benefits from the utilization of temporal information as well. This is rarely explored in VFR literature.
Before proceeding to solve the underlying problems pinpointed earlier through the review literature and various motivations to this work, it is of utmost importance to lay the foundations required to ease the comprehension of this multi-problem work. Due to the extensiveness of this work which proposes various solutions within the framework, the common aspects shared ought to be described prior to the discussion of the problems solved.

In this chapter, an overview of the frameworks utilized in this work is first provided in Section 3.1. This includes a brief look into the novel cluster-centric framework for VFR, in which its mechanics are further discussed at length in Chapter 6. Section 3.2 defines the problem setting in the form of general mathematical notation used throughout this thesis in the proposed algorithms. In Section 3.3, all the datasets used in the experiments are elaborated in detail, including pre-processing settings and various assumptions made. Finally, Section 3.4 formally describes the augmented test set generation protocol established in this work, which lays out a systematic and balanced approach to experimental evaluation.

3.1 Framework

In this section, a high level overview of the two suggested frameworks for video-based face recognition – exemplar-based framework and cluster-centric framework, is provided. For each framework, the steps and processes involved are described, highlighting also the steps where novel approaches are proposed in this thesis. The integration of spatio-temporal characteristics or influences into various parts of the said frameworks is also indicated in line with the thematic emphasis of this thesis.
3.1.1 Exemplar-based Framework

The first of the two frameworks used in this work is the conventional exemplar-based framework, first coined by Krüeger and Zhou (2002), which had seen much widespread interest in recent years. In this framework, the exemplars or more specifically, the exemplar set, which is simply a collection of representative face images of all the subject classes, play the main role as the feature of attention (See, Eswaran, & Ahmad Fauzi, 2011). In the initial step, clustering is first performed on the face manifold of training videos to extract local clusters that define different groupings of face views or facial conditions. In typical cases, an exemplar is selected from each extracted cluster by taking the cluster mean or the face that is closest to the cluster mean. This constitutes the first part of characterizing a video-to-video classification matching (involving training data), where an entire training video is reduced or summarized to a single training image or a reduced set of training images.

The exemplar set is then projected onto a chosen feature space before the classification step performs matching between a new projected test image and the exemplar set images. Since each subject class may be represented by a few exemplars, all the matching or confidence scores can be aggregated by sum or product rule, or other suitable combination strategies (Kittler et al., 1998). However, this still depicts a conventional image-to-image matching between training and test data. To stay faithful to the complete characterization of a video-to-video matching, the second part (involving test data) requires combining individual image-to-image matchings using various voting strategies or Bayesian estimation methods, in order to arrive at a classification decision for the whole video sequence. The choice of method dictates the manner in which classification decision is determined – whether the frames are classified disjointedly before combining, or the frames are dynamically linked temporally during classification.

Figure 3.1 shows a block diagram of the exemplar-based framework for video-based face recognition. The red stars indicate steps within this framework where novel approaches are proposed and evaluated in this thesis. The feature representation and
clustering tasks will be discussed in detail in Chapters 4 and 5 respectively. Also, the use of spatio-temporal information can be leveraged at the clustering and classification steps. For the clustering task, a new spatio-temporal approach is proposed in Chapter 5, while the classification task can employ simple Bayesian estimation methods that function in a recursive, accummulative fashion.

3.1.2 Cluster-centric Framework

By building upon the exemplar-based framework discussed earlier, a cluster-centric framework for video-based face recognition, which utilizes features from both the point (image) and cluster (image set) levels is proposed. The term "cluster-centric" here highlights the source from which the rich set of features are extracted and used. Ultimately, the novelty of this framework lies in the proposal of a distinct, two-way expansion-contraction mechanism to VFR – firstly, the expansion by deconstruction of the video manifold into both types of features via local clusters derived from the manifold, and secondly, contraction by integration of similarity metrics based on the relationship between these features via probabilistic classification. This differs from most other approaches proposed by other researchers, that only rely on building a single generative or discriminative model based on a particular selected feature. Com-

![Exemplar-based framework for video-based face recognition](image)

Figure 3.1: Exemplar-based framework for video-based face recognition. The red stars indicate tasks where novel approaches are proposed in this thesis.
putationally, this would result in an obvious increase in cost, although there are elegant workarounds to the formulation (especially in the classification task discussed in Chapter 6) that will be able to mitigate this issue.

Figure 3.2 shows a block illustration of the novel cluster-centric framework for video-based face recognition. Similarly, the red stars indicate the steps within this framework where novel approaches are proposed and evaluated in this thesis. Taking the exemplar-based framework a step further, the spatio-temporal properties inherent in video sequences can be harnessed at different parts of the entire process. To exploit the spatio-temporal nature of video sequences, temporal cluster segments that correspond to each video frame is introduced. These test cluster segments are able to characterize a short temporal window or "snapshot" of adjacent frame variations for matching with the trained image set sub-manifolds. With these additional image set features, the classification step (Chapter 6) also requires a novel dual-feature approach while maintaining the ability to perform temporally accumulated classification of the face identity in video. In the earlier training phase, a novel spatio-temporal approach is proposed in the clustering step (Chapter 5) to enable more precise clustering of training data, similar to that described in the exemplar-based framework.

3.2 Problem Definition

Generally, a sequence of face images extracted from a video $V_c$ of subject $c$ is defined as an array of observed data in $\mathbb{R}^D$,

$$X_c = \{x_{c,1}, x_{c,2}, \ldots, x_{c,N_c}\}$$

(3.1)

where $x_{c,i}$ is the $i$-th image of the video sequence. $N_c$ is the number of face images in the video sequence, with the subject label of a $C$-class problem, $c \in \{1, 2, \ldots, C\}$. Each video is assumed to contain faces belonging to the same subject.

The default notation for a sequence of face images, $X_c$ indicates a training
Figure 3.2: Cluster-centric framework for video-based face recognition. Tasks involving both point and cluster features are shown on the left and right branches respectively. The red stars indicate tasks where novel approaches are proposed in this thesis.

video sequence. For each training video\(^1\), \(M\) number of clusters are extracted for each subject,

\[
\mathbf{Z_c} = \{z_{c,1}, z_{c,2}, \ldots, z_{c,M}\}, \text{ where } \quad (3.2)
\]

\[
z_{c,m} = \{x_{c,1}, x_{c,2}, \ldots, x_{c,N_m}\} \quad (3.3)
\]

\(^1\)Assumes one training video per subject. If more than one training video is used, image frames from all same-class videos can be aggregated sequentially before extracting clusters.
is the \( m \)-th cluster, a vector of \( N_m \) images. Subsequently, an exemplar image is selected from each cluster, resulting in an exemplar set for the \( c \)-th subject,

\[
E_c = \{ e_{c,1}, e_{c,2}, \ldots, e_{c,M} \} \tag{3.4}
\]

where \( e_{c,m} \subseteq z_{c,m} \) (indicating that an exemplar is selected from among the images in each cluster). Figure 3.3 illustrates how the training features are denoted. Hence, the resulting overall exemplar set consists of the extracted exemplars from all training videos (covering all \( C \) subjects), which can be succinctly summarized as

\[
E = \{ e_{c,m} | c = 1, \ldots, C; m = 1, \ldots, M \} \tag{3.5}
\]

For each test video sequence \( X'_k \) (extracted from test video \( V' \)) with \( N'_k \) number of frames, each face image \( x'_{k,i} \) (similar as notation in (3.1)) is a point feature of an unknown class \( k \) (unobserved data). Clusters are derived by partitioning the test video sequence into temporally ordered segments of length \( L \), whereby the test cluster image set

\[
\Theta_k = \{ \theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,M'} \} \text{, where } \theta_{k,m} = \{ x'_{k,1}, \ldots, x'_{k,L} \} \tag{3.6}
\]

\[
\Theta_k = \{ \theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,M'} \} \text{, where } \theta_{k,m} = \{ x'_{k,1}, \ldots, x'_{k,L} \} \tag{3.7}
\]

![Figure 3.3: Illustration of cluster-centric image set and exemplar features constructed from the training data](image-url)
is the \( m \)-th cluster segment consisting of \( L \) images each. However, it is remedial to note that if the number of images in the test sequence is indivisible by \( L \), then the last cluster is simply left with a segment length less than \( L \). Figure 3.4 illustrates how the test features are denoted.

At this juncture of the thesis, it is also necessary to note that in the classification matching task, the image sets at the cluster feature level (both \( Z \) and \( \Theta \)) are to be matched for each successive frame of the test video to accommodate requirements of the proposed classification approach (elaborated in Chapter 6). Thus in a straightforward manner, \( \theta_{k,j} \) is assigned as the cluster feature of the \( i \)-th instance of the test video sequence, if the frame belongs to the cluster itself, \( i.e. \ x'_{k,i} \in \theta_{k,j} \). In notation terms, the aggregated test cluster set is denoted as

\[
\Theta_k = \{ \theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,N_k'} \}
\]  

(3.8)

3.3 Datasets for Experiments

3.3.1 Datasets

In order to ensure that a comprehensive evaluation is conducted, three standard video face datasets were used: CMU MoBo (Gross & Shi, 2002), Honda/UCSD

Assume \( L=3 \)

\[
\begin{align*}
X'_{k} & x'_{k,1} \quad x'_{k,1+L} \quad x'_{k,1+2L} \quad \cdots \quad \cdots \quad x'_{k,N_k'} \\
\Theta_k & \quad \quad \quad \theta'_{k,2} \quad \theta'_{k,3} \quad \theta'_{k,4} \quad \cdots \quad \theta'_{k,M'} \\
\end{align*}
\]

\textbf{Figure 3.4: Illustration of cluster-centric image set and exemplar features constructed from the test data}
(K. Lee et al., 2005) and NICTA ChokePoint (Wong et al., 2011). This is by no means an exhaustive attempt to substantiate the proposed framework and its approaches, but to conduct experiments on a good selection of notable datasets that have been widely-accepted and well-received by the research community in this area of work.

**CMU Motion of Body (MoBo)**: The first dataset is one of the earliest and most commonly used dataset for VFR, adapted from its original use for human identification from distance. Its usefulness for VFR is warranted by a variety of facial appearances that can be captured from the video of a person walking on a treadmill in an indoor environment, while performing four different types of walking (slow walk, fast walk, inclined walk and slow walk holding a ball). A total of 96 sequences of 24 different subjects (each person has 4 videos) from the three most near-frontal views are used, with each video containing from 100 to 300 frames. Figure 3.5 shows some sample video frame shots from this dataset.

**Honda/UCSD**: This second dataset was collected specifically for unconstrained face recognition from video and it is one of the most challenging public datasets up to date. During the data collection, the subjects were asked to move their faces in different combinations of rotation, expression and speed. Only the first subset, which consists of 59 video sequences of 20 different persons, divided into a training subset (of 20 persons) and a test subset (of 19 persons) is considered in this work. Each video contains

![Figure 3.5: Sample video frame shots of three near-frontal views from the CMU MoBo dataset](image-url)
about 300-500 image frames, comprising of large pose and expression variations with significant 2-D (in-plane) and 3-D (out-of-plane) head rotations. Figure 3.6 shows some sample video frame shots from this dataset.

NICTA ChokePoint: The third dataset is recently collected for person identification under real-world surveillance conditions. An array of cameras was placed above several portals (natural "choke points" in terms of pedestrian traffic) to capture subjects walking through each portal in a natural way. The first subset (Portal 1), which consists of 600 sequences (one subject appearing in each sequence) of 25 different subjects extracted from 48 recorded videos, is considered in this work. Each subject has a variety of 24 different sequence instances, consisting of a combination of 4 sequence shots, 3 camera angles, and 2 movement modes (entering and leaving portal). The length of sequences ranged from 80 to 160 frames. The cropped face images provided with the dataset contain variations in terms of illumination conditions and pose, video quality, as well as misalignments due to the presence of slight occlusions. Only sample sequences taken from Camera 1 (single camera) are chosen in order to simplify the experimental setup for this dataset. Two different configurations are adopted in this work (for experiments in Chapter 6) based on how training-test samples are partitioned according to the movement modes: Leaving-Leaving (training and test videos from the same portal) and Entering-Leaving (training and test videos from different portals). Figure 3.7 shows some sample video frame shots from this dataset, with subjects captured entering and leaving different portals in the building.

![Figure 3.6: Sample video frame shots from the Honda/UCSD dataset](image-url)
3.3.2 Pre-processing Settings

For each video sequence (from both training and test data), faces are extracted\(^2\) using the Viola-Jones cascaded face detector (Viola & Jones, 2001), and then resampled to 32 × 32 pixel grayscale images. To cope with demanding head poses and rotations, the undetected faces are re-tracked and localized using the robust Incremental Visual Tracker (IVT) tracker (Ross, Lim, Lin, & Yang, 2008) to ensure that the experiments are thoroughly evaluated using all possible views in the sequences.

Following standard procedures used in literature, the detected face images are histogram-equalized to normalize illumination conditions and no further pre-processing tasks were applied prior to recognition. Though it may be possible to apply state-of-

\(^2\)The pre-process of detecting faces in video sequences is only performed on the CMU MoBo and Honda/UCSD datasets. The ChokePoint dataset readily provides the cropped face images.
the-art illumination normalization techniques such as one proposed by Tan and Triggs (2010) to aid recognition performance, the widely-accepted practice of normalizing pixel intensities by histogram equalization across all frames in a video sequence is applied. This is a common practice in many notable works in literature (Arandjelović et al., 2005; R. Wang et al., 2008; Harandi et al., 2011).

Figure 3.8 shows the detected, cropped and pre-processed frames of a sample video subsequence taken from the respective datasets used. In the experiment setup, a single video for each subject in the dataset is used for training, while the remaining videos are for testing. Experiments are repeated $k$-fold (or $k$ times) for $k$ number of videos per subject\(^3\). A closed world identification system is assumed here, where recognition or identification is performed to assign the face in a video sequence to a known subject class from the database.

### 3.4 Augmented Test Set Generation Protocol

In VFR literature, a majority of works do not enforce random sampling of sub-sequences from the test sequences provided in the datasets, only making use of entire video sequences of original length. Also, it is common for evaluation procedures to be not explicitly described, thus leading to much ambiguity in how experiments were conducted and the cogency of the reported performance of the evaluated approaches.

X. Liu and Chen (2003) proposed a test scheme that mimics a practical situation where any subject can come into the recognition system at any given time for any given duration, a subject video is randomly chosen (by sampling from a certain starting frame for a certain length) to form a subsequence for testing. Among the flurry of VFR works, this is the only work that suggested a practical scheme that can evaluate the robustness of recognizing video sequences. Motivated by this random sampling procedure, an extensive evaluation protocol is formulated to generate an augmented test set to introduce a variety of sequences with different starting frame positions and

---

\(^3\)Cross validation is implemented for CMU MoBo & ChokePoint only. Honda/UCSD has uneven training-test sample composition and a one-pass evaluation is preferred for simplicity.
sequence lengths. To accomplish this systematically without introducing additional biases, the augmented test set is constructed by randomly sampling $W$ subsequences of $T$ different lengths from the original test videos belonging to each subject class, for all $C$ subject classes. Hence, a total of $W \times T \times C$ subsequences in the augmented test set are sampled for experiments. The illustration in Figure 3.9 shows the generation of the augmented test set from the original test video sequences, based on the protocol parameters $\{W, T, C\} = \{3, 3, 20\}$.

In the following chapters, $L = \{\cdot\}$ is used to denote a set of video subsequence

---

**Figure 3.8:** Sequentially ordered image frames of a sample video subsequence from the evaluated datasets. These face images were detected, cropped and pre-processed according to the described procedure.
Protocol parameters $\{W, T, C\} = \{3, 3, 20\}$

Augmented test set with subsequences $X'_{W,T,C}$

Figure 3.9: Illustration of how the augmented test set is generated from the original video sequences. The example above is based on the protocol parameters $\{W, T, C\} = \{3, 3, 20\}$. Due to limitation in display, note that all $W$ subsequences of a certain same length and class are randomly sampled from the original test video, and thus are not the same.

lengths, where its cardinality $|L| = T$. A rule of thumb to follow when selecting the predetermined set of subsequence lengths is by selecting $T$ subsequence lengths spread out at equal intervals, from around the shortest possible length to the longest possible length. To formalize this for the experiment sections later, datasets used are denoted by $\text{Dataset}_{W,T,C}$.

This evaluation scheme is designed to reduce video length and subject coverage biases and to better mimic realistic scenarios with arbitrary sets of views. It is known that longer sequences are able to produce better recognition accuracy, especially with spatio-temporal approaches (Hadid & Pietikäinen, 2005). By using video sequences of different lengths, an impartial evaluation environment is enabled to test for the robustness of recognizing faces in video. To ensure all subjects are tested, the evaluation scheme runs the random sampling for every subject in the test set, generating $W \times T$
number of subsequences for each subject. On the contrary, a fully random sampling may result in unevenly skewed data, which may not uniformly evaluate all possible subjects in the test set.

3.5 Experimental Methodology

In the subsequent chapters, the proposed novel algorithms for various tasks in a VFR pipeline will be discussed at length, along with experiments for evaluation. Here, the organisation of experiments to test the proposed algorithms are briefly explained according to an *incremental experimentation* methodology.

This methodology carefully evaluates each proposed algorithm (for each task) in the pipeline in a gradual, incremental manner while fixing the remaining tasks with existing methods. Its "incremental" nature is depicted in Figure 3.10 whereby the best method found by experiments for one task is chosen and fixed in experiments for the subsequent tasks. For instance, some existing methods for clustering (HAC) and classification (Naive Bayes or NB) are first used in experiments for Task I. The novel method (that is also the best method from previous experiments), NDMP is the best choice for feature extraction and it is then used in experiments for Task II while classification remains with the NB method. Next, the novel clustering method, STHAC is then selected for use in Task III experiments, together with the earlier selected NDMP. Finally, the last task confirms the best possible results using all three novel proposed methods. It is important to note that this is not an exhaustive attempt to find the best possible combination of methods. There is possibility that the recognition performance could improve further if other untested methods are applied. But considering the slew of methods found in the literature and all possible combination of methods, it may be a highly improbable experiment to attempt. The experiments for all three tasks will be presented in the same order as shown in Figure 3.10 following their respective chapters (Chapters 4, 5 and 6).

Due to the tediousness of going through all possible combinations of different methods, including all the free parameters used in each method, it would be an
immense undertaking to organize and analyze the outcome of an exhaustive set of experiments. Instead, this experimental methodology provides a straightforward and intuitive scheme for structuring experiments, a process of competition and elimination akin to "survival of the fittest" concept. This is neither optimal nor rigorous, but it provides a tailored solution for each task while obtaining the best combination of methods that achieves the best possible recognition accuracy.
CHAPTER 4

LEARNING NEIGHBORHOOD DISCRIMINATIVE MANIFOLDS IN FACE VIDEOS

In recent years, manifold learning has become an increasingly growing area of research in computer vision and pattern recognition. With the rapid development in imaging technology today, it plays an important role in many applications such as human activity analysis, multimodal biometrics and also in video-based face recognition, where the abundance of data often demands better representation.

Typically, an image can be represented as a point in a high-dimensional image space. However, it is common assumption that the perceptually meaningful structure of the data lies on or near a low-dimensional manifold space (Brand, 2003). The mapping between high- and low-dimensional spaces is accomplished through dimensionality reduction. This remains a challenging problem for face data in video, where large complex variations between face images can be better represented by extracting good features in the low-dimensional space.

In this chapter, a novel supervised manifold learning method called Neighborhood Discriminative Manifold Projection (NDMP) is proposed for feature extraction in video-based face recognition. NDMP builds a discriminative eigenspace projection of the video face manifold based on the intrinsic geometry of both intra-class and inter-class neighborhoods. The concept of preserving the neighborhood structure of each data point is inspired by a well known algorithm called the Locally Linear Embedding (LLE) proposed by Roweis and Saul (2000). In addition to that, neighborhood structures of points with class labels can be leveraged to produce discriminative capabilities.

For each training video, a set of face representative exemplars are automati-
cally constructed through clustering by selecting the face that is closest to the mean of each cluster. With these exemplars, an optimal low-dimensional projection is learned by solving a constrained least-squares (quadratic) objective function using local neighborhood and global structural constraints. A compact generalized eigenvalue problem is formulated, where new face data can be linearly projected to the feature space. Finally, the test video sequences are classified using a probabilistic-based Bayes classifier. Benchmark face video databases were used to evaluate the performance of various extracted features within an exemplar-based recognition setup.

For the sake of clarity, it is essential to deal with the aspect of feature extraction first in order to facilitate the presentation of experiments, although the feature extraction step clearly lies second in order within the recognition framework. This would enable the choice of features to be presumed ahead of discussions on other contributions in the clustering and classification steps.

The organization of the chapter is as follows: Section 4.1 provides a brief survey of related literature on feature extraction approaches, and the motivations that led to the idea. In Section 4.2, the proposed NDMP method and its formulation is fully elaborated in detail. Section 4.3 briefly describes the framework setup of other tasks involved in the recognition process and their respective selected methods. Section 4.4 reports and analyzes the experimental results obtained through comparative evaluation and rank-based identification. Finally, 4.5 summarizes the chapter.

4.1 Background Work

4.1.1 Related Literature

In general, the problem of dimensionality reduction\(^1\) can be summarized as the extraction of low dimensional structure from high dimensional data. The "curse

\(^1\)It is common to refer to *dimensionality reduction* using other terminologies such as *manifold learning* or *spectral projection* or *embedding*. These terms are widely used interchangeably throughout literature to mean the same thing, although they can also be differentiated distinctly in terms of context and usage.
of dimensionality" arises when data in high-dimensional space becomes too sparse that it contains irrelevant dimensions resulting in poorly organized data. Generally, the problem can be described by means of a mathematical framework. Given a high dimensional data matrix \( X = (x_1, \ldots, x_n) \) of input patterns where \( x_i \in \mathbb{R}^D \), how can we produce a "faithful" low dimensional representation of the original data set by computing its corresponding output patterns \( y_i \in \mathbb{R}^d \) such that \( Y : \mathbb{R}^D \to \mathbb{R}^d \) for \( d \ll D \)? "Faithful" representation here is loosely defined as a transformation or mapping that minimizes the reconstruction error of the output patterns \( i.e. \) nearby inputs are mapped to nearby outputs, while faraway inputs are mapped to faraway outputs (Saul, Weinberger, Ham, Sha, & Lee, 2006).

Classical linear dimensionality reduction methods such as Principal Component Analysis (PCA) (Jolliffe, 1986; Turk & Pentland, 1991), Multidimensional Scaling (MDS) (Cox & Cox, 2001) and Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997) are among the most popular techniques in pattern recognition literature and they have been applied in various applications to a good degree of success. These linear methods are clearly effective in learning data in simple Euclidean structure. Typically, a linear subspace projection

\[
Y = A^T X
\]

maps the original data points \( X \) to the projected coordinates in embedded space \( Y \) by a linear transformation matrix \( A = \{a_i \in \mathbb{R}^{D \times d} | i = 1, \ldots, d \} \).

Without going into great detail, the most commonly known definition of PCA due to Hotelling (1933) postulated the learning of a projection (with \( d \) principal axes) that maximizes its variance,

\[
\max_{\|a\|=1} \text{Var} \{a^T X\} = \max_{\|a\|=1} a^T \Sigma a
\]

where \( \Sigma \) denotes the sample covariance of \( X \) or \( \Sigma = XX^T \in \mathbb{R}^{D \times D} \). PCA is arguably the most well-known algorithm used for face recognition, and it is sometimes known

MDS (Cox & Cox, 2001) computes a low dimensional representation of a high dimensional data set that most faithfully preserves pairwise distances between data points in the new embedded space. The MDS is equivalent to PCA (with the same optimal solution possible for both) in Euclidean space with the exception that optimizing its minimization error can be achieved by spectral decomposition of the Gram (inner product) matrix \( G = X^T X \in \mathbb{R}^{n \times n} \). The Gram matrix can be easily derived from the squared pairwise distances in input space, \( D = \| x_i - x_j \|^2 \), which leads to the flexibility of applying MDS whenever a distance metric exists between data points (points are not required to be vectors).

With additional class information, LDA (Belhumeur et al., 1997) learns a linear projection that maximizes the ratio of the between-class scatter to the within-class scatter,

\[
\max \frac{a^T S_B a}{a^T S_W a} \equiv \max \frac{\text{tr}(A^T S_B A)}{\text{tr}(A^T S_W A)}
\]

where \( S_B \) and \( S_W \) denote the between-class scatter and within-class scatter matrices respectively. This relates similarly to the Fisher’s criterion (or generalized Rayleigh quotient) which is maximized in the original Fisher’s linear discriminant\(^2\) (Fisher, 1936). LDA is regarded as a supervised dimensionality reduction method in that it works to increase the separation between classes after projection to the low-dimensional embedding by using class label information. It can be solved in the form of a generalized eigenvalue problem,

\[
S_B A = \lambda S_W A
\]

which allows us to find the eigenvectors that correspond to the projection \( A \).

The biggest drawback of these methods is that they fail to discover the intrinsic dimensionality—or in other words, the number of underlying modes of variability, of

\(^2\)LDA or Fisher’s linear discriminant or Fisherfaces are often used interchangeably. Also, the original work by Fisher (1936) does not assume classes to be normally distributed in its formulation, unlike the LDA.
the image space due to its assumption of linear manifolds. If the data lies on (or near) a low dimensional submanifold, then its structure may possibly be highly nonlinear; in such cases, linear methods may fail. Also, overfitting remains a common problem in these methods due to the high number of feature variables needed to produce a faithful output representation.

The emergence of manifold learning methods, ushered in by two prominent techniques – Locally Linear Embedding (LLE) (Roweis & Saul, 2000) and Isomap (Tenenbaum et al., 2000), brought a fresh perspective towards the learning of nonlinear manifolds. These manifold learning methods, also known as graph-based spectral methods (Saul et al., 2006), are able to discover the underlying high-dimensional nonlinear structure of the manifold in a lower dimensional space. These methods start by building a sparse graph in which the nodes represent input patterns and the edges represent neighborhood relations. The graph (represented in matrix form) can be regarded as a discretized approximation of the submanifold of the input patterns. Finally, spectral decomposition is performed on the graph to reveal the low dimensional structure of the submanifold.

LLE seeks to learn the global structure of a nonlinear manifold through a linear reconstruction that most faithfully preserves the linear structure of the local neighborhood. The algorithm obtained the embedded outputs from the bottom of the eigenvectors of a sparse matrix. The local neighborhood graph is constructed by computing the k-nearest neighbors of each input point while weights are assigned to the edges of the graph. The intuition of LLE lies in the assumption of small linear "patches" on a low dimensional submanifold represented by each input point and its neighboring points. Several works have demonstrated the vast potential of LLE for various classification tasks – face identification (Pang & Kasabov, 2006), facial expression recognition (Liang, Yang, Zheng, & Chang, 2005), handwritten digit classification (de Ridder & Duin, 2002). In all these works, the nonlinear LLE is found to outperform the linear PCA when it is used to learn the feature embedding of nonlinear data for classification tasks. The LLE method was later improved by de Ridder and Duin (2002) using class
labels of the training samples to determine neighborhood selection. This supervised
version of LLE named Supervised LLE (SLLE), uses a straightforward formulation to
add distance between samples of different classes. This creates "disconnected" classes,
each of which will be mapped by conventional LLE.

Isomap is similar to the LLE in spirit, except that the low dimensional em-
bedding is determined by computing the projection that most faithfully preserves the
pairwise distances between input patterns as measured along the submanifold. As
such, geodesic distances along the submanifold between points are used instead of
standard Euclidean distances (which are from the MDS method). A highly tractable
solution to estimate geodesic distances is realized by using Dijkstra’s algorithm (with
polynomial-time complexity $O(n^2 \log n + n^2 k)$) to compute the pairwise distances be-
tween all nodes along shortest paths through the graph.

Another method, Laplacian eigenmaps (Belkin & Niyogi, 2003) finds a low di-
mensional embedding that most faithfully preserves proximity relations, where nearby
input patterns are mapped to nearby output patterns based on the nearness measured by
a weight matrix. Positive weights are assigned to the edges of the graph, whereby con-
stant or exponentially decaying ("heat kernel") functions are some suggested choices
for the weight function. For video-based face recognition, several works reported good
recognition rates by using LLE (Hadid & Peitikäinen, 2004; Fan et al., 2005) and
Isomap (albeit only the geodesic distances) (Fan & Yeung, 2006a) to build a view-
based low-dimensional embedding for exemplar selection, but they stopped short of
extending it for feature representation.

The main disadvantage of these methods is that they cannot deal with the out-
of-sample problem – where new data points cannot be projected onto the embedded
space. This limits their potential usage for classification and recognition tasks. How-
ever, some works attempt to address this problem by performing a direct mapping of
new input samples to the projected output space. New unseen samples can be mapped
directly by first finding its $k$ nearest neighbors in $X$, calculating new reconstruction
weights and then linearly mapped it to the new embedding $Y$. de Ridder and Duin (2002) observed that LLE only performs better than PCA when mapping down to very small number of dimensions (whereby classifier performance is low) as to avoid overfitting data. Another critical drawback lies in its higher computational cost of mapping new samples, $O(nd)$ as compared to $O(d)$ for PCA and other linear projection techniques. In another work by Bengio et al. (2004), an out-of-sample extension for various embedding algorithms is proposed to provide an explicit mapping for new samples from the input space to the embedding space without needing to recompute eigenvectors. A unified framework is introduced in which the extended algorithms are seen as learning eigenfunctions of a kernel.

More recently, various manifold learning methods such as Locality Preserving Projections (LPP) (He, Yan, Hu, Niyogi, & Zhang, 2005), Orthogonal Locality Preserving Projections (OLPP) (Cai, He, Han, & Zhang, 2006) and Neighborhood Preserving Embedding (NPE) (He, Cai, Yan, & Zhang, 2005) begin to surface to solve the out-of-sample problem. These methods resolve this limitation by deriving optimal linear approximations to the embedding that can be solved in closed form as generalized eigenvalue problems, while attempting to maintain its original local neighborhood structure using neighborhood graphs. He, Yan, et al. (2005) extended Laplacian eigenmaps by introducing a linear projection mapping called Locality Preserving Projections (LPP) or Laplacianfaces, which seeks an embedding that best preserves local proximity information to represent a face subspace in a low dimensional submanifold. The OLPP method (Cai et al., 2006) is based on the LPP algorithm, but produces orthogonal basis functions which allows data to be easily reconstructed while possessing more locality preserving power as well. On the other hand, NPE (He, Cai, et al., 2005) is directly formulated based on LLE’s objective functions to alleviate the out-of-sample problem. A linear subspace projection is presented by optimizing the similar objective function of LLE, thus inheriting LLE’s neighborhood preserving property. However, NPE’s lack of discriminative ability does not effectively distinguish within-class and between-class neighborhood structures for consideration in the mapping.
Some supervised methods directly incorporate additional discriminative information to the embedding. The Local Discriminant Embedding (LDE) (H. Chen, Chang, & Liu, 2005) and Marginal Fisher Analysis (MFA) (Yan et al., 2007) both utilize affinity weight matrices, which are constructed based on adjacency graphs (using simple binary or heat kernel functions). However, their discriminative ability is only limited to neighborhood point relationships by graph embedding. It does not distinguish between neighborhood structures created by intra-class and inter-class groups of points within the same manifold, which can be very disparate in nature. Along the same vein, a newer work by X. Li, Lin, Yan, and Xu (2008) proposed to use high-order tensors to enhance discriminant locally linear embedding through a graph-embedding framework. This "tensorization" of the discriminative LLE is reportedly more effective, and a high accuracy was obtained for the task of human gait recognition.

An interesting unified algorithm based on LLE and LDA (ULLELDA) was presented by J. Zhang, Shen, and Zhou (2004) to achieve effective face recognition by exploiting the strengths from both approaches. First, training samples are mapped into low-dimensional embedding space and then LDA algorithm is used to project samples onto discriminant space to enlarge between-class distances and decrease within-class distances between samples. Then, new unknown samples can be directly mapped into discriminant space (via LDA projection) without recomputing the embedding. Although the two tasks can be unified into a single step, experimental results did not show much improvement over different combinations of the PCA, LDA and LLE methods as structural information may be possibly lost when the weight matrix from LLE is projected onto LDA discriminant space. Moreover, for each new input sample, the nearest neighbors and their respective weight coefficients have to be determined followed by mapping to the low dimensional space, an expensive series of computations.

4.1.2 Motivations

The rich literature of manifold learning methods offers many potential avenues for improving existing shortcomings. There are two main motivations to this work:
1. While recent works (He, Yan et al., 2005; He, Cai et al., 2005; H. Chen et al., 2005; Yan et al., 2007) have addressed the out-of-sample problem by deriving optimal linear embeddings and incorporating additional discriminative power, there is little attention paid to the potential of optimizing and discriminating the manifold structure. Graph-based neighborhoods contain local patches of points that are used to build an optimal reconstruction of the input patterns in the new embedding space. Without using class label information, the true structure of each class patch or submanifold is not well-embodied (as in the case of simply selecting nearby points to represent neighborhood relations). Thus, the aforementioned supervised approaches do not discriminate between within-class and between-class neighborhood structures while keeping the global manifold fitted. Earlier, de Ridder and Duin (2002) suggested a simple procedure to rehash the original LLE to accommodate class labels, where the neighbors of each point can be selected from nearby points that belong to the same class. The proposed NDMP method proceeds along the same spirit to form the local neighborhoods. Although the proposed method bears some similarities to the NPE in terms of its theoretical foundation (based on local linear reconstructions of LLE), the final objective function is essentially different.

2. While many of the latest manifold learning techniques reported very good performance results, they were evaluated on still image data sets with limited variations in face pose, expression and illumination such as ORL, CMU PIE, Yale and UMIST data sets (Gross, 2004). Moreover, these data sets are restricted to only a limited variety of face images. On the contrary, video sequences consist of a series of frames with possibly a much greater amount of face variations. This may prove to be a challenging task owing to the nonlinearity of its data manifold. As such, the formulation of a method that is able to both uncover the underlying nonlinear manifold structures and discriminate between those belonging to different classes is essential for effective recognition of faces in video sequences.

The reasons above provide the necessary impetus for a novel formulation of the new Neighborhood Discriminative Manifold Projection (NDMP) method to ex-
tract meaningful features for video-based face recognition. In short, the NDMP seeks to learn a low dimensional embedding from high dimensional input patterns that most faithfully preserves the global manifold structure while maximizing the class-wise discrimination between local neighborhood structures.

4.2 Locally Linear Embedding (LLE)

The LLE algorithm can be succinctly described in three main steps. Assuming the data lies on a nonlinear manifold which locally can be approximated linearly, and $X = \{x_i \in \mathbb{R}^D | i = 1, \ldots, N\}$ represent the input data in Euclidean space consisting of $N$ face samples in $D$ dimension, belonging to one of $C$ classes $\{c | c \in \{1, \ldots, C\}\}$, the procedures are as follows:

**Step 1**: For each point $x_i$, find its $K$ nearest neighbors.

**Step 2**: Compute the reconstruction weights $W$ that best reconstruct each point $x_i$ from its neighbors $x_j$ by minimizing the cost function,

$$
\varepsilon_{\text{rec}}(W) = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{K} W_{ij} x_j \right\|^2
$$

where $W_{ij}$ is the reconstruction weight vector, subject to constraints $\sum_{j=1}^{K} W_{ij} = 1$ and $W_{ij} = 0$ if $x_i$ and $x_j$ are not neighbors.

**Step 3**: Compute the embedding coordinates $y_i$ best reconstructed by the optimal reconstruction weight $W$ by minimizing the cost function

$$
\varepsilon_{\text{emb}}(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{K} W_{ij} y_j \right\|^2
$$

subject to constraints $\sum_{i=1}^{N} y_i = 0$ and $\sum_{i=1}^{N} y_i y_i^T / N = I$ where $I$ is an identity matrix. The new coordinates in embedded space, $Y = \{y_i \in \mathbb{R}^d | i = 1, \ldots, N\}$ is a $d \times N$ matrix, where the dimensionality of the new embedded points, $d < D$. 

75
In Step 2, the optimal reconstruction weights can be computed in closed form (Saul & Roweis, 2003), whereby the cost function,

$$
\varepsilon_{rec}(W) = \sum_{j=1}^{K} \sum_{m=1}^{K} w_j w_m G_{jm}
$$

(4.7)

where $G_{jm} = (x - x_j) \cdot (x - x_m)$ is the "local" Gram matrix. From that, the reconstruction error can be easily solved as a constrained least squares problem, with a regularization term added to the reconstruction cost for cases that do not have a unique solution (i.e. if Gram matrix is singular or near singular, or $K < D$). This is the least expensive step in the LLE algorithm, with time complexity scaling as $\mathcal{O}(DNK^3)$ for each data point.

The optimal embedding is found by minimizing Equation (4.6) in Step 3 subject to the constraints specified. Equation (4.6) can be rewritten in quadratic form,

$$
\varepsilon_{emb}(Y) = \sum_{i=1}^{N} \sum_{j=1}^{K} M_{ij} (Y_i \cdot Y_j) = tr(YMY^T)
$$

(4.8)

where the square $N \times N$ matrix $M = (I - W)^T (I - W)$ is sparse, symmetric, and semi-positive definite. By constrained minimization of the trace term from Equation (4.8), the generalized eigenvector problem can be solved to obtain the optimal embedding $Y$ from the smallest eigenvectors of matrix $M$. More precisely, the bottom $(d + 1)$ smallest eigenvectors of the matrix $M$ are taken as the optimal linear projection matrix with the smallest eigenvector (corresponding to the mean of $Y$) discarded to enforce the translational constraint $\sum_{i=1}^{N} y_i = 0$. This final step of LLE is the most computationally expensive, scaling as $\mathcal{O}(dN^2)$ without further optimization.

LLE is elegant in its simplicity of both theory and implementation. While the LLE is disadvantaged by the lack of a "telltale gap" that indicates the actual intrinsic dimensionality of the underlying manifold (compared to the Isomap algorithm) (Saul et al., 2006), the resulting embedding radically unravels a highly nonlinear set of points
by projecting them as linearly as possible. Figures 4.1(a), 4.1(b), 4.1(c) and 4.1(d) compare the embedding of image frame data from a single sample video in PCA, MDS, Isomap and LLE spaces respectively.

### 4.3 Neighborhood Discriminative Manifold Projection (NDMP)

In this section, the detailed formulation of the proposed NDMP method is elaborated, in three parts. First of all, the intra-class and inter-class neighborhoods are extracted for each point. Then, two sets of reconstruction weights – one each for the neighborhood type, are computed. In the final step, the optimal linear projection is learned to enable the data points to be embedded in the low-dimensional NDMP feature space. In an exemplar-based VFR scheme, "points" or "data points" refers to the extracted exemplars from the training exemplar set, throughout the explanation of the algorithm in this section. Further details on the VFR setup and the other tasks involved can be found in Section 4.4.

#### 4.3.1 Intra-class and Inter-class Neighborhoods

For each point $x_i$, let all other points in a local neighborhood $\Psi$ comprise of two disjoined neighborhood subsets – *intra-class* subset $\{\psi_{ip}|p = 1, \ldots, K\}$ and *inter-class* subset $\{\psi_{iq}|q = 1, \ldots, K'\}$. Concretely, the $K$ nearest *intra-class* neighbors and $K'$ nearest *inter-class* neighbors are computed for each point $x_i$. A point $x_{c',j}$ is an *intra-class* neighbor of point $x_{c,i}$ if they belong to the same class ($c = c'$). Similarly, point $x_{c',j}$ is an *inter-class* neighbor of point $x_{c,i}$ if they do not belong to the same class ($c \neq c'$).

Generally speaking, it is possible to specify a local neighborhood to have an unequal number of intra-class and inter-class nearest neighbors. However, for the sake of uniformity of weight matrices, $K = K'$ is assumed. Due to class set limitation, the number of intra-class and inter-class neighbors is restricted to a maximum of $(M - 1)$, where $M$ is the number of training exemplars (points) in each class.
4.3.2 Neighborhood Reconstruction Weights

Based on Equation (4.5), two reconstruction cost functions are formulated to obtain neighborhood reconstruction weights that best reconstruct each point $x_i$ with respect to its type of neighbors. Both intra-class reconstruction weight matrix $W^r$ and inter-class reconstruction weight matrix $W^e$ can be computed by minimizing the respective cost functions:

$$
\varepsilon_{rec}(W^r) = \sum_{i=1}^{N} \left\| x_{c,i} - \sum_{j=1}^{K} W^r_{i,j} x_{c,j'} \right\|^2
$$

(4.9)

Figure 4.1: Spectral plots of data points from a sample video sequence in different embedding spaces, described by their respective first three coordinates.
Both Equations (4.9) and (4.10) can be computed in closed form by determining the optimal weights by solving a constrained least squares problem through the local Gram matrix (refer to Equation (4.7)). Alternatively, it can also be solved using a linear system of equations with regularization.

Quite clearly, the first two steps of the proposed algorithm alludes to the first two steps of LLE algorithm, with the major difference being the usage of two neighborhoods (intra-class and inter-class), which in turn result in two sets of reconstruction weights.

### 4.3.3 Optimal Projection

In this step, an optimal projection is learned to embed the data points to the NDMP feature space. Typically, a linear subspace projection \( Y = A^T X \) maps the original data points \( X \) in input space to the projected coordinates in embedded space \( Y \) by a linear transformation matrix \( A = \{ a_i \in \mathbb{R}^{D \times d} | i = 1, \ldots, d \} \).

Similar to NPE algorithm (He, Cai, et al., 2005), the *intra-class* cost function can be formulated by expanding the least-squares term. By some algebraic steps, it
can be determined that

$$\varepsilon_{\text{emb}}(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{K} W_{ij}^r y_j \right\|^2$$

$$= \left\| Y - \sum_{i=1}^{N} \sum_{j=1}^{K} W_{ij}^r y_j \right\|^2$$

$$= tr \left\{ \left( Y - \sum_{i=1}^{N} \sum_{j=1}^{K} W_{ij}^r y_j \right) \left( Y - \sum_{i=1}^{N} \sum_{j=1}^{K} W_{ij}^r y_j \right)^T \right\}$$

$$= tr \left( Y \left[ \delta_{ij} - 2W^r + \|W^r\|^2 \right] Y^T \right)$$

$$= tr \left( Y(I - W^r)^T (I - W^r) Y^T \right)$$

$$= tr (YM^r Y^T)$$

(4.11)

where $tr\{,\}$ refers to the trace of the matrix and the orthogonal *intra-class* weight matrix

$$M^r = (I - W^r)^T (I - W^r)$$

(4.12)

Likewise, the *inter-class* cost function and its orthogonal *inter-class* weight matrix is derived as

$$\varepsilon_{\text{emb}}^e(Y) = tr(YM^e Y^T)$$

(4.13)

where

$$M^e = (I - W^e)^T (I - W^e)$$

(4.14)

Substituting the linear projection $Y = A^T X$ to both cost functions yield the following:

$$\varepsilon_{\text{emb}}^r(X) = tr(A^T XM^r X^T A)$$

(4.15)

$$\varepsilon_{\text{emb}}^e(X) = tr(A^T XM^e X^T A)$$

(4.16)

Motivated by Fisher’s discrimination criterion (Belhumeur et al., 1997), the objective function can be formulated to incorporate discriminative property that enables the compaction of intra-class neighborhood and the dispersion of inter-class neighborhood (see Figure 4.2 for illustration of neighborhood-level interactions). The intra-class cost function, $\varepsilon_{\text{emb}}^r$ can be minimized so that the overall weighted pairwise distances
between intra-class neighbors in embedded space are reduced. Since the total sum of weights is subjected to $\sum_j W_{ij} = 1$ or $tr(W) = I$ as seen in Step 2 of the LLE algorithm, the inter-class cost function, $\varepsilon_{emb}$ can be formulated as a local constraint,

$$A^T X M^e X^T A = I$$

(4.17)

while considering Equations (4.13) and (4.14). This owes to the fact that the reconstruction weights determining the embedding are to be bounded within the local neighborhood of each point to enforce translational invariance.

In order to maintain global rotational invariance within the embedding structure, another constraint is aptly required for further optimization of the problem. In Step 3 of the LLE, rotational invariance is achieved by subjecting $\sum_{i=1}^{N} y_i y_i^T / N = I$ or $YY^T / N = I$, resulting in what can be expressed as a global constraint,

$$A^T X X^T A = NI$$

(4.18)
Putting them altogether, the following constraint minimization problem is defined:

\[
\begin{align*}
\min & \quad A^T X M' X^T A \\
\text{s.t.} & \quad A^T X M' X^T A = I \\
& \quad A^T X X^T A = N \mathbf{I}
\end{align*}
\] (4.19)

Optimization is performed by introducing Lagrange multipliers to minimize Equation (4.19) subject to the given constraints. Thus, the problem can be modeled as the following Lagrangian with multiple constraints:

\[
L(A, \lambda_\ell, \lambda_g) = A^T X M' X^T A + \lambda_\ell (I - A^T X M' X^T A) + \lambda_g (N \mathbf{I} - A^T X X^T A)
\] (4.20)

Since there is only a single variable to solve, both constraints can actually be minimized separately (by setting the gradients with respect to \( A \) to zero) using two Lagrangians. Alternatively, they can be easily combined by unifying the constraints in Equations (4.17) and (4.18), thus reducing it to an ordinary (single) Lagrange multiplier problem,

\[
L'(\lambda, A) = A^T S A + \lambda ((N + 1)I - A^T C_\ell A - A^T C_g A)
\] (4.21)

with the following abbreviated matrix terms denoted as:

\[
S = X M' X^T 
\] (4.22)

\[
C_\ell = X M' X^T 
\] (4.23)

\[
C_g = X X^T 
\] (4.24)

By setting the gradient of Equation (4.21) to zero,

\[
\frac{\partial L'}{\partial A} = 0 \Rightarrow 2 S A - \lambda [2 C_\ell A + 2 C_g A]
\] (4.25)
it can be rewritten as a generalized eigenvalue problem,

\[ \mathbf{SA} = \lambda \left[ \mathbf{C}_\ell + \mathbf{C}_g \right] \mathbf{A} \]  

(4.26)

Note that matrices \( \mathbf{S} \), \( \mathbf{C}_\ell \) and \( \mathbf{C}_g \) are all symmetric and semi-positive definite. The optimal embedding \( \mathbf{A} \) is solved by taking \( d \) eigenvectors associated with the bottom \( (d+1) \) smallest eigenvalues \( (\lambda_1 < \ldots < \lambda_d, d \ll D) \), with the smallest eigenvector discarded to enforce zero mean for the output patterns\(^3\). If \( \left( \mathbf{C}_\ell + \mathbf{C}_g \right) \) is invertible, Equation (4.26) can be further simplified to a standard eigenvalue problem where optimal columns of \( \mathbf{A} \) are the eigenvectors of the matrix \( \left( \mathbf{C}_\ell + \mathbf{C}_g \right)^{-1} \mathbf{S} \). With this embedding, new frames from a test video \( \mathbf{X}' \) can be projected to the NDMP embedded space via the linear transformation \( \mathbf{Y}' = \mathbf{A}^T \mathbf{X}' \).

In a generalized case, a constraint tuning parameter \( \beta = \{ \beta \mid 0 \leq \beta \leq 1 \} \) can be introduced to allow both local and global constraints to be adjusted according to priority,

\[ \mathbf{SA} = \lambda \left[ \beta \mathbf{C}_\ell + (1 - \beta) \mathbf{C}_g \right] \mathbf{A} \]  

(4.27)

For instance, \( \beta = 0.5 \) can be used if one intends to obtain equal contribution from both constraints.

### 4.3.4 Summary of Algorithm

The steps of the proposed NDMP algorithm is summarized here, as well as in the published work (See & Ahmad Fauzi, 2011a). In both VFR frameworks, the purpose of NDMP is to extract features from the face exemplar images obtained from the clustering step. Hence, the input data matrix \( \mathbf{X} \) of the algorithm elaborated so far takes the form of the exemplar set \( \mathbf{E} \).

1. For each data point \( \mathbf{x}_i \in \mathbf{X} \) belonging to a certain class \( \{ c \mid 1, \ldots, M \} \), construct an intra-class neighborhood \( \mathcal{N} \) by \( K \) nearest intra-class neighbors (basically, from

\(^3\)By a different explanation (Saul & Roweis, 2003), the bottom eigenvector is the unit vector with all equal components, representing the free translation mode of eigenvalue zero.
other exemplars of the same class), and an inter-class neighborhood $\psi'$ by $K'$ nearest inter-class neighbors (from any other class).

2. Compute a pair of reconstruction weight matrices, $W^r$ and $W^e$ that best reconstructs the intra- and inter-class neighborhoods (can be solved in closed form by Eq. (4.7)).

3. Compute a pair of orthogonal weight matrices $M^r$ and $M^e$, which are defined by

$$
M^r = (I - W^r)^T (I - W^r) \quad \text{and} \quad M^e = (I - W^e)^T (I - W^e)
$$

respectively.

4. Compute optimal embedding (projection) matrix $A$ by solving the generalized eigenvalue problem of Eq. (4.26) or $XM^rX^T A = \lambda [X (M^e + I) X^T A$. The $d$ eigenvectors associated with the bottom $(d + 1)$ smallest eigenvalues are taken (while discarding the smallest eigenvector which is the mean signal).

5. Project input data $X$ to the new embedded NDMP-space $Y$ by linear projection, $Y = A^T X$. More importantly, test video data can be similarly projected for the purpose of classification matching.

A more concise pseudocode of NDMP is shown in Algorithm 1.

4.4 VFR Setup

In the setup to investigate a variety of manifold learning methods, an exemplar-based framework for VFR (previously discussed in Section 3.1.1) is used, whereby exemplars (as still images) are chosen as the features used for matching. In this case, a video-to-video recognition setting can be realized from simple image-to-image matching by learning a meaningful representation of training video data through the extraction of representative exemplars, and then performing an aggregated classification of test video frames. In this section, the selection of appropriate methods used for the two other tasks in the VFR pipeline – clustering (for exemplar selection) and classification, are described. Based on the experimental methodology described in Section 3.5, existing methods should be selected for these two tasks to create an unbiased experimental environment that is not influenced by other new algorithms proposed for the pipeline.
Algorithm 1 Neighborhood Discriminant Manifold Projection (NDMP) Algorithm

**Input:** Image points in original space, \( X = \{x_1, x_2, \ldots, x_N\} \). \( X \leftarrow E = \{e_1, 1, e_1, 2, \ldots, e_{C, M}\} \).

**Output:** Projected image points in embedded space, \( Y = \{y_1, y_2, \ldots, y_N\} \)

**Training:**

1. **for all** \( x_i \in X \) **do**
2. Construct intra-class neighborhood \( \psi \) by \( K \) nearest intra-class neighbors
3. Construct inter-class neighborhood \( \psi' \) by \( K' \) nearest inter-class neighbors
4. **end for**
5. Compute \( W^r \) and \( W^e \) by minimizing Eqns. (4.9) and (4.10)
6. Compute \( M^r \) and \( M^e \) from Eqns. (4.12) and (4.14)
7. Compute \( A \) by solving generalized eigenvalue problem in Eq. (4.26)
8. \( Y \leftarrow A^T X \)

**Testing:**
9. Projection of test video to embedded space \( Y' \leftarrow A^T X' \)

4.4.1 Clustering for Exemplar Extraction

The objective of the clustering and exemplar extraction step in the VFR setup is to perform summarization of face video sequences (from the training set) into a smaller representative set of face images, called *exemplars*. Typically, the task of clustering produces groups of face appearances that most closely resemble each other, so as to simplify the selection of exemplars. In other words, this task enables the myriad of appearance variations in a video sequence to be sufficiently characterized by a selected set of prominent appearances. Considering the abundance of face variations in each training video, a suitable dimensionality reduction method is also necessary to uncover the intrinsic low dimensional structure of the data manifold. Various works in literature (Krüeger & Zhou, 2002; Hadid & Peitikäinen, 2004; Fan et al., 2005; Fan & Yeung, 2006a) have utilized exemplars in a VFR setting to some degree of success.

In this setup, the LLE algorithm (Roweis & Saul, 2000) is first applied to
project the face appearances in a low-dimensional embedded space, for each training video. Hierarchical agglomerative clustering (HAC) algorithm is then applied to perform clustering in the new embedded feature space. HAC is a hierarchical method of partitioning data points by constructing a nested set of partitions represented by a dendrogram (see Section 5.2.2 for further details). Its optimality in clustering poses an advantage over the classical k-means algorithm, which is a primary choice for many previous works due to its simplicity in implementation. The reason is that the k-means require both the number of clusters and the initial selection of cluster seeds to be determined arbitrarily.

4.4.2 Classification

In many multiple-instance classification approaches, a majority voting scheme is a common method used to decide on the correct class label of an entire set of test patterns by determining the class that obtains the most votes across all test patterns. On the other hand, probabilistic strategies have been proposed in various VFR works (Hadid & Peitikäinen, 2004; Fan et al., 2005; W. Liu et al., 2006), and are noted to produce more reliable and robust measures compared to simple majority voting, in particularly when dealing with patterns of multiple instances.

In the experiments, a probabilistic-based scheme in the form of a Naive Bayes classifier is utilised. The subject identity in a test video $X'$ is evaluated by determining the class $c$ that maximizes the posterior probability:

$$c^* = \arg \max_c p(c|X') = p(c) \prod_{i=1}^{N} \frac{p(x'_i|c)}{p(X')}$$  (4.28)

where $p(c|X')$ is the posterior probability, $p(c)$ is the class prior while the probability of sample evidence $X'$ is constant and class-independent. Assume that observations are i.i.d, the class likelihood can be estimated using a normalized class probability score, computed for each $i$-th frame of the test video,

$$p(x'_i|c) = \frac{\sum_{j=1}^{M} 1/d_{l2}(x'_i, e_{c,j})}{\sum_{i=1}^{C} \sum_{j=1}^{M} 1/d_{l2}(x'_i, e_{s,j})}$$  (4.29)
The normalized class probability score is the inverse $L_2$-distance between test frame $x'_i$ and each training exemplar $e_{c,j}$.

A cumulative probabilistic voting strategy is also adapted from the Naive Bayes classifier by accumulating the class probability scores (or class likelihood from Equation 4.29) across all frames in the video by simple sum rule. Then, voting is performed at each frame by taking the class with the highest cumulative probability score. This strategy is employed in Section 4.5.2 for rank-based identification.

4.5 Experimental Results

Experiments were conducted based on the defined VFR setup, to evaluate the performance of the proposed method in comparison with other manifold learning methods. Two public benchmark databases – CMU MoBo (Gross & Shi, 2002) and Honda/UCSD (K. Lee et al., 2005) were used in the experiments. Details of both datasets are elaborated in Section 3.3. For both datasets, assume that the face images were extracted, resized and preprocessed based on the standard protocol defined in Section 3.4.

For each subject, one video sequence is used for training while the remaining video sequences in the dataset are for testing. The number of exemplars ($M$) selected are determined heuristically from the residual error curve of clustering distance criterion$^4$ (Duda et al., 2000) whereby $M = 6$ for CMU MoBo and $M = 7$ for Honda/UCSD. For all the experiments, the same set of parameters for the proposed NDMP method are used – fixing the intra-class and inter-class neighbors as $K = K' = M - 1$, and the tuning parameter $\beta = 0.75$ (found empirically from Fig. 4.5). The optimal feature dimension of NDMP is equivalent to the total number of classes minus one, $d = C - 1$, due to its close affinity to the formulation of Fisher’s discriminant. Meanwhile, the optimal number of feature dimensions for all other evaluated methods were determined empirically through trial experiments.

$^4$Further details can be found in Chapter 5.
Experiments were performed on the following manifold learning methods:

- Principal Component Analysis (PCA) or Eigenfaces
- Linear Discriminant Analysis (LDA) or Fisherfaces
- Locality Preserving Projections (LPP) or Laplacianfaces
- Orthogonal Locality Preserving Projections (OLPP)
- Marginal Fisher Analysis (MFA)
- Neighborhood Preserving Embedding (NPE)
- Neighborhood Discriminative Manifold Projection (NDMP)

Evaluation of these algorithms were conducted in a two-part experiment – the first is a comparative evaluation conducted to examine the performance of different manifold learning algorithms, while the second involves a rank-based identification setting to assess the reliability of the algorithms across match rank. To maintain uniformity in the test data, we perform all experiments on sequences that are of the same fixed length. For both evaluated datasets, 20 subsequences of 100 frames long were randomly sampled from each subject video in the test set, i.e. CMU MoBo_{20,1,20} and Honda/UCSD_{20,1,20}, denoted with parameters defined by the sampling protocol. A fixed subsequence length is used here to ensure that this experiment is unbiased from temporal influence of varying subsequence lengths, thus allowing better observation of the impact of feature choice towards recognition accuracy.

4.5.1 Comparative Evaluation

In the first part of the evaluation, the recognition accuracy of the proposed NDMP method is compared against classical linear projection methods (PCA, LDA) and also recent state-of-art graph-based manifold learning methods (LPP, OLPP, MFA, NPE). Classification of the sampled video subsequences was performed using the Naive Bayes probabilistic classifier. The overall recognition performance on the two datasets is summarized in Table 4.1.
Table 4.1: Average recognition rates (%) and confidence intervals of various manifold learning methods on the evaluated datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>PCA</th>
<th>LDA</th>
<th>LPP</th>
<th>OLPP</th>
<th>MFA</th>
<th>NPE</th>
<th>NDMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honda/UCSD</td>
<td>60.7</td>
<td>68.9</td>
<td>56.8</td>
<td>64.6</td>
<td>57.4</td>
<td>71.7</td>
<td>86.9</td>
</tr>
<tr>
<td></td>
<td>±1.1</td>
<td>±0.9</td>
<td>±1.5</td>
<td>±1.6</td>
<td>±1.8</td>
<td>±0.6</td>
<td>±0.3</td>
</tr>
<tr>
<td>CMU MoBo</td>
<td>86.6</td>
<td>92.6</td>
<td>89.3</td>
<td>90.2</td>
<td>91.4</td>
<td>96.3</td>
<td>97.7</td>
</tr>
<tr>
<td></td>
<td>±1.1</td>
<td>±0.9</td>
<td>±1.5</td>
<td>±1.6</td>
<td>±1.8</td>
<td>±0.6</td>
<td>±0.3</td>
</tr>
</tbody>
</table>

From the results, the proposed NDMP method is clearly superior to the other evaluated methods in its ability to recognize the identity of persons in the test subsequences. It also achieved a relatively small confidence interval of ±0.3% (as can be seen from the CMU MoBo results), which is indicative of its consistency and reliability. Meanwhile, the effectiveness and robustness of NDMP over the rest of the methods are more apparent in the challenging Honda/UCSD dataset (with standard deviation of ≈10.5% across all methods), where the test videos possess a wide range of complex face poses and out-of-plane head rotations. On the contrary, the results of the CMU MoBo dataset are spread across a standard deviation of ≈4% across all methods, possibly an indication of a relatively easier set of facial variations. Nonetheless, the NDMP remains the best performing algorithm among the evaluated methods. Surprisingly, LPP and MFA performed rather poorly in this dataset. One possible reason could be the insensitivity of the heat kernel weights defined for their affinity matrices, which do not model the local structure of the face manifold very naturally from the constructed adjacency graph.

4.5.2 Rank-based Identification

To further evaluate the reliability of the NDMP method in a rank-based identification setting, the performance of the evaluated methods is assessed based on Cumulative Match Characteristic (CMC) curves. CMC curves are commonly used to depict the rate of increase in identification with increase in the rank of the system, which is a reliability indicator for identification (one-to-many matching) systems (Bolle, Connell,
This time, a cumulative probabilistic voting strategy is adopted. The normalized class probability scores (Equation 4.29) are cumulatively aggregated by simple sum rule. A vote is taken at each frame and the class with the majority vote is classified as the matched subject of the test subsequence. In this setting, the test subsequence can also be matched with the $R$ highest voted classes, which corresponds to rank $R$ in a $C$-person dataset (where $R \leq C$). Interestingly, this classification method does not classify as well as the Naive Bayes used in the previous experiment, but it affords us further insight into their performances across rank.

From the CMC curves of various manifold learning methods evaluated on the CMU MoBo (Figure 4.3) and Honda/UCSD (Figure 4.4) datasets, it is observed that the proposed NDMP algorithm consistently yielded better recognition rates across the ranks. Again, the NDMP is particularly reliable in the more challenging Honda/UCSD dataset while it also managed a perfect recognition score (100%) for the CMU MoBo dataset.

![Figure 4.3: Comparison of cumulative match characteristic (CMC) curves of various manifold learning methods on the CMU MoBo dataset.](image)
dataset within the top 3 matches. It is not conclusive why the CMC curves on the Honda/UCSD dataset appear to be more erratic, unlike typical smooth exponential trend of CMC curves. One possible reason lies in the abundance of difficult facial variations in this dataset, further exacerbated by the large random variety of subsequences sampled (whereby 100-frame subsequences are randomly sampled from original long videos of about 250-350 frames long).

The proposed algorithm also achieved a significant improvement over the NPE algorithm, its "predecessor" or the most related method that shares the same theoretical foundation. The cutting edge demonstrated by the NDMP algorithm highlights the undeniable importance of preserving the global manifold structure while providing discriminatory power between local neighborhood structures. One can observe as well that purely "globalized" manifold learning methods such as PCA and LDA performed rather poorly due to their inability to learn the nonlinear manifold of appearance variations that is inherent in videos. Meanwhile, the performance of other graph-based "localized" manifold learning methods (MFA, LPP, NPE, NDMP) tend to improve
rapidly as the rank increases, owing much to the preservation of local neighborhood geometry.

4.5.3 Analysis and Discussions

In the experiments, the constraint tuning parameter $\beta$ of the generalized NDMP (see Equation 4.27) is a free parameter that can be adjusted to values between 0 and 1 to alter the balance between contributions of the local neighborhood constraint and global invariance constraint. For the experiments in the previous section, $\beta = 0.75$ is chosen by empirical means, that is by analysing the effect of varying this parameter on the overall recognition performance. Note also that in the experiments, the number of exemplars $M$ are heuristically determined in the clustering step (see Section 5.2.4 for more details) and the largest possible neighborhood size $K = K' = (M - 1)$ is used.

Analysis is further extended by varying the neighborhood sizes with $K = \{5, 6, 7\}$.

There are some insightful observations from the results in Figure 4.5. Firstly, it may appear at a glance that both constraints seemed equally important, but marginally better results can be expected by imposing more influence towards constraining the local neighborhood structure (larger $\beta$ values) than the global manifold invariance (smaller $\beta$ values). Secondly, the recognition capability of the proposed algorithm outclasses the NPE algorithm, regardless of the choice of neighborhood size $K$ based on this limited experiment (where only three $K$ values were considered). This points towards the fundamental strength in the formulation of the NDMP algorithm, in which the preservation and discrimination of local neighborhood structures is essential for optimal embedding in a lower dimensional feature space.

4.6 Summary

In this chapter, a novel supervised manifold learning algorithm called Neighborhood Discriminative Manifold Projection (NDMP) for feature representation in video-based face recognition is presented. The NDMP constructs an optimal discrimin-
Figure 4.5: Comparison of different values of $\beta$ on the Honda/UCSD dataset.

inative eigenspace projection of a high dimensional face manifold in a lower dimensional embedded space based on the preservation of both local neighborhood structure and global manifold structure. The local geometry is preserved through the use of intra-class and inter-class neighborhood information, while the global manifold structure is safeguarded by imposing rotational invariance within the embedding. These two tasks are well-posed as a constraint minimization problem, which can be further reduced to a straightforward generalized eigenvalue problem, which possesses an optimal solution for linearly projecting new data onto the embedding. Extensive experiments on benchmark video face datasets demonstrated the superiority of the proposed NDMP method compared to conventional linear projection methods and recent state-of-the-art methods. Its robustness in handling the diverse collection of face variations inherent in video sequences is evidential from the empirical results reported.

This preliminary work is by no means exhaustive, and there is much room for further progress. For instance, the NDMP method can be further extended to prevent cases of over-generalization or underfitting of data, by mapping to a much richer kernel feature space (via "kernel trick" for implicit mapping to feature space). Alternatively,
kernel matrices can be derived from the sparse weighted graph rather than utilizing a pre-defined kernel function (as suggested by Ham, Lee, Mika, and Schölkopf (2004)). For better practicality, the proposed method should be further tested with larger face video datasets or real-world video footage (with significant amount of untrained patterns) in order to further substantiate its robustness and capability in large-scale recognition.
In this chapter, a new clustering technique that incorporates temporal dynamics into the selection of meaningful clusters, called Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC) is formulated. The abbreviation STHAC is coined after the original Hierarchical Agglomerative Clustering (HAC) in pattern classification literature (Duda et al., 2000).

Two variants for STHAC are proposed, both of which are formulated to exploit the inherent ordering of video frames or data points, by means of their distances in spatial and temporal dimensions:

1. A *global fusion* variant (STHAC-GF), that blends the contribution of spatial and temporal distances among points in space. This method involves the normalized combination of both measures in global space-time dimension.

2. A *local perturbation* variant (STHAC-LP), that perturbs the spatial and temporal distances based on a local spatio-temporal neighborhood criterion. Spatial and temporal local neighborhoods for each point are mutually considered to form a common spatio-temporal neighborhood where neighbor and non-neighbor (foreign) points are individually perturbed to emphasize their respective similarities and dissimilarities.

To overcome the nonlinearity formed by face appearances in video sequences, Locally Linear Embedding (LLE) is first applied to each training video to uncover the intrinsic embedding of faces in lower-dimensional space. This is a preparatory step before

---

1 Video frames or face images are often loosely referred to as ’points’ in high dimensional space.
applying clustering so that face appearances of the same person under variations in pose, illumination, expression, etc. are more distinguishable. By way of an exemplar-based recognition framework established in Section 3, the following steps involved feature representation by selected state-of-art dimensionality reduction methods, and classification of subjects using a probabilistic-based Bayes classifier. Extensive experiments on benchmark face video databases testify of the importance of both spatial and temporal information in the selection of clusters features.

The organization of the chapter is as follows: Section 5.1 reviews some related work on the extraction of cluster features (image set or exemplars) for VFR, and some interesting spatio-temporal approaches. In Section 5.2, the basic HAC method is introduced followed by a detail elaboration of the proposed STHAC method. Section 5.2.3 discusses matters concerning the selection of number of clusters. Section 5.3 briefly mentions the setup of other tasks in the framework. Section 5.4 describes the experiments performed on various methods for extracting cluster features, and finally, Section 5.5 concludes the chapter.

5.1 Related Work

The use of clustering in the face of large collections of data is primarily motivated by the need to classify or categorize data into groups of quasi-similar data by an automatic, unsupervised manner. In the case of the variety of face appearances in a video, clustering is a useful task (before feature extraction) that groups together face images of somewhat similar appearances. Feature-wise, this is beneficial as the face images within each cluster depict tightly formed local manifolds that can be characterized as linear subspaces, or similarly referred to in various works as "local models" (K. Lee et al., 2005; Fan & Yeung, 2006b) or "local linear patches" (R. Wang et al., 2008). Some other works further summarize each cluster by extracting a representative face image, or exemplar, which can sufficiently characterize the appearances within the cluster. This simplification of representation is clearly motivated by the ease of performing conventional image-based matching, and also the reduction of storage space for training data from long video streams.
A typical face video sequence may contain large variations in pose, expression, illumination, video quality and other possible occlusions. Generally, the distribution of these video frames (data) in its original high-dimensional sample space constitutes a complex manifold that is nonlinear in nature. This poses a challenge to clustering in Euclidean space. Brand (2003) hypothesized that it can be presumed that data more often, lies on or near a low-dimensional manifold embedded in the sample space, and there exists a nonlinear mapping between the two spaces. This motivates the use of nonlinear dimensionality reduction techniques such as LLE (Roweis & Saul, 2000) and Isomap (Tenenbaum et al., 2000) to "unroll" the data before proceeding with clustering. In literature, it is common to refer to this process as spectral clustering (Ng, Jordan, & Weiss, 2001), or more concretely, the task of clustering data points using eigenvectors of matrices derived from the data.

Among the many methods in literature reviewed earlier, most image set-based methods attempt to model a set of face images either parametrically using probability distributions (Shakhnarovich et al., 2002; Arandjelović et al., 2005) or nonparametrically via subspace learning techniques (Fukui & Yamaguchi, 2003; T. Kim et al., 2007; Cevikalp & Triggs, 2010). In most general scenarios, image sets may consist of multiple unordered or arbitrarily selected observations, where the modeling of each image set entirely as a single feature vector disregards the local variabilities present within a set. Some approaches (K. Lee et al., 2005; Fan & Yeung, 2006b) construct a set of local linear models based on the cluster sets found from the proceeding clustering task. This procedure well-embodies the variabilities in the training data and better captures the person-specific features for each class.

VFR works in literature that are exemplar-based have demonstrated promising results with a majority of them employing a combination of different dimensionality reduction and clustering algorithms. One of the early works by Krüeger and Zhou (2002) proposed a method for automatic selection of exemplars by learning a set of clusters from each gallery video using a probabilistic online-learning approach. The exemplars are then used as centers of mixture distribution models for both tracking and
recognition processes. The authors argued that the major drawback of *k*-means clustering revolves around the prior selection of *k*, thus a dynamic choice of the number of clusters is essential. Further to that, the proposed probabilistic framework characterizes the kinematics (temporal) and identity (spatial) of face information and applies the sequential importance sampling (SIS) algorithm to estimate the posterior distribution of the identity variable (Zhou et al., 2003).

Hadid and Peitikäinen (2004) proposed an appearance-based scheme which embeds the face manifold in a lower dimensional space using the LLE algorithm. The data are then partitioned using *k*-means clustering in the LLE feature space, and the exemplars are defined as the set of cluster centers. Finally, a probabilistic voting strategy is used to evaluate the recognition ability of their system. Their experimental analyses point towards the capability of the LLE in building a robust low-dimensional embedding that enables good exemplars to be selected. Fan et al. (2005) also opt for the similar configuration except that classification of test videos is performed using a Bayesian inference model to exploit temporal dynamics between frames. While the simplicity of applying *k*-means clustering is appealing, it has some glaring limitations – notably it is sensitive to the initial seeds used, and secondly, that in turn may sometimes produce sub-optimal results due to its inability to find the global minima.

Better recognition performance can be obtained by extracting exemplars using hierarchical agglomerative clustering (HAC) (Fan & Yeung, 2006a) compared to conventional *k*-means clustering. The authors also preferred the use of geodesic ("shortest path") distances as opposed to Euclidean distance, an idea employed by the Isomap algorithm for low-dimensional embedding. In their evaluation, this combination only appears to perform slightly better than spectral methods (using LLE, Isomap) that explicitly embed the original data resulting in a certain degree of information loss. In a more recent work by R. Wang and Chen (2009), hierarchical divisive clustering (HDC) algorithm is used to discover local linear models on a face manifold.

While the dearth of methods that leverage temporal information for cluster-
ing in pattern recognition literature is obvious, there are a few interesting works that introduce the usage of "spatio-temporal clustering" per se. W. Liu et al. (2006) formulated the recognition problem to learn a Spatio-Temporal Embedding (STE) from raw videos. Keyframes or exemplars across different video sequences are extracted in a synchronized fashion by a modified $k$-means algorithm using a spatio-temporal objective function which iteratively updates an adaptive spatial distance metric and a simple temporal order metric. This synchronization of frame clustering assumes that face appearances are homogenously distributed across all videos for every subject. The dual-metric function in this work offers inspiration to the proposed scheme as to how spatial and temporal distances can be integrated.

In an extension to the Isomap algorithm, Jenkins and Mataric (2004) expanded the existing framework to consider temporal relationships in local neighborhoods to uncover spatio-temporal structure in data. Their main application of learning an embedding for real-world data collected from human motion and humanoid robot tele-operation showed some promising results. The evaluated spatio-temporal data can be easily flattened into low dimensional manifolds via good selection of clusters, albeit difficulties in processing large data sets.

The term "spatio-temporal clustering" is also discussed at length in data mining and other relevant scientific applications such as geographic information systems, medical imaging and weather forecasting, where long-term evolution of data can be analyzed (Kisilevich, Mansmann, Nanni, & Rinzivillo, 2010). Two spatio-temporal clustering algorithms – ST-GRID (M. Wang, Wang, & Li, 2006) and ST-DBSCAN (Birant & Kut, 2007), were proposed for analysis of sequences of seismic events, whereby finding clusters among these events involve discovering groups that lie close both in time and in space. The ST-DBSCAN algorithm in particular, introduces the concept of spatial and temporal neighborhoods, where the temporal neighbors ("observed in consecutive time units") are first retained before their corresponding spatial (geographical location) and non-spatial values (temperature) are considered for clustering.
This work is motivated by the main ideas of multi-distance integration and spatio-temporal neighborhoods laid out in (W. Liu et al., 2006) and (Jenkins & Mataric, 2004; Birant & Kut, 2007) respectively. Unlike these approaches, the novel formulation of the STHAC algorithm have several strengths that are specific to the task of training raw videos. Firstly, each video used in the training process is not treated as a single data collected over a span of time, like geographical or weather information. Instead, the frames within each video are treated as a set of data spanned across time. Thus, each frame is to be modeled as a spatio-temporal data point, and the spatial and temporal distances between points can then be normalized and neatly fused into various forms. Secondly, it leverages on the advantages of agglomerative ("bottom-up") clustering, which removes random seed initialization and distribution modeling that are characteristic of non-hierarchical clustering methods. To maintain relevancy within the scope of this thesis work, the success of the proposed clustering approach is evaluated by means of the recognition accuracy obtained in the VFR framework.

5.2 Proposed Method

In this section, the use of nonlinear dimensionality reduction by the LLE method is first presented. Then, the classical HAC algorithm is introduced, followed by a detailed elaboration of the proposed STHAC algorithm. Finally, the heuristic selection of number of clusters is discussed.

5.2.1 Dimensionality Reduction

Considering the large amount of face variations in each training video sequence, a suitable dimensionality reduction method is necessary to uncover the intrinsic structure of the data manifold which may originally lie on a complex hyperplane. Recent nonlinear dimensionality reduction techniques such as Locally Linear Embedding (LLE) (Roweis & Saul, 2000) and Isomap (Tenenbaum et al., 2000) have been proven effective at seeking a low-dimensional embedding of a data manifold. LLE in particular, is known for its capability in modeling the global intrinsic structure of the manifold while preserving local neighborhood structures to better capture various face variations such as pose, expression and illumination. Conventional unsupervised
methods such as Principal Component Analysis (PCA) (Turk & Pentland, 1991) and Multidimensional Scaling (MDS) (Cox & Cox, 2001) have the tendency of overestimating the intrinsic dimensionality of high-variability data.

For each training video, the LLE algorithm is applied to project the face variations in a low-dimensional embedded space, in preparation for the subsequent clustering step. LLE involves solving a quadratic constrained least squares problem based on local symmetries and linear coefficients, which can be efficiently computed using sparse eigenvector matrices. It also scales well with the intrinsic manifold dimensionality, $d$, thus do not require re-running the algorithm for higher-dimensional embeddings. To further demonstrate the effectiveness of applying LLE before clustering, Figure 5.1 shows the effect of clustering the data points on the LLE spectral plot (similar to that in Figure 4.1(d)) where the face images shown are the exemplars selected from the cluster means. It can be clearly observed how the LLE algorithm embeds the data points into a few distinct hyperplanes in low dimensional space. This lends a helping hand to the subsequent clustering process as the distinguishable groups of face appearances in the new embedding enable better clusters to be formed.

![Figure 5.1: Linear cluster patches (in different colors) in LLE space after applying HAC for clustering](image_url)
5.2.2 Hierarchical Agglomerative Clustering (HAC)

Hierarchical Agglomerative Clustering (HAC) is a hierarchical method of partitioning data points by constructing a nested set of partitions represented by a cluster tree, or dendrogram (Duda et al., 2000). Figure 5.2 shows a sample dendrogram. The agglomerative approach works from bottom-up by grouping smaller clusters into larger ones, as described in the following procedure:

1. Initialize each data point as a singleton cluster \( \Phi_i \). At the start, there are \( N_c \) clusters.

2. Find the nearest pair of clusters, \( \Phi_i \) and \( \Phi_j \) according to a certain linkage criterion that determines the distance between clusters. Commonly used linkage criteria are such as single-linkage, complete-linkage, average-linkage and Ward’s criterion. Merge the two nearest clusters to form a new cluster.

3. Continue merging (repeat Step 2) and terminate when all points belong to a single cluster.

The required number of clusters, \( M \) is selected by partitioning at the appropriate level of the dendrogram only after clustering task is completed. This poses an advantage over the classical \( k \)-means algorithm, which has been a primary choice of many previous works (Fan et al., 2005) due to its simplicity in implementation. The \( k \)-means has some glaring limitations, particularly in the requirement that the number of clusters be predetermined without knowledge of how the clusters are distributed proximally. Thus, initial selection of cluster seeds (which can differ in every run) is also not required. Structurally, hierarchical merging ensures that the resulting clusters are near-optimal and does not get trapped in local minima as for the case of \( k \)-means.

5.2.3 Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC)

The proposed Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC) differs from the standard HAC in terms of the computation of the nearest pair of clus-
Figure 5.2: A sample dendrogram of $M = 11$ clusters. To select $M$ clusters, the dendrogram collapses the lower branches of the tree to show no more than $M$ leaf nodes representing the clusters. The height of each partition represents the distance between each two connected clusters.

ters (Step 2). In this work, a spatio-temporal distance metric is introduced by utilizing both spatial and temporal information to influence the distance between data points in space. Since clustering methods generally utilize only the spatial distances between points, there is no consideration for temporal distances or "distance in the time domain" between points. In video sequences, temporal ordering is an inherent property that can be exploited to better characterize clusters.

Does the addition of temporal information (or distances to be exact) lead to better clustering in face video sequences? Intuitively, outliers can affect groupings if purely spatial or purely temporal clustering is applied. Figure 5.3 shows a simple illustration outlining these observations. In spatial clustering where faces are grouped by locality in "feature space", similar faces that occur at a different time (either earlier or later) may be erroneously assigned to a different cluster, or result in the creation of new clusters that contain duplicated appearances. In temporal clustering where
faces are grouped by sequence in "time space", it is possible that different faces may be assigned to the same cluster when a rapid change of appearances occurs. Both methods described thus far (k-means and HAC) are both employ spatial clustering.

Generally, spatial distance is measured by simple Euclidean distance between points,

$$d_S = \|x_i - x_j\| = \sqrt{(x_i - x_j) \cdot (x_j - x_i)} \quad (5.1)$$

Temporal distance can be measured by the time spanned between two frame occurrences ($x_i$ and $x_j$) in a video sequence,

$$d_T(x_i, x_j) = |t_{x_i} - t_{x_j}| \quad (5.2)$$

where $t$ is the discretized unit time. This straightforward definition is sufficiently intuitive to quantify temporal relationships across sequentially ordered samples.

Before elaborating further on the two variants proposed for the STHAC scheme, the first step involves computing the normalized pairwise spatial and temporal distances between all samples in a video, represented here by the $N \times N$ matrices, $D_S(x_i, x_j)$ and $D_T(x_i, x_j)$ respectively.

![Figure 5.3: Illustration of how data points might be grouped](image-url)
5.2.3 (a) Global Fusion variant (STHAC-GF)

The global variant blends the contribution of spatial and temporal distances using a tuning parameter, \( \alpha \). The tuning parameter adjusts the blending factor defined by its upper and lower bounds, \( p_{\text{max}} \) and \( p_{\text{min}} \) respectively, which act to increase or reduce the original distances. Thus, the spatio-temporal distance through global distance fusion is defined as,

\[
\overline{D}_{STg} = (p_{\text{max}} - \alpha) \overline{D}_S + (p_{\text{min}} + \alpha) \overline{D}_T, \quad 0 \leq \alpha \leq 1 \tag{5.3}
\]

5.2.3 (b) Local Perturbation variant (STHAC-LP)

The local variant perturbs the spatial and temporal distances based on a local spatio-temporal neighborhood criterion. The criterion is formulated by spatio-temporal relationships between a point and its neighbors. For each point \( x_i \), the following neighborhoods are defined:

- **Spatial neighborhood,**

  \[
  Q_{x_i} = \{x_{i,1}, x_{i,2}, \ldots, x_{i,k}\} \tag{5.4}
  \]

  which is a set of points containing the \( k \)-nearest neighbors of \( x_i \) computed by Euclidean distance.

- **Temporal neighborhood,**

  \[
  S_{x_i} = \{x_{i-w}, \ldots, x_i, \ldots, x_{i+w}\} \tag{5.5}
  \]

  as a set of points of length \( (2w + 1) \) representing a temporal window segment centered upon \( x_i \).

A point \( x_j \) is identified as a *common spatio-temporal neighbor* (CSTN) of point \( x_i \) if it belongs to both spatial and temporal neighborhood point sets. Hence, the criterion
defining the CSTN set of point $x_i$,

$$\text{CSTN}_{x_i} = Q_{x_i} \cap S_{x_i}$$  \hspace{1cm} (5.6)$$

The remaining points of the temporal neighborhood, $S_{x_i}$, that do not fulfill this criterion are non-neighbors or foreign points within the defined spatio-temporal neighborhood (for easy reference, the term non-CSTNs would suffice to identify). Figure 5.4 shows a graphical example of how the CSTN and $\neg$CSTN sets are constructed. The spatial neighborhood $Q_{x_i}$ considers points within the yellow elliptical area, while the temporal neighborhood $S_{x_i}$ is indicated by the blue thick line that connects the points. The CSTN set consists of points that belong to both both neighborhoods (intersection of the two sets), shown with green borders. On the other hand, the non-CSTN (or $\neg$CSTN) set comprises the remaining points in $S_{x_i}$ that are not in the CSTN set, shown with red borders.

Perturbations can be applied to the spatial distances without cumbersome modeling of points in time-space dimension. This procedure artificially alters the original distances by "rewarding" similarities (CSTN set) and "penalizing" dissimilarities (non-CSTN set).
CSTN set). For each point $x_i$, a corresponding point $x_j$ from either the CSTN$_x$ or $\neg$CSTN$_x$ sets is applied an appropriate perturbation to either decrease (reward) or increase (penalize) the distance between points $x_i$ and $x_j$. Hence, the perturbation affinity matrix is defined as,

$$
P_{ij} = \begin{cases} 
1 - \lambda_{sim}, & \text{if } x_j \in \text{CSTN}_x, \\
1 + \lambda_{dis}, & \text{if } x_j \in (S_x \setminus \text{CSTN}_x) \\
1, & \text{otherwise}
\end{cases}
$$

(5.7)

where $\lambda_{sim}$ and $\lambda_{dis}$ are the similarity and dissimilarity perturbation constants respectively, taking appropriate values of $0 < \{\lambda_{sim}, \lambda_{dis}\} < d(x_i, x_j)$. It makes logical sense to use only one perturbation parameter, $\lambda = \lambda_{sim} = \lambda_{dis}$, to ensure perturbations for both sets of points are fairly applied, while further simplifying parameter tuning.

In short, $P_{ij}$ seeks to accentuate the similarities and dissimilarities between data samples by artificially reducing or increasing spatial distances between all samples based on a spatio-temporal neighborhood. By matrix multiplication, the spatio-temporal distance through local neighborhood perturbation is defined as,

$$
\overline{D}_{ST} = P_{ij}(\overline{D}_S + \overline{D}_T)
$$

(5.8)

The choice of linkage criteria for merging clusters in the agglomerative procedure is the Ward’s distance criterion,

$$
d_{\text{WARD}}(\Phi_i, \Phi_j) = \frac{n_in_j}{n_i + n_j} \|\mu_i - \mu_j\|^2
$$

(5.9)

where $\mu_i$ and $\mu_j$ are means of cluster $i$ and $j$, while $n_i$ and $n_j$ are the number of points in their respective clusters. The Ward’s criterion takes into account the total within-cluster variance where the merging cost of combining two clusters considers the distribution of samples. Linkages are based on the averaged relevance of all points within the clusters unlike the single-link (min) and complete-link (max) distance criterion.
In this work, the parameters for both STHAC-GF and STHAC-LP are set to optimal values, as empirically determined by experiments. These values for use in both variants are shown in Table 5.1. In Section 5.4.3 of the experiments, the result of varying some of these parameters is further examined and discussed.

5.2.4 Heuristic Selection of Number of Clusters

The selection of correct or optimum number of clusters has little theoretical foundation and meaning. Very often, suitable heuristics are devised to provide good choices. In this case, the cluster merging cost of Ward’s criterion from Equation 5.9 takes the form of a sum-of-squares term, which is known to provide a good statistical measure for residual error (Duda et al., 2000).

Figure 5.5 shows the residual error (or cluster merging cost) curve obtained by extracting different number of clusters from three different training videos of the same dataset. A simple but effective heuristic is to find the "elbow" of the curve, at which the curve stops decreasing significantly with the increase of clusters. To facilitate an unbiased representation of features for classification, the number of clusters extracted from each class of the same dataset is fixed.

In VFR, there are some cluster-based approaches that allow each training video to be summarized into a different number of clusters (Fan & Yeung, 2006a; R. Wang et al., 2008), as many as it is needed to represent the face variations. This option may seem advantageous in dealing with video of different sequence lengths or varying degree of face variations, but it remains inconclusive if this configuration has substantial effect on the actual task (learning/recognition) or application. On the other hand, there

| Table 5.1: Optimal values for parameters used in STHAC-GF and STHAC-LP |
|-----------------|-----------------|-----------------|
| STHAC-GF        | STHAC-LP        |
| $p_{min}$       | 0.5             | $k$             | 7               |
| $p_{max}$       | 1.5             | $w$             | 7               |
| $\alpha$        | 0.75            | $\lambda$      | 0.25            |
Figure 5.5: Residual error plot of three different training videos (drawn with different colors) partitioned with different number of clusters. The "elbow" of the curve is approximately at values 5 to 9.

are other approaches (Fan et al., 2005; R. Wang & Chen, 2009) that employ a similar strategy as this work by choosing to fix the number of clusters extracted from all training videos of the same dataset. This is often useful to establish a consistent set of features in order to learn feature embeddings or perform component analyses in the next step.

5.2.5 Summary of Algorithm

The steps of the proposed STHAC algorithm (in both variations—Global Fusion and Local Perturbation) are summarized here, as well as in the published work (See & Eswaran, 2011). In this approach, the clustering step is to be performed once for all C number of classes. Each training video is assumed to contain a single subject (class). In the case of multiple training videos, all the videos belonging to the same subject can be concatenated before performing clustering.

1. Initialize each data point $x_i \in X$ as a singleton cluster $\Phi_i$. There are $N_c$ number of clusters at the beginning. Also, the eventual number of clusters $M$ is determined.
2. Compute normalized pairwise spatial distance matrix $\overline{D}_S(x_i, x_j) \in \mathbb{R}^{N_c \times N_c}$ and normalized pairwise temporal distance matrix $\overline{D}_T(x_i, x_j) \in \mathbb{R}^{N_c \times N_c}$ between all data points in $X$.

(Go to step 3 if Global Fusion method is used, and step 4 if Local Perturbation method is used.)

3. Global Fusion variant
   a. Compute global spatio-temporal distance matrix $\overline{D}_{STg} = (p_{max} - \alpha)\overline{D}_S + (p_{min} + \alpha)\overline{D}_T$

4. Local Perturbation variant
   a. For all data points $x_i \in X$,
      i. Compute spatial neighborhood $Q_{x_i}$ using $k$-nearest neighbors by Euclidean distance.
      ii. Compute temporal neighborhood $S_{x_i}$ based on a window segment of $(2w + 1)$ points centered upon $x_i$.
      iii. Extract common spatio-temporal neighbor (CSTN) subset of points whereby $\text{CSTN}_{x_i} = Q_{x_i} \cap S_{x_i}$. The remaining points from both neighborhoods $Q_{x_i}$ and $S_{x_i}$ (non-neighbors) form the non-CSTN subset.
   b. Compute perturbation affinity matrix $P_{ij}$ in Eq. (5.7) by rewarding CSTN points and penalizing non-CSTN points
   c. Compute local spatio-temporal distance matrix $\overline{D}_{STl} = P_{ij}(\overline{D}_S + \overline{D}_T)$

5. Find clusters whose merger changes the linkage criterion the least (or generally by taking the closest pair of clusters). Ward’s criterion (Eq. (5.9)) is chosen in this algorithm.

6. Merge the two found clusters $\Phi_i$ and $\Phi_j$, and update the cluster set $O$.

7. Repeat Steps 5 and 6 until $M$ number of clusters are left.

A more concise pseudocode of the STHAC algorithm, with both global and local variants, is shown in Algorithm 2.
Algorithm 2  Spatio-Temporal HAC Algorithm

Input: Training video sequence, \( X_c = \{x_1, x_2, \ldots, x_{N_c}\} \forall c \in \{1, \ldots, C\} \)

Output: Assembly of final clusters, \( O = \{\Phi_1, \ldots, \Phi_M\} \)

Initialize: \( M \), Current number of clusters \( \hat{c} \leftarrow N_c \)

1: \textbf{for} \( i \leftarrow 1, N_c \) \textbf{do}
2: \hspace{1em} \( \Phi_i \leftarrow \{x_i\} \)
3: \textbf{end for}

Clustering:

4: \textbf{for} \( j \leftarrow 1, N_c \) \textbf{do}
5: \hspace{1em} \( \overline{D}_S(x_i, x_j) \leftarrow ||x_i - x_j|| \)
6: \hspace{1em} \( \overline{D}_T(x_i, x_j) \leftarrow |t_{x_i} - t_{x_j}| \)
7: \textbf{end for}

8: \textbf{if} Global method selected \textbf{then}
9: \hspace{1em} \textbf{procedure} GLOBAL-FUSION\((p_{\text{min}}, p_{\text{max}}, \alpha)\)
10: \hspace{2em} \( \overline{D}_{STg} \leftarrow (p_{\text{max}} - \alpha)\overline{D}_S + (p_{\text{min}} + \alpha)\overline{D}_T \)
11: \hspace{1em} \textbf{end procedure}
12: \textbf{else} Local method selected
13: \hspace{1em} \textbf{procedure} LOCAL-PERTURBATION\((k, w, \lambda)\)
14: \hspace{2em} \textbf{for all} \( x_i \in X \) \textbf{do}
15: \hspace{3em} Compute \( Q_{x_i} \) and \( S_{x_i} \)
16: \hspace{3em} Extract local CSTN_{x_i} subset by Eq. (5.6)
17: \hspace{2em} \textbf{end for}
18: \hspace{2em} Compute \( P_{ij} \) by Eq. (5.7)
19: \hspace{2em} \( \overline{D}_{STl} \leftarrow P_{ij}(\overline{D}_S + \overline{D}_T) \)
20: \hspace{1em} \textbf{end procedure}
21: \textbf{end if}

22: \textbf{repeat}
23: \hspace{1em} \( \hat{c} \leftarrow \hat{c} - 1 \)
24: \hspace{1em} \( (i, j) \leftarrow \arg \min_{\forall (i, j) \in \{1, \ldots, \hat{c}\} \land (i, j) \neq j} d_{\text{WARD}}(\Phi_i, \Phi_j) \)
25: \hspace{1em} Merge \( \Phi_i \) and \( \Phi_j \), Update clusters in \( O \)
26: \textbf{until} \( \hat{c} = M \)
5.3 VFR Setup

Befitting the scope of this work, the exemplar-based framework for VFR (discussed in Section 3.1.1), which utilizes features from exemplars for recognition, is used. A video-to-video recognition setting is generalized from simple still-to-still matching by extracting exemplars from training video data after learning a meaningful representation in low dimensional space, followed by an aggregated classification of test video frames. In this section, the selection of appropriate methods used for the two other related tasks in this VFR pipeline – feature representation and classification, are briefly described. Based on the experimental methodology described in Section 3.5 (where clustering is Task II), the new feature representation approach proposed in Chapter 4 is now chosen while an existing approach should still be selected for the classification task to maintain an unbiased experimental environment for evaluating the clustering method.

5.3.1 Feature Representation

Traditional linear projection methods such as PCA and LDA have been widely used to great effect in characterizing data in smooth and well-sampled manifolds. More recently, manifold learning methods such as Locality Preserving Projections (LPP) and Neighborhood Preserving Embedding (NPE) have gained much attention due to their ability in approximating the low-dimensional embedding of nonlinear manifolds. NPE in particular, has an attractive neighborhood-preserving property due to its formulation based on LLE. For improved feature representation, the proposed Neighborhood Discriminative Manifold Projection (NDMP) (See & Ahmad Fauzi, 2011a, 2011b), a supervised discriminative variant of the NPE, can be applied. Details of the NDMP are elaborated extensively in Chapter 4.

During the training phase, the NDMP features of all $M \times C$ exemplars extracted from the training videos are learnt. By solving the generalized eigenvalue problem introduced in Equation (4.26), the projection matrix $A$ of the low-dimensional embedding can be obtained. In the test phase, new frames from a test video $X'$ can be
projected to the embedded space via the linear transformation $Y' = A^T X'$.

### 5.3.2 Classification

Similar to the choice of classification method for experiments on manifold learning (in Chapter 4), a probabilistic-based scheme in the form of a Naive Bayes classifier is also employed. The subject identity in a test video is evaluated by determining the class that maximizes the posterior probability (shown in Equation (4.28)), which in turn can be estimated by a normalized class probability score (shown in Equation (4.29)) computed for each frame of the test video. Likewise in some parts of the experiments, a cumulative probabilistic voting strategy is also used, as preferred over the conventional majority voting as it is more robust and it encourages the probability scores at each frame to be carried over to subsequent frames through accumulation. This somewhat promotes the retention of temporal information across frames in a video as opposed to disjointed voting in each frame.

### 5.4 Experimental Results

Experiments were conducted based on the VFR setup defined in Section 5.3 on two public benchmark datasets – CMU MoBo (Gross & Shi, 2002) and Honda/UCSD (K. Lee et al., 2005). Details of both these datasets have been elaborated in Section 3.3. For both datasets, face images were extracted, resized and preprocessed based on the standard protocol defined in Section 3.4.

For each subject, one video sequence is used for training and the remaining video sequences in the dataset for testing. Based on the "elbow" observation of residual error curves, the number of exemplars per training class are selected as $M = 6$ for CMU MoBo and $M = 7$ for Honda/UCSD. STHAC tuning parameters were set to optimal values determined empirically, as shown in Table 5.1. Figure 5.6 shows selected exemplars of three subjects from CMU MoBo and Honda/UCSD datasets. Experiments are performed on the following exemplar extraction methods for compar-
Figure 5.6: Selected exemplars of CMU MoBo (top three rows) and Honda/UCSD (bottom three rows) datasets extracted using the STHAC-GF method, showing the most representative faces of each person.

• Random exemplar selection
• PCA + $k$-means
• LLE + $k$-means (as proposed by Hadid and Peitikäinen (2004))
• Geodesic distances$^3$ + HAC (as proposed by Fan and Yeung (2006a))
• LLE + HAC
• LLE + STHAC-GF
• LLE + STHAC-LP

Two sets of experiments were conducted – the first is conducted by sampling fixed-length subsequences to assess the effectiveness of the approaches (unbiased by

$^2$All methods, except the first, are denoted by the combination of dimensionality reduction and clustering method used.

$^3$Only the first step of the Isomap algorithm, which calculates the geodesic "shortest path" distances in a data manifold, is used. Low-dimensional embeddings were not constructed.
sequence length), while the second is conducted by sampling variable-length sub-
sequences to evaluate the robustness of the approaches.

5.4.1 Experiments on Fixed-length Subsequences

In the first experiment, the more challenging Honda/UCSD dataset is used, randomly sampling 50 subsequences of 200 frames long from each subject in the test set, i.e. Honda/UCSD$_{50,1,20}$. To comprehensively assess with different features, four feature representations are used – PCA, LDA, NPE and NDMP. The optimal number of feature dimensions for these methods were determined empirically by trial experiments. Table 5.2 shows the average recognition rates of various exemplar extraction methods evaluated in this experiment. The Naive Bayes probabilistic classifier in Equation 4.28 is applied for the classification of test video subsequences.

The summary of recognition rates in Table 5.2 evidently indicates better selection of exemplars using spatio-temporal clustering approaches (STHAC-GF, STHAC-LP) compared to conventional spatial clustering approaches (HAC and $k$-means). The same effectiveness can be observed across different feature representations. This shows the importance of incorporating temporal information to enhance clustering of temporal-oriented data such as moving faces in video sequences. Among the two STHAC variants presented, the local variant (STHAC-LP) marginally outperforms the global variant (STHAC-GF) for most feature types, although it appears statistically insignificant. This may be attributed to the ability of CSTN in discovering the spatio-temporal rela-

<table>
<thead>
<tr>
<th>Method / Feature</th>
<th>PCA</th>
<th>LDA</th>
<th>NPE</th>
<th>NDMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random selection</td>
<td>63.7</td>
<td>64.8</td>
<td>65.7</td>
<td>66.1</td>
</tr>
<tr>
<td>PCA + $k$-means</td>
<td>65.0</td>
<td>71.2</td>
<td>66.5</td>
<td>72.2</td>
</tr>
<tr>
<td>LLE + $k$-means</td>
<td>68.5</td>
<td>70.4</td>
<td>65.4</td>
<td>73.7</td>
</tr>
<tr>
<td>Geodesic + HAC</td>
<td>73.7</td>
<td>71.3</td>
<td>66.1</td>
<td>76.8</td>
</tr>
<tr>
<td>LLE + HAC</td>
<td>66.2</td>
<td>71.2</td>
<td>70.5</td>
<td>86.2</td>
</tr>
<tr>
<td>LLE + STHAC-GF</td>
<td><strong>74.9</strong></td>
<td><strong>76.9</strong></td>
<td><strong>80.1</strong></td>
<td><strong>95.0</strong></td>
</tr>
<tr>
<td>LLE + STHAC-LP</td>
<td><strong>81.9</strong></td>
<td><strong>87.2</strong></td>
<td><strong>90.8</strong></td>
<td><strong>94.5</strong></td>
</tr>
</tbody>
</table>

Table 5.2: Average recognition rates (%) of various exemplar extraction methods with different feature representations
tionships between data points within the local neighborhood. It is worth re-iterating the advantages of NDMP as the feature choice as it clearly outperforms the PCA, LDA and NPE methods, owing much to its elegant neighborhood discriminative formulation.

For further performance testing, evaluation by rank-based identification is performed by adopting a simple probabilistic voting strategy. The normalized class probability scores (Equation 4.29) are cumulatively aggregated by simple sum rule. A vote is taken at each frame and the class with the majority vote is classified as the matched subject of the test subsequence. This strategy allows for a subject to be matched within the $K$ highest voted classes, which corresponds to rank $K$ of a $C$-person database. The Cumulative Match Characteristic (CMC) curve in Figure 5.7 reinforces the effectiveness of the STHAC methods over spatial clustering methods in the aspect of exemplar extraction, especially across the top-most ranks. With this classification approach, the global variant of STHAC remains superior over all other evaluated methods across the top 10 ranks.

![Figure 5.7: Cumulative Match Characteristic (CMC) curve of various exemplar extraction methods. Probabilistic voting strategy is applied for classification with NDMP features.](image-url)
5.4.2 Experiments on Variable-length Subsequences

Next, a more extensive experiment is provided by sampling test video subsequences of varying lengths from both CMU MoBo and Honda/UCSD datasets. This assesses the robustness of the evaluated methods, by reducing to a certain measure, the sequence length bias from evaluation. Using parameters defined by the sampling protocol, CMU MoBo \textsuperscript{12,5,20} and Honda/UCSD \textsuperscript{10,5,20} are used with the subsequence length sets for both datasets denoted respectively by $L_{\text{MoBo}} = \{11, 20, 30, 40, 50\}$ and $L_{\text{Honda}} = \{50, 100, 150, 200, 250\}$.

Table 5.3 shows the recognition rates of the evaluated exemplar extraction methods on the two datasets. NDMP features were used and classification was performed with the Naive Bayes probabilistic classifier. The STHAC methods are shown to be capable of selecting the most representative exemplars from the training data, compared to conventional spatial clustering methods. As evident from the results, the Honda/UCSD dataset poses much greater difficulty in an exemplar-based recognition setup owing to its large variety of head pose and rotations. Nonetheless, both STHAC-GF and STHAC-LP methods achieves a significant improvement over their spatial counterparts in this challenging dataset. It is worth noting that although the global variant appears to perform better than the local variant in the first experiment, the local variant tend to generalize better across sequences of varying lengths. Exploiting spatio-temporal distances at the local neighborhood level seemed to enable spatial distances to be artificially adjusted to reflect the relationship between data points in time and space.

5.4.3 Further Discussions

For both variants, the influence of spatial and temporal distances towards the formation of clusters can be tuned using various parameters, as described earlier. These parameters are chosen empirically by performing trial run experiments to determine best values. By plotting the recognition rates against different values of $\alpha$ and $M$ (Figure 5.8), it can be observed that the usage of more exemplars (8 or 9) is not necessarily
Table 5.3: Average recognition rates (%) of various exemplar extraction methods on different datasets (with variable-length subsequences)

<table>
<thead>
<tr>
<th>Method / Dataset</th>
<th>CMU MoBo_{12,5,20}</th>
<th>Honda/UCSD_{10,5,20}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random selection</td>
<td>88.4</td>
<td>43.8</td>
</tr>
<tr>
<td>PCA + k-means</td>
<td>94.2</td>
<td>45.1</td>
</tr>
<tr>
<td>LLE + k-means</td>
<td>91.8</td>
<td>48.3</td>
</tr>
<tr>
<td>Geodesic + HAC</td>
<td>95.2</td>
<td>50.9</td>
</tr>
<tr>
<td>LLE + HAC</td>
<td>91.3</td>
<td>63.8</td>
</tr>
<tr>
<td>LLE + STHAC-GF</td>
<td>94.5</td>
<td>74.1</td>
</tr>
<tr>
<td>LLE + STHAC-LP</td>
<td>95.8</td>
<td>75.4</td>
</tr>
</tbody>
</table>

beneficial and could perform worse than the HAC. Meanwhile, the effective number of exemplars (6 or 7) seemed to be substantiated by observing the "elbow" of the residual error curve or cluster merging cost (an example shown in Figure 5.5). Stronger temporal influence ($\alpha > 0.5$) also tended to produce better exemplars than strong spatial influence ($\alpha < 0.5$). This is likely in cases where certain ordering of face poses may be characteristically similar in both training and test sets, resulting in the higher recognition rates.

Figure 5.8: Performance of STHAC-GF (Honda/UCSD, variable-length subsequence experiment) using different $\alpha$ values with $M$ number of exemplars extracted.
Varying the $k$ and $w$ parameters for STHAC-LP also brings up some interesting observations (see Figure 5.9). While it remained inconclusive as to the optimum size of the respective neighborhoods ($k$ for spatial, $w$ for temporal), the value of the perturbation parameter $\lambda$ between 0.2 and 0.3 (corresponding to 20% and 30% perturbation of the original distances) clearly produces better results. Over-perturbing the distances (close to 50%) could result in a deterioration of performance. Overall, the use of local spatio-temporal neighborhoods is able to produce better selection of exemplars compared to the spatial-only HAC method.

Further analysis into the cluster distribution on the spectral plot (see Figure 5.10) reveals some insight into the selection of data points. In the HAC plot, a typical result of clustering based on spatial distances alone (whereby each cluster clearly groups data points based on closeness in proximity) is observed. In contrary, the STHAC plots are less visually clear-cut in terms of how data points are partitioned as the assignment of clusters are made with respect to both spatial and temporal distances between points. One apparent observation is that the local variant (STHAC-LP) tends to create tightly knitted groups that have a balanced contribution between spatial and temporal distances, e.g. the cluster at the north-west corner of the plot (in yel-
low) features more distinctly in the STHAC-LP while the temporal influence in the STHAC-GF is much stronger, visible from the streak of adjacently connected yellow points. There are possible strengths of the STHAC-GF in uncovering a good variety of potentially important views, such as one with the look-upwards face pose (black cluster). Interestingly, both STHAC methods identified the bottom-most small group of points as an important cluster that stands alone, unlike the HAC which merely grows that cluster to a substantial size regardless of temporal properties.

In future, more in-depth work can be afforded to investigate how spatial and temporal distances influence the outcome of clustering. Also, an analytical or systematic procedure for selecting the parameters used for STHAC-GF and STHAC-LP (Table 5.1) can be designed to eliminate the need for empirical experimentation.

5.5 Summary

In this chapter, a new spatio-temporal clustering method is presented for the selection of meaningful clusters from a video face manifold – Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC). In particular, both spatial and temporal relationships between frames are incorporated in training videos by exploiting their distances in both dimensions. Two variants of the new method were proposed.

In the first variant STHAC-GF, the normalized spatial and temporal distances between data points (which are the video frames) in manifold space were blended using a weighted method. This straightforward combination of both measures across all frames in a video is taken as a global approach to fusion. The second variant STHAC-LP introduces the perturbation of spatial and temporal distances based on a common spatio-temporal neighbor (CSTN) criterion. Spatial distances between points are perturbed to accentuate the similarities between neighboring points and dissimilarities between non-neighboring points. The experiments conducted on benchmark datasets demonstrate that promising results in an exemplar-based recognition setup can be achieved by utilizing both spatial and temporal information for the clustering of video frames in the training set.
Figure 5.10: Cluster spectral plot of the 'hide' video (from Honda/UCSD dataset) in LLE manifold space. Clusters are color-coded with exemplars for each cluster shown. Adjacent video frames are connected by lines.
Previously, novel techniques for learning a nonlinear discriminative embedding in reduced-dimension space (Chapter 4), and clustering data points of a face manifold in a spatio-temporal manner (Chapter 5) have been discussed. Both of which, dealt specifically with how meaningful features can be extracted and learnt from the face manifold in an exemplar-based recognition setup.

In this chapter, these ideas are incorporated into an original cluster-centric recognition framework, whereby features of both the image set and exemplar set sub-manifolds are engaged in a novel probabilistic classification approach that elegantly fuses together relevant feature similarities. Both sets of features are directly derived from the appearance-based clusters extracted from the face video manifold. The proposed classifier necessitates a dual-feature representation of sub-manifolds— image set sub-manifolds that are capable of characterizing set variations at the cluster level, and an exemplar set sub-manifold that encodes finer appearance-based information at the point (image) level. On a lesser scale of feature usage, an exemplar-driven classifier is also formulated prior to arriving at the highlighted dual-feature classifier. Although using only exemplar features, this classifier attempts to encode the prominence of each exemplar with respect to its exemplar class set. This concept is also included in the formulation of the dual-feature classifier. For both classifiers, the subjects in video are recognized using a Bayesian maximum-a-posteriori classifier, where a joint probability function that captures relevant dependencies between the features is formulated using similarity metrics.

By way of the two recognition frameworks established in Section 3, the preceding steps involved locating clusters from the face manifold of training videos, fol-
lowed by building the respective feature representations of the exemplar set (for both frameworks) and image set sub-manifolds (for cluster-centric framework). An extensive set of experiments on three face video datasets – Honda/UCSD, CMU MoBo and NICTA ChokePoint are performed to evaluate and compare the proposed scheme against other related approaches in terms of accuracy and efficiency.

The organization of this chapter is as follows: Section 6.1 briefly reviews literature relating to the usage of image set and exemplar set features for VFR, and also a brief look at various probabilistic classifiers for VFR. Primary motivations that contribute to the conception of the proposed method are also mentioned. In Section 6.2, the methods used for clustering and feature representation (for both image set and exemplar features) steps in the framework are briefly mentioned. Section 6.3 introduces the proposed exemplar-driven and dual-feature Bayesian classifiers in detail. Experiments are comprehensively reported in Section 6.4, highlighting the performance of the evaluated approaches in terms of recognition accuracy and computational cost. Finally, a summary of conclusions and future directions are covered in Section 6.5.

6.1 Background Work

6.1.1 Related Literature

In this section, some related approaches are reviewed; primarily where image sets, exemplar sets, or a combination of both are used as features for VFR in various forms of feature representation. In most of these methods, the topological metric space represented by the image sets and exemplar sets are often loosely referred to as subspaces, manifolds or patches. For the thematic sake of this thesis, the term sub-manifolds is used to provide a clearer illustration of these spaces being a derived subset of the larger face manifold in videos. Also, some probabilistic classifiers that have been formulated for VFR are reviewed.
6.1.1 (a) Image set-based Approaches

The notion of a video-to-video matching is typically epitomized as a recognition problem over image sets, whereby a collection of images or frames can be in the form of temporally ordered or unordered sets, consisting of whole sequences or only selected images. Image sets derived from both training and test videos are usually represented by subspaces or manifold models learned from the original vectors. Classification or matching is then performed by computing subspace distances or similarity metrics between these learned subspaces.

In an early, simplistic attempt at direct video-to-video matching, Satoh (2000) proposed the matching of two video sequences by selecting the pair of frames that are closest across the two videos. While this simplification of the problem is naturally a trivial image-to-image matching between images, it is a severely underwhelming solution that misrepresents the distance between distributions. This idea lends credence to later works that attempt to measure the similarity between two distributions of face images in a more meaningful manner.

One particular idea is to perform probabilistic modelling of the distribution of face patterns in video sequences. The probabilistic modelling of (Shakhnarovich et al., 2002) uses a single Gaussian distribution on the face space, which is highly restrictive in a complex manifold of face patterns. Arandjelović and Cipolla (2004) chose to model input videos using Gaussian mixture models (GMM) for a more holistic representation, while later extending it using kernel functions to learn probability densities confined to highly nonlinear but intrinsically low-dimensional manifolds (Arandjelović et al., 2005). In these two methods, the dissimilarity between estimated probability densities is measured by the Kullback-Leibler divergence (KLD). However, these methods are wrought with difficulty of parameter estimation under limited or highly nonlinear data. They are also likely to perform poorly when training and test sets are not strongly correlated statistically.

The Earth Mover’s Distance (EMD) metric, which is based on an average Eu-
clidean distance between sets of images from two videos was proposed for VFR by J. Li et al. (2005). Their algorithm was made more effective by applying Linear Discriminant Analysis (LDA) beforehand to increase the separability of each class, producing feature signatures which are then computed by EMD for matching. Another distance metric, Hausdorff distance (HD) was first applied to still-image face recognition by Takács (1998) to measure the similarity of faces rapidly with high tolerance towards non-rigid distortions. In short, the HD between two sets is simply the greatest of all distances from one point in one set to the closest point in the other set. The proposed variant of the HD, known as modified Hausdorff distance, introduces the notion of neighborhood function and associated penalties to further improve the metric. A recent method by S. Chen and Lovell (2010) extends the HD to be applied in feature space instead of image space, showing reasonable improvement to the overall classification performance. In another interesting approach, Cevikalp and Triggs (2010) characterized image sets by a convex geometric region (affine or convex hull) spanned by points in an affine feature space. Matching for recognition is accomplished by finding the closest convex set based on a set dissimilarity metric based on geometric distances between convex models Kernel extensions were also suggested to handle complex and nonlinear manifolds of face images.

Among the wide spectrum of similarity metrics, one of the most popular choice for VFR is principal angles, which have a long standing interest in the research community due to its simplicity in theory and efficiency in implementation. Briefly, the distance between subspaces of two image sets can be measured by the largest principal (or canonical) angle that spans the two subspaces. Mutual Subspace Method (MSM) (Yamaguchi et al., 1998), the seminal work on the principal angles for face recognition takes the cosine of the smallest principal angle, or largest canonical correlation as a similarity measure between two sets of image sequences. Subsequently, further transformations from the original feature space were proposed to gain better recognition results. Fukui and Yamaguchi (2003) introduced a Constrained Mutual Subspace Method (CMSM) by projecting the feature spaces to a constraint subspace before measuring the principal angles, while Kernel Principal Angles (KPA) by Wolf
and Shashua (2003) offers a kernel approach towards the computation of principal angles in feature space using nonlinear kernels. Further on, discriminant methods that maximize the separation between within-class and between-class image sets were proposed to obtain increased optimality in classification margins, namely Discriminative Canonical Correlations (DCC) (T. Kim et al., 2007), and more recently, a graph embedding discriminant analysis based on Grassmannian manifolds proposed by Harandi et al. (2011).

6.1.1 (b) Exemplar set-based Approaches

The use of exemplar sets for VFR are much less popular than image sets but there are a significant number of works in recent literature that stand by its advantages. Exemplar-based representations deal with the large array of face images in training videos by selecting a smaller set of appearance-specific representative face images or exemplars to summarize each subject class. This reduction is by no means disadvantageous as it naturally removes redundancy in feature representation and is computationally lightweight. In common configurations, clusters or patches are extracted from the data manifold and sample means are chosen as exemplars. Although naturally regarded as a image-to-video matching approach, it usually requires combining the matching scores between all exemplars (in training set) and multiple query frames (in test set) in order to realize a full video-to-video matching.

Numerous approaches to clustering face video data have been proposed for the purpose of exemplar selection. Notable methods include the use of radial basis function (RBF) network (Krüeger & Zhou, 2002; Zhou et al., 2003), k-means clustering (Hadid & Peitikäinen, 2004), and Hierarchical Agglomerative Clustering (HAC) (Fan & Yeung, 2006a; See & Ahmad Fauzi, 2011a). R. Wang et al. (2008) introduced Maximal Linear Patch (MLP) which partitions data on the manifold into local linear models by means of a nonlinearity degree measured by the deviation of Euclidean distances and geodesic distances in the manifold. R. Wang and Chen (2009) later explored the use of Hierarchical Divisive Clustering (HDC), which is purportedly more efficient than the bottom-up HAC since a complete construction of the hierarchy all the way
down to individual samples is not necessary. Overall, relatively promising results have been reported for exemplar set approaches although there are obvious drawbacks in terms of their over-reliance on the effectiveness of clustering face patterns and also the optimal number of exemplars selected.

Despite claims of superiority of video-to-video methods in literature, an analysis from a recent evaluation of video-to-video face verification (Poh et al., 2010) observed that the performances of both image-to-video and video-to-video matching are not significantly different and success in recognition generally depends on a few crucial factors such as choice of manifold representation, distance/similarity metrics and length of video sequence.

6.1.1 (c) Probabilistic Classifiers

In view that this chapter remains focused towards the classification task, some related works that employ probabilistic-based classifiers in various forms are highlighted as well.

Some early VFR work (K. S. Huang & Trivedi, 2002; X. Liu & Chen, 2003) successfully adopted Hidden Markov Models (HMM) to capture facial dynamics in video sequences for the purpose of recognition. In (X. Liu & Chen, 2003) which is an improvement over the basic concept in (K. S. Huang & Trivedi, 2002), each test video sequence was used to update the model parameters iteratively by applying a maximum-a-posteriori (MAP) adaptation technique. Later on, M. Kim et al. (2008) also applied HMM as the modelling paradigm for face recognition in video, but using well-defined discriminative pose features. Impressive recognition results were reported on various public datasets. Tistarelli et al. (2009) further introduced a dynamical face model based on a combination of HMMs, with its states automatically determined from the data by unsupervised clustering of facial expressions in video. Generally, the learning of temporal dynamics during the training process can be computationally demanding, while its non-scalability is an obstacle to practical usage if further optimization or iterative updating is not applied. By nature of being parametric-based, a truly optimal
parameter set is paramount for a consistent, robust recognition performance. Despite showing promising results, the severity of these disadvantages could possibly outweigh the positives.

There are methods that perform face recognition in video by modelling joint probability distributions or transitional probabilities involving spatio-temporal information. In an early landmark VFR work by Zhou et al. (2003), sequential importance sampling (SIS) algorithm is applied to model a joint probability distribution of identity and head motion for simultaneous tracking and recognition. This probabilistic approach works by integrating motion and identity information over time using the SIS algorithm. K. Lee et al. (2005) approximated a nonlinear appearance manifold as a set of linear sub-manifolds, and transition probabilities were learned to model the connectivity between sub-manifolds. In this approach, a clustering algorithm is applied to partition the training images into clusters of highly-similar face poses. Joint conditional probabilities are used to incorporate temporal dynamics within a Bayesian framework and recognition is then accomplished by the maximum likelihood estimate that yields the matched subject. More on Bayesian inference modelling, Fan et al. (2005) also proposed one that fits onto the recognition task, with the maximum likelihood estimation neatly transformed to some approximated distance measures—"Distance-from-feature-space" (DFFS) and "Distance-in-feature-space" (DIFS), in the learned sub-manifolds. The transition probability is defined by counting the actual transitions between different sub-manifolds observed during training.

In another method, Y. Zhang and Martínez (2006) proposed a weighted probabilistic approach for recognizing faces from multiple images and video sequences. Variations in facial expression and pose are handled by formulating probability terms that are weighted according to the similarity of images from the training and test sets. Their algorithm is also tailored to reject outliers by the computation of a robust covariance matrix using median absolute deviation criterion. W. Liu et al. (2006) derived a statistical formulation to the recognition problem with a probabilistic fusion model that performs temporal and spatial learning called Spatio-temporal Embedding (STE),
via the use of Bayesian keyframe learning and non-parametric discriminant embedding (NDE). An STE of a video sequence is defined as its condensed version capturing the essence of space-time characteristics of the video. Impressive results were reported on the XM2VTS face video database.

6.1.2 Motivations

With the growing popularity shown by the two category of approaches (i.e. image set features and exemplar features), the next direction in this research begs the following questions:

1. How can we fuse together both exemplar and image set features to exploit temporal dynamics in the classification of faces in videos?

2. How can we model exemplar and image set matching such that recognition capability can be improved without deterioration in its efficiency?

Approaches that utilized both image set and exemplar set features are few in literature. Most approaches only make use of one or the other, mainly for reasons of efficiency. This idea also may seem redundant and an overkill in terms of feature representation unless fusion can be accomplished with minimal increase of computation cost or with some clever formulation. One notable work by R. Wang et al. (2008) proposed a weighted manifold-manifold distance (MMD) measure involving both image set (subspace) distance and exemplar (point) distance by constructing local linear patches for recognition on the face manifold. While the fusion into a single distance measure is straightforward and efficient, the optimum setting for the compensating weight parameter is difficult to determine and does not fully leverage various similarity measures at the feature levels.

The method by R. Wang et al. (2008) builds the subspace feature from entire image sets, thus capturing only "globalized" variations that tend to generalize sequences with complex face motions. To truly attempt at incorporating temporal dy-
namics, these complex sequences ought to be characterized by modelling "localized" changes through the representation of shorter subsequences or segments throughout the video. Some earlier methods (X. Liu & Chen, 2003; Zhou et al., 2003; K. Lee et al., 2005) have shown how temporal information of face videos can be captured and represented, especially within a probabilistic-based model. Nevertheless, these methods only deal with single image (or point) features.

Motivated by the intuitiveness of probabilistic modelling, some methods ((Fan et al., 2005; Y. Zhang & Martínez, 2006; W. Liu et al., 2006)) offer some insights into how probabilities can be efficiently formulated using distance metrics or inversely, probabilistic similarities in the feature spaces. In comparison with density distribution-based techniques of characterizing the probability terms, using distances and similarity measures are much more computationally efficient and its non-parametric nature eliminates the need to learn distribution parameters or enforce additional assumptions.

6.2 VFR Setup

In this setup, a cluster-centric scheme for video-based face recognition (VFR) is introduced, whereby features of both image set and exemplar sub-manifolds are derived from extracted local clusters for use in the classification of faces. In the training step, a similar procedure is adopted, as that used in the exemplar-based setting for VFR (in Chapters 4 and 5) by grouping the face images of each video sequence into local appearance-based clusters. These local clusters offer two types of features that can be readily exploited for the recognition task – subspaces (from clusters containing image sets) and points (from exemplar images). Both subspaces and point features are represented in a dual-feature mode in both training and test sequences. Intuitively, subspaces of the clusters describe coarse set-based variations while the selected points within a cluster reflect fine appearance variations. The general notation of the variables involved have been laid out in Section 3.2.

Here, a video-to-video recognition setting can be realized by first extracting cluster subspaces which also produces the exemplars, performing relevant feature rep-
resentations, followed by a temporally-aggregated probabilistic-based classification of test video frames via the matching of the various extracted features. In this section, the selection of appropriate methods used for the two other related tasks in this VFR setup – clustering and feature representation, are described.

6.2.1 Clustering & Exemplar Extraction

With the large amount of face variations in each training video, the nonlinear dimensionality reduction method, Locally Linear Embedding (LLE) (Roweis & Saul, 2000) is applied to learn a low-dimensional embedding from its original data space. It is well known that LLE is capable of modeling the intrinsic structure of a nonlinear data manifold to seek a meaningful embedding that "unrolls" the nonlinearity of face variations such as pose, expression and illumination.

Subsequently, for each training video, the projected faces in LLE feature space are partitioned into clusters using the proposed spatio-temporal hierarchical agglomerative clustering (STHAC) algorithm (see Chapter 5), where consideration for both spatial and temporal distances tend to produce better clusters. The global spatio-temporal fusion scheme (STHAC-GF) is selected in the setup of this work. For each cluster, the face image nearest to cluster mean is selected as the exemplar. A plot previously shown in Figure 5.10 presents the face clusters and their associated exemplars of a sample training video belonging to a subject, as extracted by STHAC in LLE feature space.

6.2.2 Feature Representation

From each extracted cluster, its image set sub-manifold features can be characterized using a subspace representation spanned by all the images in the cluster where the distance between subspaces or similarity measures can be determined for matching. In this approach, each cluster \( z_i \) of the \( i \)-th set is represented as a \( d \)-dimensional linear subspace by an orthonormal basis matrix \( P_i \in \mathbb{R}^d \), obtained by the Singular Value Decomposition (SVD) of \( z_i z_i^T = P_i \tilde{\Lambda} P_i^T \), where \( \tilde{\Lambda} \) and \( P_i \) are the eigenvalue and eigenvector matrices of the \( d \) largest eigenvalues, respectively. Assume \( P_i \in \mathbb{R}^d \) and \( P_j \in \mathbb{R}^d \).
denote the orthonormal bases for two subspaces $z_i$ and $z_j$, the SVD of $P_i^T P_j \in \mathbb{R}^{d \times d}$ is given as

$$P_i^T P_j = Q_{ij} \tilde{\Lambda} Q_{ji}^T \quad s.t. \quad \tilde{\Lambda} = \text{diag}(\sigma_1, \ldots, \sigma_d)$$

(6.1)

where the singular values $\sigma_1, \ldots, \sigma_d$ are the cosines of principal angles, also commonly known as canonical correlations. This forms the basic workings of the Mutual Subspace Method (MSM) where the cosine of the smallest principal angle (or largest canonical correlation) is used as the similarity measure between two subspaces. The use of principal angles or canonical correlations are available in other popular kernelized or discriminant variants.

Kernel Principal Angles (KPA) (Wolf & Shashua, 2003), denoted in SVD terms, is given as

$$V_i^T V_j = Q_{ij} \tilde{\Lambda} Q_{ji}$$

(6.2)

where $V_i$ and $V_j$ are the orthonormal set of eigenvectors for the kernelized sets $\phi(z_i)$ and $\phi(z_j)$, mapped to a higher dimension with a kernel function $k(z_i, z_j) = \phi(z_i)^T \phi(z_j)$.

The Discriminative Canonical Correlations (DCC) proposed by T. Kim et al. (2007) is given as

$$P_i^T T T^T P_j = Q_{ij} \tilde{\Lambda} Q_{ji}$$

(6.3)

where $T$ is the discriminative matrix found by iterative optimization to maximize similarities of pairwise sets of similar classes while minimizing similarities of pairwise sets of different classes. Comparison between these methods are reported later in the comparative evaluation presented in Section 6.4.2.

For the exemplar sub-manifold feature, the earlier proposed nonlinear dimensionality reduction method called Neighborhood Discriminative Manifold Projection (NDMP) (see Chapter 4) is applied. In brief, NDMP seeks to learn an optimal low-dimensional projection by considering both intra-class and inter-class reconstruction weights. Global structural and local neighborhood constraints are imposed within a constrained optimization problem, which can be solved as the generalized eigenvalue
problem (given in Equation (4.26)). New test samples $X'$ can be projected to embedded space $\mathbb{R}^d$, by the linear transformation $Y' = A^T X'$ where $d \ll D$. A detailed theoretical formulation of NDMP is provided in Chapter 4.

Figure 6.1 summarizes the cluster-centric approach to the extraction of subspace and point features from a training video sequence through local class-specific appearance-based clusters. Note how the image set sub-manifolds capture the cluster-specific variational-based properties from each image set while the solitary discriminative exemplar sub-manifold essentially differentiates between the individual appearance-based properties from each cluster.

### 6.3 Bayesian Classifiers for VFR

Considering that the proposed classification algorithm is Bayesian in nature, the formulation of the probabilistic model is shown in the manner by which the Naive

---

**Figure 6.1**: Cluster-centric scheme for VFR illustrating the extraction of subspace and point sub-manifold features from a training video sequence. In terms of feature representation, cluster image sets are characterized by linear subspaces while the exemplar set consists of exemplar images from across all classes projected onto NDMP-space.
Bayes and Exemplar-Driven Bayes classifiers (that precede the proposed classifier) are also deliberated. Finally, the computation of the probability terms involved are elaborated at length.

6.3.1 Naive Bayes Classifier

The proposed classification approach is modelled based on the evaluation of class hypotheses set \( H = \{H_c\}, \ c = 1 \ldots C \), by Bayesian inference, where the subject identity \( c^* \) of a new sequence \( X \) can be found by estimating the maximum-a-posteriori (MAP) decision rule\(^1\):

\[
c^* = \arg \max_c p(c|X) \quad (6.4)
\]

Generally, a conventional Naive Bayes (NB) classifier can be defined by way of a MAP decision rule. Assuming conditional independence between all observations, \( i.e. \ x_i \perp\perp x_j | c \) where \( i \neq j \) and that the unobserved samples are \( i.i.d. \), the posterior distribution over the class hypotheses at time frame \( N \) at the end of test sequence \( X \) can be expressed by Bayes rule,

\[
p(c|X) \equiv p(c|x_1, \ldots, x_N)
\equiv \frac{\prod_{i=1}^{N} p(x_i|c)p(c)}{p(x_i)} \quad (6.5)
\sim p(c) \prod_{i=1}^{N} p(x_i|c) \quad (6.6)
\]

where \( p(c) \) is the class prior while \( p(x_i|c) \) is the likelihood probability of the unknown sample \( x \) of instance \( i \) given class \( c \). The sample marginal distribution (denominator term) in Equation (6.5) is a constant normalization factor that ensures that the sum of the likelihoods over all possible classes equals to one \( i.e. \ p(x) = \sum_c p(x_1, \ldots, x_N|c)p(c) \). Hence, the posterior distribution is not dependent on the normalization term (Equation (6.6)).

---

\(^1\)For the sake of clarity, the notation denoting the test sequence as \( X \) is abused, contrary to earlier notation of \( X' \) defined in the problem setting (in Section 3.2)
In order to accommodate an exemplar-based scheme (where each training subject/class is represented by a set of \( M \) exemplars), the likelihood term \( p(x|c) \) can be modified to consider the likelihood of an input frame \( x \) given the exemplar \( e \) of class \( c \). After which, the probability scores obtained from all \( M \) terms of \( p(x|c, e) \) can be aggregated by simple sum rule:

\[
p(c|X) \propto p(c) \prod_{i=1}^{N} \sum_{j=1}^{M} p(x_i|c, e_{c,j}) \quad (6.7)
\]

### 6.3.2 Exemplar-Driven Bayes Classifier

The Naive Bayes classifier can be further extended to leverage the natural properties of an exemplar-based scheme by introducing a joint probability function that considers the relationship between individual exemplars and their respective classes. This addresses the lack of attention towards modelling the causal relationship between classes and their associated exemplars within the Bayesian framework. As such, the "prominence" of exemplars within each exemplar set subspace can be quantified and exploited to provide a fair weightage to the contribution of each exemplar to the likelihood probabilities.

Briefly, the joint probability function for the Exemplar-Driven Bayes (EDB) classifier is formulated as,

\[
p(c, E, X) = p(X|c, E)p(E|c)p(c) \quad (6.8)
\]

where the \( p(X|c, E) \) term is the likelihood of video sample \( X \) given the exemplar \( E \) belonging to class \( c \), while \( p(E|c) \) encodes the exemplar prominence, or the contributive strength of exemplar \( E \) given the class \( c \).

Therefore, the MAP classifier (Equation 6.4) is redefined by maximizing the
The decision rule is realized by maximizing the joint posterior probability \( p(c, E | X) \) which determines the probability of class \( c \) and exemplar \( e \) (of the exemplar set \( E \)) given observed sample \( X \). The probability terms involved in Eq. (6.9)— exemplar likelihood \( p(x_i | c, e_{c,j}) \) and exemplar prominence \( p(e_{c,j} | c) \) can be computed efficiently by enforcing a non-parametric approach to probability distributions using relevant similarity scores. This idea will be carried over extensively into the formulation of the dual-feature classifier (in Section 6.3.3), where further details on the computation of these terms are also discussed. The graphical model depicting the dependencies in the EDB classifier is shown in Figure 6.2(a).

**Figure 6.2:** Graphical models depicting the dependencies of the proposed Exemplar-Driven Bayes (EDB) and Dual-Feature Bayes (DFB) classifiers
6.3.3 Dual-Feature Bayes Classifier

Since the cluster-centric recognition scheme aims to utilize features from both image set and exemplar sub-manifolds, a Dual-Feature Bayes (DFB) classifier is introduced to fuse the matching tasks for these features. A new joint probability function is proposed to fuse together relevant similarities within the extracted local clusters at both point and subspace levels,

\[
p(c, E, Z, \Theta, X) = p(X|E, \Theta)p(\Theta|Z)p(Z|c)p(E|c)p(c) .
\]  

(6.10)

The conditional dependencies among the point feature \((E, X)\) and subspace feature \((Z, \Theta)\) variables are encoded in the Bayesian framework. The graphical model depicting the dependencies in the DFB classifier is shown in Figure 6.2(b).

To determine the class hypothesis at the end of a test video sequence, the joint posterior probability to be maximized can be conditioned as follows:

\[
p(c|E, Z, \Theta, X)
\]

\[
= \frac{p(X|E, \Theta)p(\Theta|Z)p(Z|c)p(E|c)p(c)}{\sum_{c'} p(X|E, \Theta)p(\Theta|Z)p(Z|c')p(E|c')p(c')}
\]

\[
\propto p(c) \prod_{i=1}^{N} p(x_i|E, \Theta)p(\Theta|Z)p(Z|c)p(E|c)
\]

\[
\propto p(c) \prod_{i=1}^{N} p(x_i|E)p(x_i|\Theta)p(\Theta|Z)p(Z|c)p(E|c)
\]  

(6.11)

with the assumption of conditional independence between observations in \(X\) and law of total probability. The latent variables \(\Theta\) and \(E\) are also assumed independent of each other as they both represent different features types and have no mutual causality.

Two additional assumptions are further asserted to simplify it compactly to the final term. With the use of local clusters, each class is now represented by \(M\) number of clusters and exemplars (see Eqs. (3.2) and (3.3)). Hence, conditional probabilities involving \(\Theta\) and \(E\) can be aggregated by sum rule to maintain consistency in this framework, a somewhat similar procedure to that used in the exemplar-based schemes.
Finally, \( p(x_i|\Theta) \) and \( p(Z|c) \) are assumed non-informative and are treated as constants since they are intuitively insignificant to the decision rule. With that, the following expansion is obtained in vectorized form:

\[
p(c|E, Z, \Theta, X) \\
= p(c|e_{c,j}, z_{c,j}, \theta_i, x_i) \\
\propto p(c) \prod_{i=1}^N p(x_i|\theta_i) \sum_{j=1}^M p(\theta_i|z_{c,j})p(x_i|e_{c,j})p(e_{c,j}|c)p(z_{c,j}|c) \\
\propto p(c) \prod_{i=1}^N \sum_{j=1}^M p(\theta_i|z_{c,j})p(x_i|e_{c,j})p(e_{c,j}|c) \quad (6.12)
\]

Due to the limited sample size in the problem setting, good estimations of distribution can be challenging and easily result in over-fitting or under-fitting of data. Hence, relevant similarity metrics involving point and subspace features derived from within the local clusters are exploited, as illustrated in Figure 6.3. Concretely, these similarity metrics are computationally inexpensive and structurally intuitive for formulating the relevant probabilities. The likelihoods \( p(\theta_i|z_{c,j}) \) and \( p(x_i|e_{c,j}) \) reflect the probabilities of test point \( x \) and subspace \( \theta \) at frame \( i \) given their respective trained image sets and exemplars. Meanwhile, the conditional probability \( p(e_{c,j}|c) \) weighs the importance of each exemplar with respect to its class-wise exemplar subspace. Since there are no information on the correctness of hypothesis at the start of the observation sequence, class priors \( p(c) \) are presumed to be uniformly distributed over all the class hypotheses.

6.3.3 (a) Cluster likelihood \( p(\theta_i|z_{c,j}) \)

The cluster likelihood term describes the matching between image sets. The training and test sub-manifolds containing the image set features can be used to determine the similarity between clusters (or image sets) in the form of principal angles or canonical correlations (as discussed in Section 6.2.2). The cluster likelihood, or likelihood of the observed cluster subspace \( \theta_i \) (corresponding to face image \( x_i \)) given the
Figure 6.3: Graphical illustration of similarity metrics involving exemplar points (circles) and image set subspaces (colored regions) within local class-specific clusters. Two sample classes are shown, each with only two extracted clusters.

training cluster subspace $\mathbf{z}$ is defined as

$$p(\theta_i | \mathbf{z}_{c,j}) = (1 - \alpha)S^{CL}_i(\theta_i, \mathbf{z}_{c,j}) + \alpha$$  \hspace{1cm} (6.13)

where the normalized subspace similarity metric is defined by the average of first $r$ canonical correlations,

$$S^{CL}_i(\theta_i, \mathbf{z}_{c,j}) = \sum_{l=1}^{r} \frac{\sigma_l}{r}$$, where $r < d$  \hspace{1cm} (6.14)

In the experiments, the value of $r$ is fixed to obtain consistent correlation values. Parameter $\alpha$ is the lower bound of the similarity metric (asserting value range of $[\alpha, 1.0]$), defining its degree of sensitivity. It is not necessary to compute the observed cluster subspace $\mathbf{\Theta}_i$ at each frame instance since it can be computed once every $L$ number of image frames, resulting in a total of $M'$ times for the observed video sequence (refer to Equations 3.6 and 3.7). This amounts to a substantial saving of computational cost during classification, while allowing each $i$-th frame instance to be assigned a cluster.
6.3.3 (b) Exemplar likelihood $p(x_i|e_{c,j})$

Point-to-point matching between a training exemplar and observed face images can be represented by the exemplar likelihood term. As such, a point similarity metric is defined as the inverse squared Mahalanobis distance between the observed face image $x_i$ and the $j$-th exemplar of class $c$ in NDMP-projected embedded space,

$$S_{EL}^{EL}(x_i,e_{c,j}) = \frac{1}{\left[ (x_i - e_{c,j})\Sigma^{-1} (x_i - e_{c,j})^T \right]}$$  \hspace{1cm} (6.15)$$

where $\Sigma$ is the common covariance matrix for all class samples (inclusive of the test sample). The similarities are sum-normalized across all classes. The choice of Mahalanobis metric here is assumed to intuitively embody well-behaved multinormal sample distributions with a common covariance matrix.

The exemplar likelihood, or likelihood of the observed face image $x_i$ given the training exemplar $e$ is formulated under stochastic selection rule as,

$$p(x_i|e_{c,j}) = \frac{S_{EL}^{EL}(x_i,e_{c,j})}{\sum_{k=1}^{C} \sum_{j=1}^{M} S_{EL}^{EL}(x_i,e_{c,j})}$$ \hspace{1cm} (6.16)$$

6.3.3 (c) Exemplar prominence $p(e_{c,j}|c)$

Causal relationship between exemplars and their respective class-wise exemplar subspaces can be represented by the conditional probability $p(e_{c,j}|c)$, or exemplar prominence term. Intuitively, these conditional probabilities act as prominence weights for the exemplar likelihoods $p(x_i|e_{c,j})$, or strength of influence of the exemplar within its own subspace. A point-to-subspace similarity metric is defined by the inverse $L2$-Hausdorff distance from each exemplar $e_{c,j}$ to its corresponding class-wise exemplar subspace $E_c$ in NDMP space,

$$S_{PR}^{PR}(e_{c,j},E_c) = 1/(\min_{e \in E_c} ||e_{c,j} - e||)$$ \hspace{1cm} (6.17)$$
The $M$ similarities of each $c$-th class are normalized by min-max normalization which is a linear mapping to the range $[0, 1]$. Hence, the exemplar prominence term is formulated as

$$p(e_{c,j}|c) = \frac{S^{PR}_{c,j}(e_{c,j}, E_c)}{\sum_{j=1}^{M} S^{PR}_{c,j}(e_{c,j}, E_c)}$$

(6.18)

This term can be pre-computed during training since it does not depend on observation sample $X$.

6.3.4 Summary of Algorithms

The steps of the two proposed Bayesian classifiers—Exemplar-Driven Bayes (EDB) and Dual-Feature Bayes (DFB) classifiers are summarized here, as well as in the published works (See, Ahmad Fauzi, & Eswaran, 2011; See, 2011; See, Eswaran, & Ahmad Fauzi, 2012; See, Ahmad Fauzi, & Eswaran, 2013).

Exemplar-Driven Bayes (EDB) classifier:

1. Initialize class posterior probability, $p(c,E|X)$ (or shortened to $\mathcal{J}_c$ in Algorithm 4) to the class prior $p(c)$, which is estimated as a uniform sum distribution $U(0, 1)$ (also known as Irwin-Hall distribution).

2. Pre-compute the exemplar prominence term $p(e_{c,j}|c)$ for all $C \times M$ exemplars in $E$, before the testing stage.

3. Compute the product of the exemplar likelihood term $p(x_i|c, e_{c,j})$ and the exemplar prominence term, for each $i$-th image from the test video $X$, summing across all exemplars of each class $c$. This term should be normalized by simple min-max normalization after summation.

4. Update posterior probability $\mathcal{J}_c$ for all $C$ classes, at the current frame instance $i$, by product rule. The sum of all $C$ (normalized) posterior probabilities should be 1.

5. Assign the subject identity $c^*$ with the class $c$ that gives the maximum posterior at the end of the sequence (frame instance $N'$).
Dual-Feature Bayes (DFB) classifier:

1. Initialize class posterior probability, \( p(c, E, Z, \Theta, X) \) (or shortened to \( J_c \) in Algorithm 4) to the class prior \( p(c) \), which is estimated as a uniform sum distribution \( U(0, 1) \) (also known as Irwin-Hall distribution).

2. Pre-compute the exemplar prominence term \( p(e_{c,j}|c) \) for all \( C \times M \) exemplars in \( E \), before the testing stage.

3. Compute the cluster likelihood \( p(\theta_i|z_{c,j}) \) associated to each \( i \)-th image from the test video \( X \), for all \( C \times M \) clusters in \( Z \). This is only required to be computed once every \( M \) number of frames.

4. Compute the product of three probability terms— exemplar likelihood term \( p(x_i|e_{c,j}) \), the exemplar prominence term and the cluster likelihood term, for each \( i \)-th image from the test video \( X \), summing across all exemplars of each class \( c \). This term should be normalized by simple min-max normalization after summation.

5. Update posterior probability \( J_c \) for all \( C \) classes, at the current frame instance \( i \), by product rule. The sum of all \( C \) (normalized) posterior probabilities should be 1.

6. Assign the subject identity \( c^* \) with the class \( c \) that gives the maximum posterior at the end of the sequence (frame instance \( N' \))

Concretely, the pseudocodes of the EDB and DFB approaches are shown in Algorithms 3 and 4. It must be pointed that the careful breakdown of steps (which are not mentioned in the formulations) are done for clarity and easy following. Practically, the implementation of these classifiers can be optimized by converting most of the for-loops into a one-shot matrix multiplication or vectorized procedure.

6.4 Experiments

To conduct a thorough performance evaluation of the proposed method against other classification schemes, experiments were conducted on three public benchmark
Algorithm 3  Exemplar-Driven Bayes (EDB) Classifier

**Input:** Test video sequence, $X = \{x_1, x_2, \ldots, x_{N'}\}$. Exemplar set images, $E = \{e_{1,1}, e_{1,2}, \ldots, e_{C,M}\}$ from training.

**Assumption:** Features for $E$ have been extracted, while $X$ is properly projected to the similar feature space.

**Output:** Matched subject identity, $c^*$ of test video sequence $X$

**Class hypothesis:** $\mathcal{H} = \{c | 1, \ldots, C\}$

**Pre-Classification:**

1. Initialize class posterior $\mathcal{J}_c \leftarrow p(c) \sim \mathcal{U}(0,1)$

2. for all $e_{c,j} \in E$ do

3.   $\mathcal{G}_{c,j} \leftarrow p(e_{c,j}|c)$ \hspace{1cm} \triangleright Exemplar prominence term

4. end for

**Testing:**

5. for all $i \in X$ do

6.   for all $c \in \mathcal{H}$ do

7.     $\mathcal{J}_c \leftarrow \mathcal{J}_c * \frac{1}{N} \sum_{j=1}^{M} \mathcal{G}_{c,j} * p(x_i|c, e_{c,j})$ \hspace{1cm} \triangleright Exemplar likelihood term added

8. end for

9. end for

10. $c^* \leftarrow \arg \max_C \mathcal{J}_c$

---

Databases – CMU MoBo (Gross & Shi, 2002), Honda/UCSD (K. Lee et al., 2005) and a more recently created NICTA ChokePoint (Wong et al., 2011). Details of these datasets are elaborated in Section 3.3. For all datasets, it is assumed that face images were extracted, resized and preprocessed based on the standard protocol defined in Section 3.4.

The ChokePoint dataset consists of a complex set of recording conditions, whereby each subject recorded 24 sequences at a predefined portal (or natural choke points in terms of pedestrian traffic), consisting of a combination of 4 sequence shots, 3 camera angles, and 2 movement modes (entering and leaving portal). While it has
Algorithm 4  Dual-Feature Bayes (DFB) Classifier

Input (new test samples): Test video sequence, \( X = \{x_1, x_2, \ldots, x_{N'}\} \). Test cluster segments, \( \Theta = \{\theta_1, \ldots, \theta_{N'}\} \).

Input (training samples): Exemplar set images, \( E = \{e_{1,1}, e_{1,2}, \ldots, e_{C,M}\} \). Cluster image sets, \( Z = \{z_{c,1}, z_{c,2}, \ldots, z_{c,M}\} \).

Assumption: Features for \( E \) and \( Z \) have been extracted or pre-trained, while \( X \) and \( \Theta \) are properly projected to similar feature spaces.

Output: Matched subject identity, \( c^* \) of test video sequence \( X \)

Class hypothesis: \( \mathbb{H} = \{c|1, \ldots, C\} \)

Pre-Classification:

1: Initialize class posterior \( J_c \leftarrow p(c) \sim \mathcal{U}(0,1) \)

2: for all \( e_{c,j} \in E \) do

3: \( G_{c,j} \leftarrow p(e_{c,j}|c) \) \text{ } \triangleright \text{ Exemplar prominence term}

4: end for

Testing:

5: for all \( i \in X \) do

6: if \( (N' \text{ mod } M) = 1 \) then

7: for all \( z_{c,j} \in Z \) do

8: Update \( F_{c,j} \leftarrow p(\theta_i|z_{c,j}) \) \text{ } \triangleright \text{ Cluster likelihood term added}

9: end for

10: end if

11: for all \( c \in \mathbb{H} \) do

12: \( J_c \leftarrow J_c * \frac{1}{N} \sum_j^M F_{c,j} * G_{c,j} * p(x_i|e_{c,j}) \) \text{ } \triangleright \text{ Exemplar likelihood term added}

13: end for

14: end for

15: \( c^* \leftarrow \arg \max_c J_c \)
its own baseline verification protocol, and suggested evaluation procedures for single and multiple camera sequences (Wong et al., 2011), it is adapted to suit the recognition evaluation used here. In the experiments, two different setups were provided for the ChokePoint based on how training-test samples are partitioned: Leaving-Leaving (LL) and Entering-Leaving (EL), using only the first subset of the dataset (Portal 1) with 25 subjects\(^2\). The first is a more controlled setting that involves recognizing faces in videos captured within the same time and environment (background, lighting). The second setting is expectedly more adversed and challenging as it features two entirely different environment conditions in the training and test sets, with the videos also recorded at a time interval apart.

Based on accepted practice in most VFR works, one video sequence is used for training while the remaining video sequences are used for testing for each subject class. The description of the generated augmented test sets is covered in Section 6.4.2 where different sampling parameters \((W, T, C)\) are applied to obtain different observations. In most scenarios, the number of sampled subsequence lengths, \(T\) should be consistent across all datasets to facilitate fair comparison between all evaluated datasets. The subsequence lengths are determined by \(T\) equally distributed intervals on the original sequence length. Meanwhile, \(W\) is arbitrarily selected on the basis of creating sufficient random samples for each evaluated subject. This should have minimal affect on the overall recognition accuracy when \(W\) is sufficiently\(^3\) larger than \(T\) or similar.

### 6.4.1 Evaluated Methods

In the experiments, the performance of the proposed Dual-Feature Bayes (DFB) classifier is compared to the following classification approaches (with abbreviations in parentheses):

- Exemplar-based majority voting (MajVote) by Euclidean nearest neighbor distance in PCA, LDA and NDMP projected spaces, with vote taken at each frame

\(^2\)The two setups for the ChokePoint dataset will be subsequently referred to as ChokePoint-LL and ChokePoint-EL.

\(^3\)As rule of thumb, \(W \geq T\) is used.
- Exemplar-based probabilistic voting (ProbVote) in NDMP projected space, where the nearest neighbor distances are normalized to probabilities and aggregated cumulatively by sum rule, with vote taken at each frame

- Mutual subspace method (MSM)

- Kernel principal angles (KPA)

- Discriminative canonical correlations (DCC), with training conducted by (i) random partitioning into two class-specific sets (DCC-RandomSplit), and (ii) partitioning into $M$ class-specific cluster sets as extracted by STHAC (DCC-Clusters)

- Manifold-manifold distance (MMD)

- Bayesian MAP classifiers – Naive Bayes (NB) and Exemplar-Driven Bayes (EDB)

By loosely categorizing the evaluated methods listed above, the MajVote, ProbVote, NB and EDB are exemplar-based methods; MSM, KPA and DCC are image set-based methods; MMD and the proposed DFB approach involved the fusion of both exemplar and image set features.

The optimal parameters for each approach were mainly determined empirically through trial experiments. For the voting methods, PCA features are set to retain 95% of data energy, while for both LDA and NDMP features, the dimension is set to the maximum possible dimension, that is number of classes minus one. The exemplar training set in all Bayesian classifiers (NB/EDB/DFB) are projected to NDMP space, with feature dimension similarly set to the number of classes minus one, the intrain class and inter-class neighbors, $K = K' = M - 1$, and the tuning parameter $\beta = 0.75$. The NDMP algorithm is the primary choice for dimensionality reduction due to its capability in extracting meaningful discriminative features from a highly nonlinear training set manifold with complex facial variations.
For methods that require clustering during training (voting methods, DCC-Clusters, NB, EDB, DFB), the number of clusters per subject were determined heuristically from the residual error curve of clustering distance criterion (details can be found in Section 5.2.4). The cluster sizes $M = 7$ for Honda/UCSD and $M = 6$ for CMU MoBo and ChokePoint are used. In Figure 6.4, sample exemplar sets extracted from training videos of all the datasets using STHAC algorithm show the distinct variability of face appearances for each subject.

For the image set-based approaches (MSM/KPA/DCC), the dimensionality of PCA subspaces learned for each image set sub-manifold is set to 10, representing about 98% of data energy of the set. The kernel function used in KPA is a Gaussian radial basis function (RBF) kernel, a promising choice suggested in the original authors’ implementation (Wolf & Shashua, 2003). For MMD, the weighting parameter is set to 0.5 for equal weights, as suggested by R. Wang et al. (2008). For the proposed DFB method which also utilizes image set features, discriminative features based on the DCC approach are chosen to represent the sub-manifolds spanned by the extracted clusters. Other parameters include the temporal segment length $L$ and subspace similarity sensitivity $\alpha$, which are both set to good empirical values of $L = 20$ and $\alpha = 0.75$.

Figure 6.4: Selected exemplars of two subjects from CMU Mobo (top two rows), Honda/UCSD (bottom two rows) and ChokePoint (middle two rows) datasets.
A thorough performance evaluation was conducted to benchmark the proposed algorithm against other related approaches on the basis of accuracy (recognition rate) and efficiency (computational cost). Experimental results in terms of recognition accuracy and computational cost are discussed and further analyzed in Sections 6.4.2 and 6.4.3 respectively.

### 6.4.2 Recognition Accuracy

#### 6.4.2 (a) Experiment I: Comparative Evaluation

In the initial experiment, the augmented test set is generated with the following sampling parameters, in \((W,T,C)\): CMU MoBo\(_{12,5,20}\), Honda/UCSD\(_{10,5,20}\)^4 and ChokePoint-LL\(_{9,5,25}\). The number of subsequence lengths sampled is fixed at \(T = 5\), with the subsequence lengths for each dataset chosen as shown in Table 6.1.

The accuracy performance of the evaluated approaches (as listed in Section 6.4.1) on the three datasets and the average, measured in terms of recognition rate (%), is shown in Figure 6.5. Clearly, the proposed DFB approach (rightmost bar in each grouping) obtained the best recognition rate among the evaluated approaches, for all three datasets. Its effectiveness is most prominent in the challenging Honda/UCSD dataset compared to other datasets, while it is also distinctly superior (92.23%) when results are averaged across all datasets (see Table 6.2). The performance of the image set-based approaches are noticeably poor in most of the datasets, possibly due to their inadequacy in generalizing complex and rapidly changing head poses belonging.

---

[^4]: In actual experiments, \(C = 19\) for test set due to one missing subject from the testing videos of the first dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(W)</th>
<th>(T)</th>
<th>(C)</th>
<th>Subsequence lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU MoBo</td>
<td>12</td>
<td>5</td>
<td>20</td>
<td>{10,20,30,40,50}</td>
</tr>
<tr>
<td>Honda/UCSD</td>
<td>10</td>
<td>5</td>
<td>20</td>
<td>{50,100,150,200,250}</td>
</tr>
<tr>
<td>ChokePoint-LL</td>
<td>9</td>
<td>5</td>
<td>25</td>
<td>{20,40,60,80,100}</td>
</tr>
</tbody>
</table>
Table 6.2: Average recognition rates (%) and standard deviation of the evaluated methods across all three datasets for Experiment I

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average Recognition Rates (%)</th>
<th>Std. deviation (across datasets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MajVote-PCA</td>
<td>72.03</td>
<td>11.4</td>
</tr>
<tr>
<td>MajVote-LDA</td>
<td>81.73</td>
<td>15.3</td>
</tr>
<tr>
<td>MajVote-NDMP</td>
<td>85.37</td>
<td>14.5</td>
</tr>
<tr>
<td>ProbVote</td>
<td>85.50</td>
<td>13.1</td>
</tr>
<tr>
<td>MSM</td>
<td>71.87</td>
<td>11.0</td>
</tr>
<tr>
<td>KPA</td>
<td>71.87</td>
<td>5.2</td>
</tr>
<tr>
<td>DCC-RandomSplit</td>
<td>72.30</td>
<td>9.4</td>
</tr>
<tr>
<td>DCC-Clusters</td>
<td>82.23</td>
<td>7.1</td>
</tr>
<tr>
<td>NB</td>
<td>87.63</td>
<td>10.3</td>
</tr>
<tr>
<td>EDB</td>
<td><strong>88.60</strong></td>
<td><strong>9.7</strong></td>
</tr>
<tr>
<td>MMD</td>
<td>88.63</td>
<td>9.0</td>
</tr>
<tr>
<td>DFB</td>
<td><strong>92.23</strong></td>
<td><strong>5.8</strong></td>
</tr>
</tbody>
</table>

to different subjects. It can be observed that approaches that employ the described Bayesian framework (NB, EDB, DFB) yielded better performance than conventional voting schemes. This is again most obvious in the difficult Honda/UCSD dataset.

Interestingly, the feature fusion approaches (MMD, DFB) seemed to produce some promising results. In the case of DFB, the incorporation of cluster-centric submanifold features (point and subspace) derived from the original video manifold, coupled with a temporally-driven probabilistic classification framework, provides a holistic representation that demonstrated a marked improvement over other related approaches. The standard deviation of the evaluated methods across the datasets shown in Table 6.2 also demonstrates the reliability and robustness of the DFB method compared to the rest.

To assess the general reliability of the evaluated methods in a one-to-many rank-based identification setting, the Cumulative Match Characteristic (CMC) curves based on all three datasets are shown in Figure 6.9. The proposed DFB method is clearly superior in the CMU MoBo and ChokePoint-LL datasets where it topped the other methods across most ranks. In the Honda/UCSD dataset, the DFB also demonstrated its reliability through the smaller ranks (1-6) but is eventually surpassed by the
Figure 6.5: Recognition rates of various classification methods on different datasets for Experiment I. The last group of bars indicate the average recognition rates across all three datasets.

MMD method in the larger ranks.

6.4.2 (b) Experiment II: On Subsequence Lengths

The second experiment centers upon the investigation of the effect of subsequence lengths on the performance of the classification methods. The accuracy performance of the proposed DFB method against selected methods from the evaluation list are compared. For that, the following sampling parameters, in \((W, T, C)\), were used to generate the augmented test set for this experiment: Honda/UCSD\(_{10,10,20}\), ChokePoint-LL\(_{9,9,25}\), and ChokePoint-EL\(_{9,9,25}\). For the ChokePoint dataset, two configurations – Leaving-Leaving (LL) and Entering-Leaving (EL) are applied; the second in particular, matches two sets of videos with entirely different environment conditions and camera angles, thus providing an ideal test for robustness in handling highly variable appearances. The CMU MoBo dataset is omitted in this experiment as many of its original sequences are short and that will not provide significant differences in the subsequence lengths as a larger number of subsequences \((T)\) is required for sampling. The choice of subsequence lengths for these datasets are shown in Table
Generally, the performance of the Bayesian MAP classifiers are superior to that of other selected methods, with the proposed DFB method still achieving the best recognition performance across all datasets, as shown in Table 6.4. Unlike crude voting strategies or rigid subspace-based methods, temporal dependencies between frames are well-exploited in a Bayesian framework. Furthermore, the dual-feature representation of cluster image sets and exemplar sub-manifolds ensures that coarser set variations and finer appearance cues both contribute towards the classification decision. This is evident from its ability to address the shortcomings of its exemplar-based counterpart.

In terms of confidence intervals of the ChokePoint datasets (LL and EL), the DFB and EDB both reported low standard deviations in the range of $0.7 - 1.3$, as compared to the MMD ($1.2 - 1.5$), subspace-based methods ($1.9 - 3.5$) and the majority voting method ($2.7 - 3.4$), which adds to its robustness towards different training-test set partitions. Like before, there was no confidence interval reported for the Honda/UCSD dataset as only a single fold of the experiment was evaluated (see Section 3.3.2).

It can be observed that the ChokePoint-EL setting appears to be more challenging than the ChokePoint-LL setting as noticeable from the recognition rates obtained under both settings. This is much expected as the Entering-Leaving (EL) mode struggles to match faces in adversely different environments, where a variety of new face poses are previously not encountered in the training set. It remains a viable research direction in the future to address issues concerning the effectiveness of recognition in practical real-world scenarios.

Table 6.3: Parameter settings and subsequence lengths used in the generation of augmented test set for Experiment II

<table>
<thead>
<tr>
<th>Dataset</th>
<th>W</th>
<th>T</th>
<th>C</th>
<th>Subsequence lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honda/UCSD</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>{25,50,75,100,125,150,175,200,225,250}</td>
</tr>
<tr>
<td>ChokePoint-LL</td>
<td>9</td>
<td>9</td>
<td>25</td>
<td>{20,30,40,50,60,70,80,90,100}</td>
</tr>
<tr>
<td>ChokePoint-EL</td>
<td>9</td>
<td>9</td>
<td>25</td>
<td>{20,30,40,50,60,70,80,90,100}</td>
</tr>
</tbody>
</table>
Table 6.4: Average recognition rates (%) of selected evaluated methods for Experiment II, with larger variety of subsequence lengths (larger $T$ values)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Honda/UCSD (10,10,20)</th>
<th>ChokePoint-LL (9,9,25)</th>
<th>ChokePoint-EL (9,9,25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Voting-NDMP</td>
<td>62.9</td>
<td>73.2</td>
<td>78.6</td>
</tr>
<tr>
<td>MSM</td>
<td>63.2</td>
<td>67.9</td>
<td>67.7</td>
</tr>
<tr>
<td>DCC-RandomSplit</td>
<td>64.3</td>
<td>73.0</td>
<td>73.5</td>
</tr>
<tr>
<td>DCC-Clusters</td>
<td>73.7</td>
<td>85.5</td>
<td>83.5</td>
</tr>
<tr>
<td>MMD</td>
<td>76.0</td>
<td>87.7</td>
<td>80.1</td>
</tr>
<tr>
<td>EDB</td>
<td><strong>75.7</strong></td>
<td><strong>92.8</strong></td>
<td><strong>78.8</strong></td>
</tr>
<tr>
<td>DFB</td>
<td><strong>84.5</strong></td>
<td><strong>93.4</strong></td>
<td><strong>84.7</strong></td>
</tr>
</tbody>
</table>

By further studying the implications of classification on different video subsequence lengths, some interesting observations can be made. The proposed DFB method seemed to gradually outperform the other methods as the subsequence length increases, albeit only registering mediocre results with shorter subsequences (as shown in Figure 6.6). This is expectedly characteristic of this Bayesian framework where temporal accumulation of frames can lead to better convergence towards a particular candidate subject. It is also worth noting that while it may appear advantageous for the performance of image set-based approaches to be highly uncorrelated with the subsequence lengths, they still not able to achieve recognition rates as high as the Bayesian methods. Since the variability of faces increase in longer videos, characterizing entire videos as image set subspaces severely limits the ability to capture finer individual appearance information. On the contrary, exemplar-based methods suffer from the rigidity of frame-wise classification, ignoring the usefulness of representing set sub-manifolds at a coarser granularity.

### 6.4.3 Computational Cost

The computational cost of the evaluated methods is assessed by measuring the time taken to perform feature modeling and classification/matching per sequence. For a more meaningful depiction of the cost for processing each subsequence, the metric
Figure 6.6: Performance of selected methods in Experiment II based on different test subsequence lengths

of seconds-per-sequence is preferred as opposed to frames-per-second which is typical convention for video frame rates. Thus, the measured time (in seconds) is averaged across all $T$ number of subsequence lengths to provide a fair basis for comparison. The machine in use is a Pentium IV 2.8 GigaHertz (GHz) running on 1.5 GigaBytes (GB) RAM.

Table 6.5 summarizes the computational cost of selected methods in terms of average feature modeling time and classification time based on the Honda/UCSD (10, 5, 20). A reasonably quick average classification time of 0.103s per sequence (average
subsequence length of 150 frames), or a rate of about 10 sequences per second was reported for the proposed DFB method. In comparison, single feature methods obtained mixed performances i.e. exemplar-based methods – NB: (0.078s), EDB: (0.093s), and image-set-based methods – MSM (0.163s), DCC-RandomSplit (0.236s), DCC-Clusters (0.755s). As expected, the exemplar-based methods are faster than image-set-based methods as straightforward matching is performed on a frame-by-frame basis.

The DFB method also fared better than the MMD (0.256s) in terms of classification speed. This is a strong implication that the proposed design of a Bayesian classification framework for recognition remains a key factor behind its efficiency. The computation of the similarity metrics involved is mostly inexpensive and does not incur much computational overhead, except for the cluster likelihood, where its subspace similarity metric (Eq. (6.14)) is the costliest term. This term is computed only once every \( L \) (i.e. size of test cluster segment) number of frames, but more importantly, it can be pre-computed prior to the classification matching process. Hence, this slight burden of computation is considered as part of the training stage where the DFB takes 83.987s, or about \( 10^2 \) times the duration needed to train the EDB (0.813s). A 4% improvement in recognition accuracy is achieved at the expense of this increase in modeling complexity. To put it in perspective, the modeling time required by the DFB method is only marginally longer than that of the DCC-Clusters approach (77.328s).

More importantly, classification speed among the Bayesian methods remained insignificantly different, and this bodes well for the DFB method as it can achieved

<table>
<thead>
<tr>
<th>Classification Method</th>
<th>Modeling Time (s)</th>
<th>Classification/Matching Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSM</td>
<td>72.483</td>
<td>0.1625</td>
</tr>
<tr>
<td>DCC-RandomSplit</td>
<td>39.218</td>
<td>0.2355</td>
</tr>
<tr>
<td>DCC-Clusters</td>
<td>77.328</td>
<td>0.7551</td>
</tr>
<tr>
<td>MMD</td>
<td>0.410</td>
<td>0.2560</td>
</tr>
<tr>
<td>NB</td>
<td>0.782</td>
<td>0.0782</td>
</tr>
<tr>
<td>EDB</td>
<td><strong>0.813</strong></td>
<td><strong>0.0937</strong></td>
</tr>
<tr>
<td>DFB</td>
<td><strong>83.987</strong></td>
<td><strong>0.1034</strong></td>
</tr>
</tbody>
</table>

Table 6.5: Computational cost of selected methods, showing modeling and classification time averaged across all subsequences with Honda/UCSD (10, 5, 20)
a far better recognition accuracy. Ultimately, it is worth noting that efficiency in the classification step (in testing stage) is essential for deployment in practical real-world applications.

6.4.4 Further Discussions

Little was discussed in the experiments on the effects of introducing exemplar prominence to the Bayesian classification framework. Both the EDB and DFB incorporated the exemplar prominence term to exploit the causal relationship between classes and their associated exemplars, as a means of weighting the contribution of each exemplar towards the likelihood probabilities.

A posterior plot can be used to profile or keep track of changes in the posterior probabilities across time. Figure 6.7 shows a sample posterior plot of a 100-frame test subsequence from the Honda/UCSD dataset on the EDB approach, showing the posterior probabilities of the seven most probable subjects. This figure also demonstrates

![Figure 6.7: Posterior plot of a subsequence sampled from the 'rakesh' video of the Honda/UCSD dataset. Posterior probabilities of the seven most probable subjects are shown in different colors. The subject (in blue line) is correctly identified at the end of the subsequence.](image)
the capability of the proposed Bayesian classification framework in arriving at the correct subject class, even when the initial frames (frame number 1-15) were incorrectly classified.

By comparing the posterior probabilities of the EDB and NB classifiers across frames on a same test sequence (in Figure 6.8), it can be observed that the EDB classifier is able to achieve a slightly higher posterior probability during classification. The additional exemplar prominence term is leveraged to decrease its susceptibility to sudden drops of confidence level, when there are other potential candidates subjects arise during classification of faces within the sequence.

6.5 Summary

In this chapter, a novel dual-feature probabilistic classification approach that gracefully fuses together cluster-centric sub-manifolds that are represented using image set and exemplar set features is introduced. Upon extraction of the appearance-based clusters from the face video manifold, the sub-manifold features at the cluster (image set) and point (exemplar image) levels are then characterized using selected state-of-the-art methods. Intuitively, the image set sub-manifolds represent group vari-
ability at the cluster level while the exemplar set sub-manifold encodes individual appearance information at the point level. A dual-feature Bayesian (DFB) maximum-a-posteriori (MAP) classifier is proposed for the recognition of subjects in video sequences, where a joint probability function that captures relevant dependencies between the features is formulated using similarity metrics that are computationally efficient. An extensive set of experiments conducted on benchmark video face datasets have shown the promising potential of the proposed classification approach in comparison with existing schemes. The proposed DFB method came out tops against most of the other evaluated methods in terms of recognition accuracy while maintaining a feasible level of computational efficiency.

For future directions, further exploration into more recent state-of-the-art feature representations (especially methods for image set representations (Cevikalp & Triggs, 2010; Harandi et al., 2011)) can be leveraged for use in the proposed classification framework to increase the upper bound of classifier performance. It also remains a definitely possibility to apply this framework to various applications that demands temporally-driven data. On the issue of robustness, further tests can be conducted to test the capability of the proposed classifiers (EDB, DFB) in dealing with real-world scenarios such as multiple identities in a sequence, and degraded low quality videos.
Figure 6.9: Comparison of cumulative match characteristic (CMC) curves of various classification methods evaluated in Experiment II on the three datasets.
CHAPTER 7

CONCLUSIONS

This thesis has set out to explore the possibility of formulating algorithms for video-based face recognition (VFR) that integrates spatio-temporal characteristics within a meaningful manifold representation for face videos. Specifically, the thesis addresses four distinct areas in VFR: (1) Manifold representation for faces in video, (2) integration of spatio-temporal information in video, (3) utilization of features derived from cluster-centric sub-manifolds, (4) experimental protocol for evaluation sampling in VFR.

The problems encountered in these areas may appear trivial in the face of promising results reported by various approaches in literature, but owing to several key motivations in the aspect of computer science (both theoretically and experimentally), psychology and cognitive neuroscience, it appears that the aforementioned issues at large deserves a good measure of attention. Also, the existing literature surveyed in Chapter 2 collectively suggests that there is still considerable room for improving video-based face recognition in a variety of aspects.

In view of the advantages offered by clustering video data, the primary ethos of this work focuses on harnessing the rich set of features readily available in video data. Two VFR frameworks—exemplar-based and cluster-centric frameworks, are proposed. The distinction showcased by these frameworks involved the integration of spatio-temporal characteristics into the clustering and classification tasks of the VFR pipeline, and the utilization of manifold representation to characterize complex facial variations in video. The exemplar-based framework uses a single feature, namely exemplar images (points in space) to represent videos in training. The novel cluster-centric framework uses two different features extracted from each cluster to encourage extensive representation of both variational (image set subspace) and appearance (exemplar im-
age points) features. There are obvious compensating advantages and disadvantages of both frameworks: The cluster-centric framework is naturally more cumbersome in computation due to the tedious extraction of features though overall performance is robust and superior; the exemplar-based framework is more lightweight but recognition ability tends to be less impressive.

The proposed algorithms have also demonstrated superior performance in comparison to existing approaches used in various tasks of the VFR pipeline, as shown in experimental results reported in earlier chapters concerning each task. Concluding remarks and future extensions are covered for each of the pipeline tasks.

**Feature extraction and representation**

The proposed Neighborhood Discriminative Manifold Projection (NDMP) algorithm described in Chapter 4 clearly achieved a far better recognition rate than other related methods across two evaluated datasets. On a fixed-length-sequence evaluation (to maintain uniformity and remove bias from sequence lengths), it achieved an average recognition rate of 97.7% on the CMU MoBo dataset, while managing a perfect recognition score (100%) within the top 3 matches (rank-3). The superiority demonstrated by the NDMP algorithm stresses the essentiality of preserving the global manifold structure while providing discriminatory power between local neighborhood structures. Other compared methods, in particular global projection methods, appear to struggle with an innate incapability of learning highly nonlinear complex manifolds. All listed errors have been corrected.

There are several possible extensions to further improve this algorithm. The NDMP algorithm can potentially underfit data (over-generalization) especially when the dimension of features are less than the number of samples. This is naturally expected from algorithms that are founded upon the basic mechanism of LLE since it generally uncovers an intrinsically low-dimensional approximation of a data manifold. Despite that, the reduction of image samples in a video through the use of exemplar images have helped to mitigate this issue. Hence, in order to cover an even larger variety
of facial views (by increasing the number of exemplars used), it is therefore necessary
to increase the dimensionality of the features used. A straightforward extension to
this would be to map the samples to a kernel feature space (via "kernel trick" to apply
implicit mapping), or to derive kernel matrices from a sparsely weighted graph. The
formulation of NDMP itself is already fundamentally well-constructed and this can po-
tentially increase its robustness further. On the practical side, the NDMP algorithm can
be further tested with large-scale face video data or real-world video footages in order
to investigate its reliability under extremely variable conditions. As a generic feature
representation method, the NDMP can also be applied to numerous object recognition
or person identification problems in computer vision research.

Clustering

In the VFR frameworks, clustering is an important first step towards formation
of meaningful cluster sub-manifolds from training videos. This addresses the issue of
data redundancy (directly adjacent frames may contain variations that are too subtle)
and provides an elegant configuration for achieving a video-to-video recognition task.
The proposed Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC) al-
gorithm described in Chapter 5 combines the benefits of "bottom-up" agglomerative
clustering which overcomes the need for seed initialization and distribution modeling,
and the extension of each image as a spatio-temporal data point, where spatial and
temporal distances can be normalized and neatly fused into various measures. Two
variants were proposed– a global fusion variant (STHAC-GF) and a local perturba-
tion variant (STHAC-LP), both of which, seek to involve the spatial and temporal
distances between points by means of global metric aggregation and local manipu-
lation based on membership in spatio-temporal neighborhoods. Fixing the methods
used in all other pipeline tasks, both STHAC variants are able to produce impressive
recognition results in comparison to other spatial-based clustering algorithms in ex-
periments on both fixed-length and variable-length subsequences. Most significantly,
STHAC methods underline their capability in uncovering the most relevant exemplars
in training across different feature representations used (PCA, LDA, NPE, NDMP).
There is little to choose between both variants, although the local variant appears to be
marginally better than its global counterpart, especially across sequences of varying lengths.

While the notion of spatio-temporal clustering may be rather new with very few known related works, there are several extensions to this work. A more in-depth study into how spatial and temporal distances influence the outcome of clustering can be conducted. Presently, the initial proposition to model the distance between points (image samples) with a single interwoven space-time metric is a simplistic effort despite it proving more beneficial than conventional spatial distances. In this work, temporal distances are simply assumed to be linearly distributed across time in relation to spatial distances (since normalization of both distances are required). Is this assumption valid? Also, how can these two distances be well balanced? Is a single unit in time equivalent to a single unit in space?

Another future point worth investigating concerns the effect of outliers in training data. Often, outliers can cause erroneous clustering of points or creation of small clusters that are located far from the rest. This is undesirable in the context of the recognition task as it is vital for the cluster sub-manifolds to contain faces that are closely similar. One likely solution is to eliminate these outliers with a pre-processing step that tests for the fidelity or relevancy of these face images, or during the clustering process itself.

Practically, an analytical or systematic procedure for selecting the parameters used in STHAC-GF and STHAC-LP can be proposed to eliminate the need for empirical experimentation. Currently, the number of clusters are also chosen heuristically based on the "elbow" of the residual error plot for cluster merging, while the other parameters are selected by experiments. This methodology is tedious and also prone to inaccuracies.
Classification

The proposed novel probabilistic classification approaches described in Chapter 6 elegantly fuse together relevant feature similarities based on Bayesian maximum-a-posteriori (MAP) estimation. The two aforementioned VFR frameworks utilize different sets of features. For the cluster-centric framework, the Dual-Feature Bayes (DFB) classifier considers both features of the exemplar set (point-based) and image set (subspace-based) sub-manifolds, both of which, are derived from the previously extracted clusters. Meanwhile, the Exemplar-Driven Bayes (EDB) classifier uses only the face exemplar features. In formulation, the EDB is modeled after a typical Naive Bayes classifier, enhanced by way of introducing an exemplar prominence measure that captures the causal relationship between each exemplar and its associated class-specific exemplar set. The DFB in turn, further extends the EDB by incorporating a new set of subspace-based features to allow coarser variational information from each sub-manifold to be modeled alongside finer facial appearance information from image features. Relevant similarity metrics that are computationally inexpensive and structurally intuitive are designed to provide a good estimation of the distribution of data constrained by small sample size problem.

In two comprehensive experiments involving three datasets and ten other methods, the proposed classifiers obtained some promising results, with the DFB performing exceptionally well. As the video subsequence length increases, the DFB gradually overtook the other (exemplar-based and image set-based) methods while the EDB only managed a mixed bag of results, thus highlighting the essentiality of both types of features for classification. Nevertheless, this leap in performance comes at a cost—training the DFB takes about 100 times the time needed for the EDB, although the difference in classification time between both methods is negligible.

There are many potential avenues for further work. Further adoption of newer feature representation methods (particularly for image set representations) such as affine/convex hull (Cevikalp & Triggs, 2010; S. Chen, Wiliem, Sanderson, & Lovell, 2014), graph embedding on Grassmannian manifolds (Harandi et al., 2011) may be
able to increase the upper bound of classifier performance. In this work, the choice of DCC for subspace features serves as a sample candidate to advocate the proposed dual-feature classifier, and also for the purpose of experimental comparisons (Note that the DCC algorithm is also one of the compared methods). Finding a feasible choice for this also requires consideration into the computational cost and type of application it is deployed for. This classification method can also be extended to various applications that process temporally-driven data. On the issue of robustness, extended experiments can be conducted to test the capability of DFB in dealing with real-world scenarios such as multiple identities in a video involving sudden change of persons, and degraded poor quality videos.

**Augmented test set generation protocol**

Many existing VFR approaches offer little clarity into their evaluation methodologies and experiments are often performed in an unbalanced and skewed manner. The bias that indirectly pervades into VFR experiments are veered towards either extremes— conducting "selective experimentation" whereby only selected videos or frames are tested, or "generalized experimentation", whereby evaluation is performed on entire video sequences, producing high accuracies under unrealistic conditions. In the quest towards balanced and realistic experimentation, an augmented test set generation protocol is devised to provide a large test set comprising of a variety of video subsequences of different lengths and different persons. These subsequences are also randomly sampled from different starting frame positions, to mimic realistic scenarios containing arbitrary facial views. This is by no means a "gold standard" for VFR evaluation, but it is with hope that future VFR research will adopt this protocol. For future considerations, this can be extended to a more challenging protocol that includes sampling videos of different quality (although this may be subjected to pre-labeling), or merging of videos containing different persons to construct composite subsequences.

* * * * *
Video-based face recognition presents some challenging problems, particularly in the aspect of feature extraction and representation, integration of temporal dynamics, and other experimental issues. This thesis has presented several novelties that help address these problems. The problem statements and research questions set up at the beginning of this thesis have been thoroughly investigated in detail and viable solutions have been proposed, supported both theoretically and experimentally. New algorithms are proposed for three important sub-tasks within the applied frameworks. Quantitatively, the performance of the proposed approaches in terms of recognition accuracy has consistently shown a marked improvement over other existing methods; the best approach being the combination of all three methods (STHAC & NDMP & DFB) on the cluster-centric framework. Also, classification speed of the best approach is competitive with only a minor concern over the longer training time. In qualitative terms, these approaches are well supported by theoretical formulation and careful exploitation of the strengths of current state-of-the-art methods, while the newly introduced framework opens up possibilities for further advances in this area. At this juncture, there remains other unresolved issues in various problems outside the scope of this work. More experiments are required to find better solutions to the VFR problem altogether. Lastly, this thesis has shown that the usage of spatio-temporal information, coupled with manifold representation of video data, is a promising principle towards effective video-based face recognition.
REFERENCES


Chen, S., & Lovell, B. (2010). Feature space hausdorff distance for face recognition. In *Proc. of*


application to dual-subspace based classification. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 1384–1390).


of IEEE Int. Conf. on Image Processing (ICIP) (pp. 3321–3324).


PUBLICATION LIST

Journal Articles


Conference Proceedings


