Department of Computing

A Framework of Face Recognition with Set of Testing Images

Ke Fan

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

December, 2014

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

Ke Fan	Date

Abstract

Face recognition is one of the most important applications in computer vision. It aims to automatically identify or verify a person using a still image or a video sequence. The main difficulty in this application is the variations of a face in pose, illumination, and expression as these factors will significantly impact the recognition performances. Conventional methods based on single testing image are often insufficient to achieve satisfactory performances due to a fact that the identity features from a single image are hard to overcome those variations. To extract more effective features, Face Recognition based on Image Set (FRIS) has been proposed by using a set of testing images to improve the recognition performance. In FRIS, more images would provide more information for the same person on different conditions. It is a more feasible way to extract identity feature out of variations.

There are two main existing categories in this field: sample based methods and structure based methods. The sample based methods would define the similarity based on differences among a small group of samples in sets. Large variation difference is a big problem for this category of approaches as two sets of the same identity images under different conditions may have large differences. Structure based methods represent each set as a model and measure the difference between models. In this case, good representation needs a wide range of images for one person to reveal the structure of someone's face manifold. That requirement is hard to be fully met under real applications. Along with existing methods, we introduce a novel framework to improve the performance of current methods. The ultimate goal of this thesis is to improve the performance in FRIS significantly by relieving the pose variations of images with a face normalizer to convert facial images in different poses into a frontal standard face. To achieve this goal, we divide the whole process into three stages: image alignment, face normalization, and feature extraction for recognition.

The manifold alignment technique is firstly introduced in the thesis. This topic has its own independent interest in computer vision though we mainly use it for face recognition. We propose two new unsupervised algorithms for the automatic alignment of two manifolds of different datasets with possibly different dimensionalities. Alignment is performed automatically without any prior information on the correspondences between two manifolds. The first proposed algorithm automatically establishes an initial set of sparse correspondences between the two datasets by matching their underlying manifold structures. Local histogram features are extracted at each point of the manifolds and matched using a robust algorithm to find the initial correspondences. Based on these sparse correspondences, an

embedding space is estimated where the distance between the two manifolds is minimized while maximally retaining the original structure of the manifolds. The problem is formulated as a generalized eigenvalue problem and solved efficiently. Dense correspondences are then established between the two manifolds and the process is iteratively implemented until the two manifolds are correctly aligned. The alignment consequently reveals their joint structure. Next we introduce another improved method by releasing the restriction of data overlapping and give a more elegant solution for alignment problem. The alignment process is achieved by iteratively increasing the sparsity of the correspondence matrix until the two manifolds are correctly aligned and consequently one can reveal their joint structure.

The second step is the face normalization with an aim to bridging the gap of the large pose variations in face recognition task. In this thesis, the problem is addressed by directly transforming an image with non-frontal pose into the frontal view image. Then the recognition is performed on the transformed frontal images. For such a purpose, we first estimate a rough head pose of the input image and then use a group of Gaussian Processes Regression (GPR) models to normalize such pose into frontal view. The GPR models are learned independently for different poses. A final joint output estimation is the product of the output Gaussian distributions.

In this thesis, a dimensionality reduction method is also proposed for image set based face recognition. This algorithm transforms each image set into a convex hull and uses Support Vector Machine (SVM) to compute margins between each pair sets. Then we use Principal Component Analysis (PCA) on the margin directions by applying dimension reduction with an aim to preserve these margins. Classification can be achieved by distance based on metric of convex hulls in low dimension feature space.

This thesis demonstrates the effectiveness of the proposed methods on different public-available datasets. Experiments are conducted on different individual stage and the whole framework in terms of alignment accuracy, computational time and face recognition accuracy. Results show that the proposed framework outperforms existing approaches significantly, particularly for the case of two image sets without pose overlap.

Contents

List of Figures

List of Tables

List of Algorithms

Publications

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

- Xiaoming Chen, Wanquan Liu, Jian-Huang Lai, Ke Fan: Feature Extraction via Balanced Average Neighborhood Margin Maximization. The proceeding of the International Conference on Neural Information Processing (ICONIP) (2) 2011: 109-116
- Ke Fan, Wanquan Liu, Senjian An, Xiaoming Chen: Margin Preserving Projection for Image Set Based Face Recognition. The proceeding of the International Conference on Neural Information Processing (ICONIP) (2) 2011: 681-689
- Ke Fan, Ajmal S. Mian, Wanquan Liu, Ling Li: Unsupervised iterative manifold alignment via local feature histograms. IEEE Winter Conference on Applications of Computer Vision (WACV) 2014: 572-579
- Xiaoming Chen, Ke Fan, Wanquan Liu, Xin Zhang, Mingliang Xue: Discriminative Structure Discovery for Dimensionality Reduction of Facial Image Manifold. (to appear in) Neural Computing and Applications.
- Ke Fan, Ajmal S. Mian, Wanquan Liu, Ling Li: Unsupervised Manifold Alignment using Soft-assign. (submit to Machine Vision and Applications).

Chapter 1

Introduction

1.1 Problem Statement

Face recognition is a challenging problem for identifying a person from images or videos. It has been an active research topic in computer vision community over the past two decades. Currently, there are many applications for face recognition. If an accurate image set based facial recognition system as defined below was properly designed, it could be useful in many applications. For instance, identification system, the most standard application for face recognition can be improved to handle pose and lighting variations. Surveillance is important application for face recognition as automated face recognition can be applied to search for 'dangerous' people or crimes in recorded data. Another important application is pervasive computing. Most of devices such as smart phones, tablets and wearable electronics have cameras which can be used to capture photos and identify their users automatically, and can thus provide personalization messages based on identification.

Traditional approaches for face recognition recognize a person from one single testing image and mainly assume that all images are taken in controlled environments. However in reality environments, face images are captured from varied video cameras with variation of conditions in pose, illumination, expression, etc. Most conventional single image approaches cannot handle facial variations in real world applications in uncontrolled environments. This is due to the fact that the difference between facial images caused by those variations are often lager than the identity difference itself. Thus the distance in feature space between two faces of different persons in the same viewpoint may be smaller than that of the same person under different poses. That makes the Nearest Neighbor (NN) classifier in conventional methods to fail easily.

Nowadays, it is easy to obtain large quantity of images for both training and testing. Theoretically, a set of images for the same individual should provide more variation information in pose, illumination and expression and thus it is more feasible to find the intrinsic identity feature to classify different subjects. Methods based on image sets are expected to achieve better performance than traditional ones based on a single testing

image, because a set of testing images can incorporate information about the variability of the individual's appearance, and make decision based on identity information collectively. In this situation, face recognition problem can be formulated as follows: *Given a set of query images for one unknown subject, we need to design a classifier based on the training image sets and use it to find the label information for the query image set. This is called Face Recognition based on Image Set (FRIS).*

FRIS is not an easy task due to the fact that the larger the data set is, the more external factors there are in effect. These factors can include the lighting of environment, the continuous change of viewpoints, and different facial actions. For example, Figure 1.1 shows images of three sets of different people at various poses, lighting conditions and facial expressions. Although the faces in each row have the same identity, they still can be very confusing to a computer system. However more images mean that more information are provided. If appropriate approaches can be developed to extract and effectively utilize features from image sets, we can expect that a well-developed computer system could achieve satisfactory performance.

Figure 1.1: An illustration of the variation of conditions in three image sets, each row is a set for one subject. The illumination, pose and facial expression are all different for images in three image sets.

1.2 Aims and Research Goals

It has been the goal of face recognition research to have the ability of recognizing identity in complex scenes. While this seems natural and simple to humans, it is still a challenging task to computer currently. Despite the research on this topic is fast developing, there are a number of limitations in the existing image set based face recognition methods.

There are two main categories of methods to solve the FRIS problem. Firstly, some approaches define the set-to-set similarity based on the local samples between sets, which is restrictive for sets having similar environment condition. If the sets are captured in similar condition, those algorithms work perfectly, because these samples have very little variation to each other. However, this kind of methods are normally sensitive to noise and outliers. Some other approaches manage to find a model (e.g., linear subspace and affine hull) for one person and are compared according to the similarity of the models. This requires that the sets are widely dispersed under different conditions to span a subspace or manifold. The advantage of model based methods is that few outliers samples may not influence the model building process and are naturally denoised in the process. However, in the real application of the face recognition, the samples are often insufficient to build a good model.

To this end, novel robust algorithms are required to take the advantage of both categories for unrestricted environments, and overcome the problems in the existing approaches. Specifically, the desired algorithm should include the following properties:

- No restriction on the relation between sets is required. The proposed algorithm should be able to handle all set intersection types for poses, i.e., no overlap, with overlap and even one set being contained in another set.
- No restriction on variation of set is required. The proposed algorithm will handle sets which have not enough variations to build a model of identity.
- The proposed algorithm is robust against image noises and outliers.

1.3 System Overview

In order to achieve these goals, a framework of face recognition based on image set is developed in this thesis, which includes many novel techniques and algorithms. The system

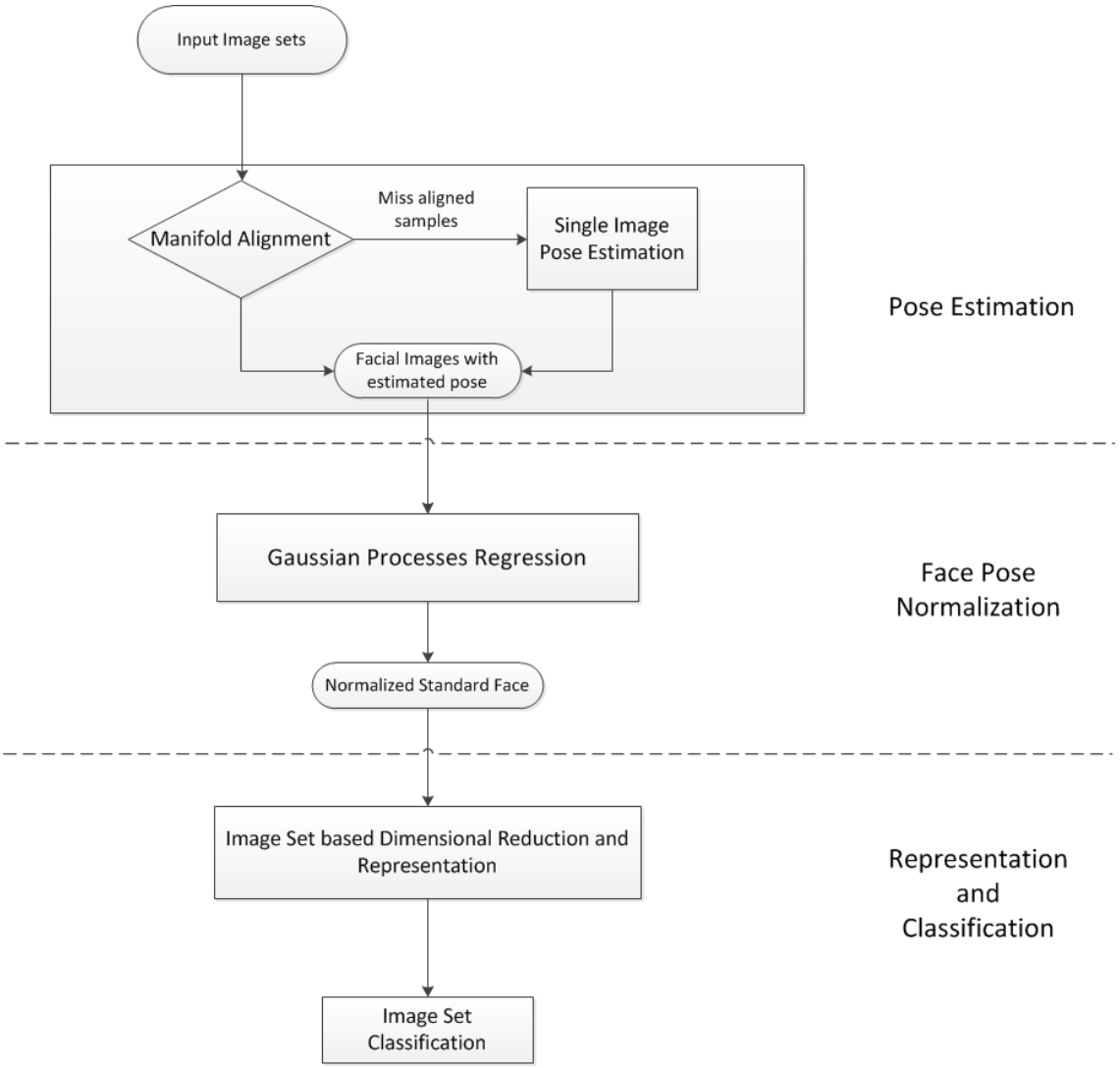


Figure 1.2: Overview of the framework. This framework contains three main parts: (a)Pose Estimation. (b)Face Pose Normalization. (c)Representation and Classification.

framework is shown in Figure 1.2. In this thesis, we build up a novel framework for general image set based face recognition problem by developing a series of techniques. This framework is designed to overcome the limitations of existing techniques and achieve the research goals in Section 1.2. The main idea of the whole system is to apply the FRIS method to a variation of uncontrolled environments and we intend to significantly decrease the influence of the variations in input images. All the input faces with different poses are required to be normalized into some standard frontal faces. That means this system have the ability to handle more wild cases in pose variations.

We explore different ways to learn features, model image sets and use these techniques to achieve these aims. In particular, we tackle the following tasks: image sets alignment, face normalization and feature extraction. The image set alignment task is to discover the intrinsic relationship of two data sets. The alignment is performed on each set with a reference set to find the correspondence. Its output is a one to one correspondence between sets. It is an effective way to automatically label the face pose. The face normalization task is a convenient intermediate step before classification in order to remove large pose variation impact. All faces are normalized into a standard frontal pose that reduces the pose variations in order to achieve better performance. The set based feature extraction task is to find more discriminant low dimension representations that can improve face recognition performance for FRIS problem.

1.4 Significance and Contributions

This thesis makes contributions to the field of computer vision in two levels: framework and techniques. From the point view of framework we propose a new system for image set face recognition which has the capability to handle FRIS in the uncontrolled environment. For each component of this system, some new techniques are proposed to improve state of the art approaches. We summarize the main technical contributions of this thesis as below.

- We propose a new unsupervised algorithm for the automatic alignment of two manifolds of different datasets with possibly different dimensionalities. It requires the reference set should cover all the variations which could appear in query set. The proposed algorithm automatically establishes an initial set of sparse correspondences between the two datasets by matching their underlying manifold structures. Local histogram features are extracted at each point of the manifolds and matched using RANSAC algorithm to find the initial correspondences. Based on these sparse cor-

respondences, an embedding space is estimated where the distance between the two manifolds is minimized while maximally retaining the original structure of the manifolds. The problem is formulated as a generalized eigenvalue problem and solved efficiently. Dense correspondences are then established between the two manifolds and the process is iteratively implemented until the two manifolds are correctly aligned consequently revealing their joint structure. We demonstrate the effectiveness of our algorithm on aligning protein structures, facial images of different subjects under pose variations and RGB and Depth data from Kinect. Comparison with a state-of-the-art algorithm shows the superiority of the proposed manifold alignment algorithm in terms of accuracy and computational time.

- We propose another improved robust unsupervised algorithm for automatic manifold alignment. The significant contribution is that the proposed alignment algorithm is performed automatically without the previous assumption on the correspondences between the two manifolds. For such purpose, we simplify the histogram-based features of the previous work. The elegance of this idea is that such a complicated problem is formulated as a generalized eigenvalue problem, which can handle the outliers by using the extended correspondence matrix. The alignment process is achieved by iteratively increasing the sparsity of the correspondence matrix until the two manifolds are correctly aligned and consequently one can reveal their joint structure. We demonstrate the effectiveness of our algorithm on different datasets by aligning facial images of different subjects under pose and lighting variations. Finally, we also compare with state of the art algorithms and the results show the superiority of the proposed manifold alignment in terms of vision effect and numerical accuracy.

- We propose a face normalization method for bridging the gap of the large pose difference in face recognition task. We address the problem by directly transforming the image of non-frontal pose into the frontal view image. Then the recognition can be performed by any state of the art classification method. For such purpose, a group of Gaussian Processes Regression (GPR) models are used to normalize pose into a frontal view. The GPR models are learned independently for different poses. A final joint output estimation is the product of Gaussian distributions. We compare with state of the art algorithms and the results show the superiority of the proposed technique in terms of normalization error and numerical accuracy.

- We propose a new dimensionality reduction method Margin Preserving Projection (MPP) for image-set based face recognition. This proposed method is designed for affine hull modeling based FRIS problem. In the proposed method, we transform each set into a convex hull and use Support Vector Machine to compute the margins between each pair of sets. Then we use PCA for dimension reduction with an aim

to preserve these margins. Finally we use convex hull distance to do classification in low dimension feature space. Experiments with benchmark face video databases validate the proposed approach.

- We propose a novel framework for FRIS. This framework can handle the extreme case when there is no pose overlap between the training set and the query set. The image sets firstly are aligned to a reference set in order to estimate the pose. The face normalization is then applied to reduce the wild pose variations. The obtained frontal standard faces are finally used for a normal FRIS problem. Experiments on different setups show our system outperforms some state of the art methods in general cases.

1.5 Structure of the Thesis

This thesis is organized as follows. In Chapter 2, a review of related works is presented. The state of the art approaches of FRIS are first briefly discussed. This is followed by exploring the related face recognition techniques,e.g., illumination robust face recognition and pose robust face recognition. Then the system related techniques are discussed respectively in more details. For example, the manifold alignment is reviewed, and it is adopted to quickly estimate the pose of each image in a set. In Chapter 3, we introduce Manifold Alignment using RANSAC technique to show the basic idea. The improved Manifold Alignment algorithm is presented in Chapter 4 as we remove the assumption of the previous version and build a more elegant algorithm to solve the alignment problem. The face normalization step in this framework is presented in Chapter 5 as we discuss about the Gaussian Processes Regression (GPR) based face normalization techniques in details. In fact, the nonlinear characteristic of GPR brings satisfactory performance in cross pose face recognition experiment. In Chapter 6, A discriminant learning algorithm is designed for convex hull model based methods. Finally, the whole system is built in Chapter 7. Conclusions and directions for future work are addressed in Chapter 8.

Chapter 2

Literature Survey

In this chapter, we will review literatures about FRIS approaches for better understanding the limitations of the existing methods. Furthermore, relevant literatures for each component are also reviewed for better analysis to show their advantages and disadvantages.

2.1 Face Recognition based on Image Set

Initially, there is no specific methods proposed for FRIS problem. Some researchers attempt to integrate results from traditional single image face recognition methods. They apply frame based techniques to all or selected frames from face sequences, and then obtain corresponding results using majority voting or other decision level fusion algorithms Zhao *et al.* (2003). This strategy ignores some important information for the correlation in the image set. Experiments in Hadid and Pietikainen (2004); Wang *et al.* (2008) show that image set based methods outperform those single frame based ones with direct applications on FRIS.

The emergence of face recognition based on image set actually is a natural development of face recognition system. Although people can find identity through a single image, human recognizing is a dynamic process in everyday life. The dynamic in computer system corresponds to multi images in a video. Researches of FRIS then start from image sequences based methods which use temporal coherence within a sequence. There is a strong requirement of those applications that the user should perform a strictly pre-defined motion with controlled lighting setup. Typical approaches are condensation method and Hidden Markov Models (HMM) based methods Hadid and Pietikainen (2004); Liu and Chen (2003). The consecutive assumption of consecutive motion is not always satisfied, because in most case data may be derived from unordered observations, e.g., multi-sensor surveillance systems and videos in different periods. Therefore the general case of FRIS is assumed that the training and testing data used in a FRIS system are organized in set of images for each subjects with random and typical variations in illumination pose and facial action, but the temporal coherence may be unnecessary in many applications.

Unlike the single image based method, the similarity is naturally defined on sample to sample distance. Considering about image set, there are some difficulties: The size of the image set normally is different and big. There should be a model to incorporate all the samples; There is no an effective metric for image set comparison. Therefore many researchers focus on exploring approaches in following aspects for FRIS problem:

- How to build models to handle image set? There are two main categories of solutions for this problem. One is parametric method with assumption that the images in a set satisfy certain distributions; the other one is model-free nonparametric methods in which the set is seen as sampling of a structure or a manifold.
- How to define the similarity between sets? One straightforward idea is to modify the sample to sample distance to for image set. More elegant idea is to build the set into a structure model and the distance is simply defined using the similarity of models.
- How to extract discriminative features and design classifier with a given similarity?

In next Section 2.1.1, we will mainly focus on reviewing the modeling technique of image set. The discussion about the two main categories on how to define the similarity measure is presented in Section 2.1.2. The details of learning the discriminative representation of the set are described in Section 2.1.3.

2.1.1 Modeling Methods

Existing techniques can be categorized according to the three main challenges of the FRIS problem. The first challenge is how to extract and represent the information from an image set. To tackle the first challenge, existing techniques include parametric and nonparametric representations. Parametric methods model distributions to represent an image set with the parameters estimated from the dataset itself. Without any assumption on data distribution, nonparametric methods represent many favorable properties in more flexible modeling manners.

2.1.1.1 Parametric Methods

Parametric approaches assume that the data satisfy a certain probability distribution. Image sets are then represented by using the parameters of the distribution, e.g., mean

and standard deviation of Gaussian distribution. Shakhnarovich *et al.* (2002) introduce the probabilistic modeling method. They represent each set by a multivariate Gaussian distribution and measure the similarity using the Kullback-Leibler divergence. Arandjelovic *et al.* (2005) introduce manifold density divergence which represents each set using Gaussian mixture models (GMM) to produce more realistic modeling. Kernel PCA is used in Arandjelović and Cipolla (2006) to build a dissimilarity measure between distributions of face sets. The resistor-average distance is then applied on nonlinearly mapped data and used as similarity. The significant limitation of the parametric methods is that the learned distribution parameters based on training data can be quite different from the testing set. The arbitrarily densities of sets can significantly affect the parameter estimation.

2.1.1.2 Nonparametric Methods

Nonparametric methods include techniques that do not assume data to satisfy any particular distribution and model the data in a more flexible way. These methods mostly attempt to represent the image set by linear subspace or by nonlinear manifold. Mutual Subspace Method (MSM) Yamaguchi *et al.* (1998) consider each image set as a linear subspace. Based on MSM, an improved Constrained MSM (CMSM) Nishiyama *et al.* (2005) project the basis of subspace onto a constrained subspace to handle pose variations. In MSM, the measure of difference between linear subspace is the sum of cosines of few smallest principal angles. A basic limitation of these methods is that real world variations usually lead to high curvature of the underlaying manifold structures and have outliers problem, thus linear subspace may lose good discrimination in this kind of situations. In Wolf and Shashua (2003) the performance is improved by modeling subspace in a nonlinear way using the kernel trick. However finding the optimal kernel function and parameters is a very complicated problem. Mixture of linear subspace Nishiyama *et al.* (2007); Fan and Yeung (2006) is another way to model the nonlinear structure of sets. Wang *et al.* (2008) model face appearance as some local approximations of manifold using a clustering method, which depends on the difference between geodesic distance and Euclidean distance. The similarity is measured by a weighted average of canonical angles and exemplar distances. It is much easier to formulate due to the linear locality, but accuracy of models will be lower than global nonlinear methods. Grassmannian manifold Chang *et al.* (2007) is another way to model face image set by considering the geometric structure of data. The kernel trick using in Grassmannian manifold methods Wang and Shi (2009); Harandi *et al.* (2011) is able to handle the non-linearity structure in data.

2.1.2 Set-to-Set Similarity

It is necessary to properly design corresponding measurement of the similarity for specific models. Some methods have used conventional metrics, e.g., Kullback-Leibler divergence to parametric method and principal angel for linear space. To capture the set nonlinearity, complicated models like nonlinear manifold normally need to define effective distance. Two types of measurement have been reviewed as below. The first one defines the set-to-set distance using some of the set samples. The second type of similarity is based on their model structure.

2.1.2.1 Local Sample based Similarity

Derived from traditional face recognition, pair-wise comparison is the most intuitive way. The similarity between sets can be defined by using minimal, maximal or mean distance of all pair distances. Earlier work Satoh (2000) directly uses the single frame criterion by matching the closest pair of samples. The result of Wolf *et al.* (2011) shows that directly using pair-wise distance can not achieve satisfactory performance. Cevikalp and Triggs (2010) model each image set as a affine hull, the similarity is the distance of nearest points between them. Those points are actually the linear combination of samples within each affine hull. However it can not be guaranteed that the synthetic points of linear combination represents a face. Sparse Approximated Nearest Points (SANP) Hu *et al.* (2011) include a additional sparse constraint with the affine hull modeling to avoid the wild linear combination. In Hu *et al.* (2012) the extension of SANP in kernel version improves the matching performance of image sets. Another improvement Yang *et al.* (2013) replaces the L1 norm constraint in SANP with L2 norm constraint with a slightly performance improvement.

Experiments show that if the data satisfy the condition that all sets have images with similar variations, locality based methods demonstrate significant superiority in both accuracy and efficiency. However, such restrictive assumption has limited the use of Locality based methods. If the data sets can strictly meet such assumption, this type of methods are still a good choice.

2.1.2.2 Structural Similarity

The limitation of Locality based approaches is that the similarity metric is based on small region of samples, and most of the other images are not utilized. Modeling each image set into a holistic model is a solution. As discussed in Section 2.1.1, the holistic set structure of each image set can be generally considered as a linear subspace Yamaguchi *et al.* (1998); Kim *et al.* (2007) or a nonlinear manifold Harandi *et al.* (2011); Wang *et al.* (2008); Fan and Yeung (2006). A computationally efficient way of computing the similarity between two linear spaces is to calculate their canonical correlation, which is defined as cosines of principal angles Kim *et al.* (2007); Wolf and Shashua (2003). However, the global structure may be a nonlinear manifold and a single subspace can not well represent it. Wang *et al.* (2008) divide an image set into multiple local linear clusters, then use principal angle and cluster exemplar distance to measure the similarity. Chen *et al.* (2013) compute the distance between different local linear subspaces to achieve better performance. Although all these techniques can discover the structure feature of set, classification is still implemented based part of the data points, and the rest of the image set data are often not effectively utilised. Mahmood *et al.* (2014) apply improved spectral cluster on a big data set containing all training data and testing data. It is divided into small groups to achieve the cluster-wise correspondence between training and testing data. The classification is determined by distribution distance using most of the data.

The limitation of structure based methods is the requirement for dataset that need to have large divergence. Only data samples with typical variations can span a low dimensional subspace. If the data is not adequate, it is hard to reveal the complete structure of data.

2.1.3 Image Set Discriminant Features

Discriminant Analysis is a powerful technique which is widely used in the traditional single image face recognition. Mutual Subspace Method (MSM) Yamaguchi *et al.* (1998) uses the Principal Component Analysis (PCA) Turk and Pentland (1991) to generate subspace of data and define the similarity based on the smallest principal angle. Directly applying Traditional method on FRIS problem can not achieve satisfactory performance, since the similarity definition is different and the set based discriminant feature may be lost in the process. Therefore learning the discriminant representation for FRIS should properly design for specific method. Based on MSM, an improved Constrained MSM (CMSM) Fukui and Yamaguchi (2005) projects the basis of subspace onto a constrained subspace to handle pose variations. Discriminant analysis of Canonical Correlations (DCC) Kim *et al.*

(2007) is developed for Canonical Correlations (principal angle) based methods. It follows the idea of Linear Discriminant Analysis (LDA) Belhumeur *et al.* (1997) by replacing the metric with canonical correlations and finally solve this problem iteratively. Grassmann Discriminant Analysis (GDA) Hamm and Lee (2008) uses two different metrics to define new Grassmann kernels for Kernel Linear Discriminant Analysis (KLDA) Scholkopft and Mullert (1999). Kim *et al.* (2006) propose the Locally Orthogonal Subspace Method (LOSM), where the class subspace is only required to be orthogonal to its local neighbors. Kim and Cipolla (2009) improve LOSM by incrementally updating the principal components of the class correlation and the total correlation matrices. Discriminant analysis is also used in the clusters local patches Wang and Chen (2009), where different class data are better separated and local clusters are more compacted.

2.1.4 Section Summary

In Section 2.1, we reviewed the three aspects of FRIS problem. Methods are carefully analysed for the advantages and limitations. Two modeling methods are designed for different case of input data. The nonparametric methods are more flexible to utilize in most of case. Two types of similarity are compared in Section 2.1.2. Sample based methods assume there are overlap of variations between sets. Structure based methods need the data to be sufficient span a low dimension space. However, both of them can not be used effectively in the uncontrolled environment. In this thesis, we will propose a novel framework that can be used in uncontrolled environment. We combine Manifold Alignment and Face normalization technique to handle the wild case. In next sections, We will introduce those techniques.

2.2 Manifold Alignment

Manifold Alignment is a technique to align two high dimensional datasets when their individual elements are quite different. It is often hard to directly find the element-wise matches, especially when there is no prior information about their correspondences. To solve this problem, many researchers have developed some techniques includes relaxation labeling Maciel and Costeira (2003), graph spectra Egozi *et al.* (2013) and tensor model Chertok and Keller (2010). Recent researches on manifold learning Tenenbaum *et al.* (2000); Roweis and Saul (2000); Zhang and Zha (2004) provide a way to discover the intrinsic structures of high dimensional datasets in a low dimensional space. Manifold alignment is a powerful technique using this idea for establishing an effective correspon-

dence between different datasets. Such correspondences are a powerful tool for knowledge transfer across disparate datasets using their underlying intrinsic manifold structures.

The main idea of manifold alignment is to optimize a problem with two criteria. First, the alignment should minimize the distance between manifolds, i.e., the two manifolds must lie close to each other in a feature space. Secondly, the alignment should preserve the structures of both manifolds, i.e., the relationship between the data elements of individual manifolds must be preserved as much as possible. Current manifold alignment techniques mainly focus on two aspects: feature extraction and joint manifold discovery. Feature extraction tries to find the structure feature of manifold to give an instruction for defining distance between manifolds. Based on features or prior information, the final aim is to obtain the joint manifold which satisfy all constraints.

2.2.1 Feature Extraction for Alignment

Feature extraction can be categorized as supervised, semi-supervised and unsupervised approaches. Supervised approaches, such as Wang and Mahadevan (2008, 2009a), build correspondences based on a set of manual point pairs selected between the two manifolds. If the known correspondences are inadequate and low quality for supervised approaches, we can use semi-supervised methods Ham *et al.* (2005) to find dense and refined correspondence estimation. For unsupervised methods, all of the features are extracted based on local properties without prior manual information. Wang and Mahadevan (2009b) introduce a similarity metric based on the permutations of the k nearest neighbor Euclidean distances, which is computationally expensive. Moreover, it only uses the simple inter-sample Euclidean distances within a manifold as features. Such features can be highly sensitive to the data sampling. Pei *et al.* (2012) describe a similarity estimation method based on features of B-spline curves, which are fitted to the local neighbors. Cui *et al.* (2012) introduce another unsupervised feature based on Canonical Correlation which is using the similarity of image appearance.

2.2.2 Alignment and Joint Manifold Discovery

There are different methods to find joint manifold structure and refine the correspondence after the similarity has been built. Two-steps methods like Procrustes alignment Wang and Mahadevan (2008) are firstly embedded into a low dimensional feature space while preserving their individual manifold structures using dimensionality reduction such as

Locality Preserving Projection (LPP) He and Niyogi (2003) or Laplacian eigenmaps Belkin and Niyogi (2003). Based on correspondences manually or automatically established in the feature extraction stage, a transformation is derived so that it optimally aligns the two manifolds by minimizing the distance errors between the corresponding elements. Iterative Closest Point (ICP) Besl and McKay (1992); Chen and Medioni (1992) is a rigid registration and can be seen as spacial case in 3D for this propose. Pei *et al.* (2012) consider the transformation as an extended affine transform and an instance matching function. The correspondences of final result are defined only based on the data elements in the original space. As such an embedding space is not learned, which makes it difficult to handle new data. Another limitation is that the embedding is performed separately without taking the similarities between the two datasets into account. Consequently, the transformation in the embedding space cannot guarantee an optimal alignment.

Other blended methods, such as Wang and Mahadevan (2009b); Cui *et al.* (2012), formulate the correspondence and embedding into feature space as a single optimization problem. Due to the non-convex nature of the optimization function, it is iteratively solved by first initializing with some rough correspondences, for example, using prior knowledge. All of these methods are directly using binary matrix to model the correspondence matrix, they are sensitive to the initial binary correspondences.

2.3 Pose Robust Face Recognition

As we mentioned in Chapter 1, the variations in pose, illumination and expression would cause larger facial changes and these can significantly drop the recognition performance. The lighting problem can be solved based on edge based features Wang *et al.* (2004); Tan and Triggs (2010). The expression is a partial deformation of face that can find stable regions like nose for face recognition Chang *et al.* (2006). Out of the three types of variation, pose variation remains the most challenging problem for face recognition. The changes caused by pose variations are non-linear and significantly larger than the identity differences. Thus the distance in feature space between two faces of different persons in the same viewpoint is smaller than that of same person under different poses causing simple Nearest Neighbor (NN) classifiers in conventional methods such as Eigenface Turk and Pentland (1991) and Fisherface Belhumeur *et al.* (1997) to fail easily. In our system, the alignment outliers are often under different poses. This problem could be solved using Cross-pose face recognition technique.

Many different approaches have been reviewed in the literature Zhang and Gao (2009) to

tackle the pose problem. Multi-view approaches simply import multiple pose face images to train the face recognition system and thus reduce the impact of large pose variations. Such approaches cannot be considered as cross pose because they can only recognize previously seen poses. Some methods e.g. Sanderson *et al.* (2006), transformed frontal face images to generate non-frontal faces for extending the training set. Then the training set is used to collaborate for face recognition under pose variations, that is, one actually can convert a multi-view problem to a FRIS problem Cevikalp and Triggs (2010); Hu *et al.* (2012). Next we will introduce some techniques dealing with pose variation. In this thesis, we will use Gaussian Processes Regression (GPR) technique to do pose normalization.

2.3.1 Assistance of 3D Models

One category of face recognition approaches dealing with pose issue is recognition with assistance of 3D models. In fact face images are 2D projections of 3D human face under different viewpoints. To handle face images in small pose variations Gao *et al.* (2001) propose a simple pose recovery method using a generic cylindrical 3D face model. Face images under any horizontal poses where mapped on the generic cylindrical model, are rotated and recovered to the frontal face for recognition. Castillo and Jacobs (2007) apply stereo vision techniques to reconstructs 3D face models. The cost of stereo matching of face image set is used as the measure. Another example is the 3D Morphable model (3DMM) Blanz and Vetter (2003) based 2D face recognition method. The 3D Morphable Model is formed using Principal Component Analysis features built from 200 scanned 3D faces with shape and texture. Then the 3DMM is morphed to generate a given 2D face image as closely as possible by iteratively minimizing pixel difference of image and the model. This is a highly non-convex optimization process which estimates a large number of parameters including 199 shape, 199 texture, camera pose and lighting condition. The final set of PCA parameters are used to encode gallery faces and recognize a probe face using NN distance between the gallery and probe parameters. Due to the non-convex nature of the optimization, a unique global solution is not guaranteed and the technique can fail. Moreover, this technique also needs manual initialization of the 3DMM pose. The 3D face morphable Blanz *et al.* (2005) can also be used to generate novel 2D non-frontal face images for training. However, the texture model is not a pure albedo and also contains the lighting conditions of the 200 training images.

In conclusion, reconstructing 3D models from a 2D image is an ill-posed problem. Generating side poses accurately from frontal views requires the person specific 3D face models which are not available in most cases. Using an average 3D face model for all faces can generate inaccurate side poses.

2.3.2 2D Feature Matching

Directly matching two 2D face images with different poses is not as convenient as 3D since approaches for 3D like ICP Besl and McKay (1992) are very robust. In 3D case, the local regions of face are considered to be robust to pose variation Kanade and Yamada (2003). From different view points, local patches are situated in different positions which motivates the need for cross pose patch alignment. Ashraf *et al.* (2008) propose an alignment method by Lucas-Kanade-like optimizing process. Li *et al.* (2009) use a generic 3D model to estimate the patches correspondence and measure the similarity by canonical correlation analysis. Another way is to find the pose invariant feature for recognition Huang *et al.* (2007); Levine and Yu (2006). However, these features are not robust under large pose variations. Sharma and Jacobs propose a partial least square (PLS) based method Sharma and Jacobs (2011). PLS can project images from different poses into a common feature space such that the variance of corresponding features is maximized in that space. Preserving only the maximum variance implicitly discards some changes due to pose variations. However, due to the complex nature of pose variations, PLS cannot achieve complete pose invariance.

2.3.3 Pose Normalization

2D pose transformation is another simplified effective method dealing with pose variations. Rather than reconstructing 3D face from 2D facial appearances, one can find a direct 2D image mapping to synthesize a virtual views across different pose gap. Some researchers Chai *et al.* (2007); Li *et al.* (2012); Zhang *et al.* (2013) found that effective regressors can be learned to bridge the coupled faces across different poses. This regression step can be seen as a preprocessing step for face recognition. Actually for every side face, there is one corresponding frontal face. Theoretically the relation of two faces under different views is estimated by a linear projection which is formed with 2D-3D projection and rotation transform in the presence of 3D model. Such fact supports that directly normalization using machine learning technique is achievable and efficient. In Chai *et al.* (2007) a simple linear regression has been used for constructing this transformation. In considering the variations in distortion, alignment and cropping errors, the theoretical linear regression transformation is not always satisfactory for the real-world data. Improved methods consider the error by coupling the bias and variance of source images and target images by using L1-norm Zhang *et al.* (2013) and L2-norm Li *et al.* (2012). However, the main function still follows the linear assumption. By considering this problem as a nonlinear regression problem, we will propose a nonlinear data driven model to handle

such transformation.

2.4 Chapter Summary

We present a review of related works in this chapter. The state of the art approaches for FRIS problem are briefly discussed. There are three key aspects in solving the FRIS problem. For modeling, the Non-Parametric methods are better performed than Parametric methods, since it have less restrict on data. Different metric criteria show different limitations in the uncontrolled environment. Either of structure based and sample based methods are limited for our aim to solve the uncontrolled FRIS problem. The methods for learning the discriminate feature are finally reviewed. This technique can further improve the recognition performance.

Following the FRIS technique review, we explore techniques related to FRIS, e.g., face recognition under varying pose and lighting condition. Then the system related techniques are discussed respectively in detail. The manifold alignment can be adopted to quickly estimate the pose of each image set. It includes two parts: Feature extraction and joint structure discovery. Pose robust face recognition includes a lot of techniques to solve pose variation issue. We will adopt the face normalization technique to reduce the pose variation in both training and testing set.

Chapter 3

Image Set Manifold Alignment

As discussed in Chapter 2, the existing methods for face recognition based on image set have been reviewed and analyzed. Among them, the works that are state of the art and most related to the aims of this thesis are Mahmood *et al.* (2014) and Cui *et al.* (2012). They have similar ideas to combine the structure based and sample based method to improve the FRIS problem. The first step is to find correspondence between sets. In Mahmood *et al.* (2014), a rough correspondence is obtained by a hierarchical sparse spectral clustering to group the two sets into tiny clusters that the corresponding samples are lied in the same group. There is no exact correspondence achieved in the process, hence the refined process on the sets may not be achievable. We intend to find better and accurate correspondence using manifold alignment technique with similar criteria like Cui *et al.* (2012). The main idea of manifold alignment is discussed in Section 2.2. Firstly, the alignment should minimize the distance between manifolds, i.e., the two manifolds must lie close to each other in a certain feature space. Secondly, the alignment should preserve the structures of both manifolds, i.e., the relationship between the data elements of individual manifolds must be preserved as much as possible.

The limitations of previous alignment approaches we discussed in Section 2.2, are 1) sensitive to the initial correspondence condition and 2) hard to take different types of similarities into account except the closing neighbors. We propose a manifold alignment method to overcome these problems and achieve better performance than state of the are methods. A two-step process is proposed for estimating the initial set of correspondences which is then followed by an iterative single-step refinement procedure that can find better and dense correspondences. Such a process combines the advantages of different type of methods. Our approach falls in the unsupervised category avoiding the cumbersome manual annotations associated with supervised approaches. The aim of this manifold alignment method is to find the correspondence samples in a complete reference set for a test set. We will discuss a more general case which requires no prior relationship assumption in Chapter 4.

More specifically, the technique contributions are summarized as follow. Our method extracts a novel 16-dimensional histogram from rotation invariant features such as the re-

lationships between normal vectors. We use RANSAC to perform robust feature matching which not only gives accurate matches but also avoids the massive computational complexity required for exhaustive comparisons Wang and Mahadevan (2009b). An initial set of correspondences is automatically established based on the local structures of the manifolds that are invariant to transformations and robust to sampling. In the single-step stage, a point-wise distance between corresponding points is employed to minimize the inter-manifold distance, while a local reconstruction constraint is imposed to preserve intra-manifold structures. This is formulated as a generalized eigenvalue problem which can be efficiently solved. An iterative optimization framework is employed to refine the alignment. Additionally, a joint manifold structure is achieved which is useful for further information transfer between the datasets.

The effectiveness of the proposed algorithm is demonstrated on facial images of different subjects under pose variations. We also show the alignment ability on different data like protein structures and RGB-Depth data obtained from the Microsoft Kinect sensor. Comparison with the manifold alignment proposed by another unsupervised method Wang and Mahadevan (2009b) shows that our method is more accurate and executes about 20 times faster.

3.1 Initial Correspondence Estimation

To establish correspondences for unsupervised manifold alignment, one way is to directly match their features in the original space. However, this is not possible if the two datasets are from different sources or their dimensions are not equal. Even if the data are in same dimensions like face images, there are still identity and variation differences. An alternate way is to find the similarity between the local structures of the manifolds. *The assumption in manifold alignment is that the two datasets have similar underlying manifold structures in feature space.* Following this idea, we firstly apply a manifold learning algorithm to discover the underlying manifold structures. This is essentially equivalent to embedding the manifolds in some feature space where the local relationship between the manifold samples is preserved. Features extracted in this embedding space can be matched directly to find correspondences between two manifolds.

Given a reference dataset $X = [x_1, x_2, \ldots, x_{N_x}] \in R^{d_x \times N_x}$ and a test dataset $Y = [y_1, y_2, \ldots, y_{N_y}] \in R^{d_y \times N_y}$, where the columns x_i and y_i are the samples, d_x and d_y are the dimensions of datasets, and N_x and N_y are the number of samples in X and Y respectively.

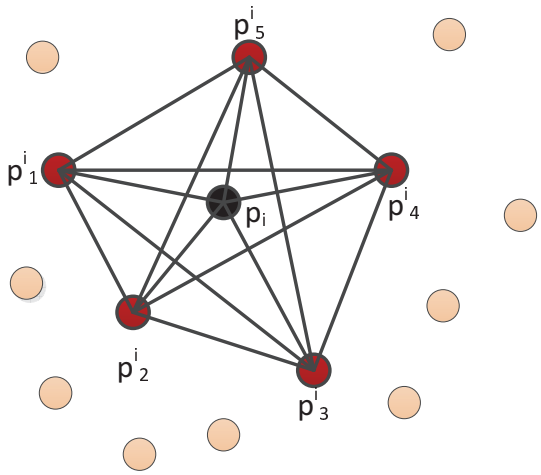


Figure 3.1: Example of a query point p_i and its k-neighbors. The query point and its neighbors are fully connected.

To measure the distance between two manifolds, a rough correspondence between their elements should be established first. We define an $N_x \times N_y$ correspondence matrix W^{xy} such that

$$W_{i,j}^{(x,y)} = \begin{cases} 1 & \text{if } x_i \text{ corresponds to } y_j \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Each column of $W^{(x,y)}$ can contain the value 1 at only one location. This means that for any point in Y, there is only one corresponding point in X ($N_x \geq N_y$).

Normally, the intrinsic dimension of the underlying manifolds is very low and the geometric features are hard to observe in the original space. Nonlinear dimensionality reduction techniques are capable of preserving the manifold structure in a low dimension embedding space. We apply a nonlinear manifold learning method called Local Generative units and Global Affine transformation (LGGA) Huang *et al.* (2009) on two datasets X and Y to find same dimension features P_x and P_y. Before feature extraction, the manifolds are normalized with respect to the scales of their respective largest principal components. The scale of the embedded points is calculated first using Principal Component Analysis (PCA) and then the complete dataset is rescaled according to the largest eigenvalue.

3.1.1 Local Histogram Feature

In order to efficiently obtain and compare the local manifold structures, we define a feature based on the histogram of measurements that encode the neighborhood's geometric

properties. We choose a histogram based feature and it is orientation invariant. This is essential because the manifolds are generally misaligned initially.

Consider a point p_i in P_x. As shown in Figure 3.1, all of its k-neighbors p_k^i in the Euclidean space are selected and organized in $R_i = [p_i, p_1^i, \cdots, p_k^i]$, termed as the *k-neighbors*. The normal vector n_i at point p_i can be approximated by the normal of the best fit plane to k-neighbors using PCA. n_i can be obtained by selecting the eigenvector corresponding to the smallest eigenvalue Hoppe *et al.* (1992).

For each pair of points p_s and p_t in the k-neighbors R of point p, and their corresponding normals n_s and n_t (calculated from their respective k-neighbors R_s and R_t), we define a unit difference vector between them:

$$v = \frac{p_t - p_s}{\|p_t - p_s\|} \ . \tag{3.2}$$

Similarly, we define some angular features for the pair as

$$f_1 = \max(n_t \cdot n_s, -n_t \cdot n_s) \tag{3.3}$$

$$f_2 = \max((n_t \cdot (v \times n_s), -(n_t \cdot (v \times n_s)) \tag{3.4}$$

$$f_3 = |\arccos(\max(n_t \cdot v, -n_t \cdot v)) - \arccos(\max(n_s \cdot v, -n_s \cdot v))| \tag{3.5}$$

$$f_4 = \|p_t - p_s\| \tag{3.6}$$

The four features measure the curvature and the angles between the normals and the difference vector. f_1 and f_2 are dot products between the unit vectors. They are in fact the cosine of the angles between these vectors. The maximum operation is performed to select only the acute angles. Similarly, f_3 shows the difference of the two angles between the two normals and the difference vector. We quantize each feature into two bins to make a 16 dimensional histogram of the curvature features. Since the features are based on mutual angles, it can be proven that they are rotation invariant. Moreover, these features are still comparable when we choose different value of k for the two manifolds.

3.1.2 Correspondence Estimation

The 16 dimensional histogram features of two manifolds can be matched using the nearest neighbor metric in the Euclidean space. However, in practice, the correct correspondences my not be the nearest ones because such a low dimensional representation of the manifold structure is not unique everywhere. Directly choosing the nearest features as correspondences may not represent the manifold distance metric very well. A solution is to select

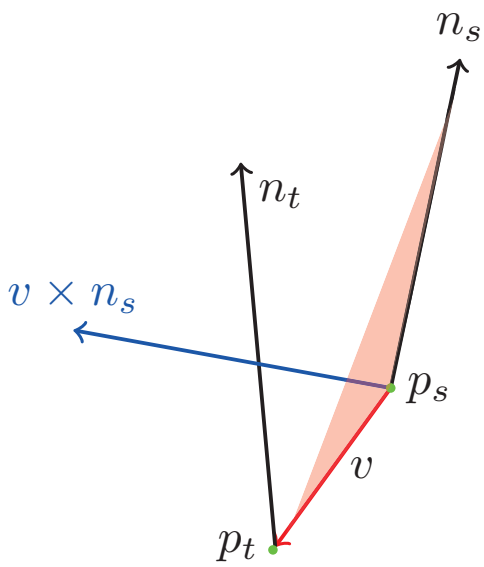


Figure 3.2: A pair of points and their normals and vectors used for constructing the feature histogram. This feature is rotation invariant.

the top few potential correspondences for each feature rather than the best one at this stage and later reject the outliers based on a robust algorithm. Next we will describe the process in detail.

Once the 16-D point features are obtained for a given k, the distance between two manifolds (structural similarity) can be represented as a distance matrix $D^k = [d_{ij}^k | i = 1, \cdots, N_x, j = 1, \cdots, N_y]$. The d_{ij}^k in the matrix is the local structure similarity between the ith point of X and the jth point of Y. A local feature correspondence set C^k can be defined as

$$C^k = \{(x_i, y_j) | d_{ij}^k < \theta_j\} \tag{3.7}$$

We only assign the pair (x_i, y_j) to the local feature correspondence set C^k when the distance d_{ij}^k is below certain thresholds θ_j for jth column. The parameters θ_j depend on the amount of ambiguity in the nearest neighbors step and can be adjusted based on the training data. We take the top 10% nearest neighbors to each point as its matching pairs, i.e., the value of each θ_j is set dynamically based on the jth column of D^k. In order to achieve reliable correspondences, they are chosen based on multiple values of k. Corresponding points should have similar feature even if the regions determined by the number of neighbors k changes. In our experiments, two different values of k are used. If a correspondence appears in multiple values of k it is accepted as reliable, otherwise it is

rejected, that is:

$$C = \bigcap_i C^{k_i} \tag{3.8}$$

where C^{k_i} indicates the set of correspondences which are selected for a given k_i. The intersection C still contains outliers which must be filtered out.

If the two manifolds have a similar structure in the embedding space, their rough alignment can be approximated by an affine transform. The RANSAC algorithm Fischler and Bolles (1981) is employed to find the inlier set by maintaining similar geometric relationships of the correspondences without exhaustively trying all combinations in C. Details are given below:

1. Select n correspondences from C as the initial inliers and compute an affine transform.
2. Apply the affine transformation on the test manifold and verify all other correspondence in C. Consider all other correspondences that fit well with the estimated transform as inliers.
3. Count the number of inliers as the confidence.
4. Save the correspondences with the maximum confidence and iterate until convergence or maximum number of iterations is reached.

A fixed number of iterations are used in our implementation. The final set of correspondences, with the maximum confidence, is selected to initialize the correspondence matrix W for the next alignment step.

3.2 Manifold Alignment

Given the small set of initial correspondences W between two manifolds, the dense correspondences and the joint manifold structure can be derived by iteratively solving a generalized eigenvalue problem as described below.

3.2.1 The Loss Function

Assume that there are two projection functions f_x and f_y that map X, Y to $F^{(x)} \in R^{d \times N_x}, F^{(y)} \in R^{d \times N_y}$ respectively, in an intrinsic space with dimension d. Note that $F^{(x)}$ and $F^{(y)}$ are the low dimensional representations of X and Y, and columns x_i and y_i are mapped to $F_i^{(x)}$ and $F_i^{(y)}$ respectively.

Our aim is to find projection functions f_x and f_y that can project the data into a joint low dimensional space. Inspired by some manifold alignment techniques in Wang and Mahadevan (2009a,b), we formulate the loss function for the mapping as follow:

$$J(F^{(x)}, F^{(y)}) = \mu J_C(F^{(x)}, F^{(y)}) + J_W(F^{(x)}) + J_W(F^{(y)}) \tag{3.9}$$

The first term J_C indicates the between-manifold distance via corresponding points across the datasets. The last two terms represent the reconstruction error of the locally generated models for each dataset. The following subsections provide the details of each term.

3.2.2 Manifold Matching Error

Measuring the distance between two manifolds can be represented as a problem of measuring the distance between their corresponding points. Thus the first term of (3.9) is as follows:

$$\begin{aligned} J_C &= \sum_{i,j} \|F_i^{(x)} - F_j^{(y)}\|^2 W_{i,j}^{x,y} \\ &= tr(F^{(x)}\Omega^x(F^{(x)})^T + F^{(y)}\Omega^y(F^{(y)})^T - F^{(x)}W^{x,y}(F^{(y)})^T - F^{(y)}W^{y,x}(F^{(x)})^T) \\ &= tr(FL^{xy}F^T) \end{aligned} \tag{3.10}$$

where Ω^x is a $N_x \times N_x$ diagonal matrix with elements $\Omega_{ii}^x = \sum_j W_{i,j}^{x,y}$ equal to the sum of corresponding rows of W. Similarly, Ω^y is a $N_y \times N_y$ diagonal matrix with $\Omega_{jj}^y = \sum_i W_{i,j}^{x,y}$, the sum of columns of W. We combine $F^{(x)}$ and $F^{(y)}$ into $F = [F^{(x)}, F^{(y)}]$ and build a new joint Laplacian matrix

$$L_{xy} = \begin{pmatrix} \Omega^x & -W^{xy} \\ -W^{yx} & \Omega^y \end{pmatrix} \tag{3.11}$$

3.2.3 Reconstruction Error

To ensure the local relationships are preserved when aligning two manifolds, we introduce a reconstruction error constraint for manifold structure discovery that is inspired by LGGA Huang *et al.* (2009). For any sample x_i, its k-neighbors $X^{(i)} = [x_1^{(i)}, \ldots, x_k^{(i)}]$ can be seen as a local linear region, and its element can be reconstructed based on PCA regarding x_i as a mean.

$$x_k^{(i)} \approx x_i + U^{(i)} v_k^{(i)} \ . \tag{3.12}$$

where $U^{(i)}$ and $v_k^{(i)}$ are obtained by PCA. For each $X^{(i)}$, this is easily achieved through Singular Value Decomposition (SVD) of the local region of x_i's K-neighbors

$$\bar{X}^{(i)} = (X^{(i)} - x_i e^T) \approx U^{(i)} \Sigma^{(i)} (V^{(i)})^T = U^{(i)} v_k^{(i)} \tag{3.13}$$

The obtained $v_k^{(i)}$. where $\Sigma^{(i)}(V^{(i)})^T = v_k^{(i)}$ is a denoised low dimensional representation of the original data with the minimized reconstructed error. Measuring the difference between projected $v_k^{(i)}$ and data in feature space indicates the reconstructed error of this low dimensional feature, which can be formulated as

$$J_i^{(f)} = \|W^{(i)} \Sigma^{(i)} (V^{(i)})^T - (F^{(i)} - f_i e^T)\|_F^2 \ . \tag{3.14}$$

where e is a column vector of all 1's. Now, let the matrix $S^{(i)}$ and S_i be the 0-1 selection matrix such that $FS^{(i)} = F^{(i)}$ and $FS_i = f_i e^T$, then the optimal $W^{(i)}$ that minimizes the error can be computed as

$$W^{(i)} = F(S^{(i)} - S_i)(\Sigma^{(i)}(V^{(i)})^T)^+ \tag{3.15}$$

where $()^+$ is the Moor-Penrose inverse matrix. The overall reconstruction error J_W is then given by

$$\begin{aligned} J_W &= \sum_i J_i^{(f)} \\ &= \sum_i \|W^{(i)} \Sigma^{(i)} (V^{(i)})^T - (F^{(i)} - f_i e^T)\|_F^2 \\ &= \sum_i \|F(S^{(i)} - S_i)\Theta_i\|_F^2 \\ &= \|FS\Theta\|_F^2 \end{aligned} \tag{3.16}$$

where

$$S = [(S^{(1)} - S_1), \ldots, (S^{(n)} - S_n)] \tag{3.17}$$

$$\Theta_i = (I - (\Sigma^{(i)}(V^{(i)})^T)^+ \Sigma^{(i)}(V^{(i)})^T) \tag{3.18}$$

$$= I - V^{(i)}(V^{(i)})^T \tag{3.19}$$

$$\Theta = diag\{\Theta_1, \ldots, \Theta_n\} \tag{3.20}$$

Equation (3.16) can now be rewritten in a compact matrix form,

$$J_W(F) = tr(FB_x(F)^T) \tag{3.21}$$

where

$$B = S\Theta\Theta^T S^T. \tag{3.22}$$

3.2.4 Solution of J

Following the above definitions of matching error and reconstruction error, the loss function (3.9) can be formulated as

$$\begin{aligned} \arg\min_F J &= \arg\min_F tr(FLF^T) \\ s.t. \quad & FF^T = I \ . \end{aligned} \tag{3.23}$$

Here, we impose the unit variance $FF^T = I$ constraint to guarantee a unique solution. It follows that F can be computed as the eigenvectors corresponding to the second to $(d+1)$th smallest eigenvalues of the middle sparse matrix

$$L = \begin{pmatrix} B_x + \Omega^x & -W^{xy} \\ -W^{yx} & B_y + \Omega^y \end{pmatrix} \tag{3.24}$$

Finally, Manifold Alignment can be solved as a general eigenvector problem.

3.2.5 Iterative Alignment

After solving (3.23), we can obtain the aligned joint manifold in feature space. The correspondences matrix W between the two manifolds can be recalculated based on this new finding joint manifold. In section 3.1, we assume that for every data point in query set, there is a correspondence point in the reference set. The correspondence update process can be achieved by finding the nearest reference sample for every point in query set. The whole process is repeated iteratively until a stopping criterion is reached. The complete procedure is summarized in Algorithm 1. Here $\text{Init}(X, Y)$ performs the initialization of correspondences as described in Section 3.1. B_x and B_y are calculated by (3.22) following the procedure in Section 3.2.3. $\text{Align}(B_x, B_y, W_t)$ is the optimization (3.23) that finds the joint feature in a low dimensional space using manifold alignment with a certain L computed from (4.8). The function $\text{NN}(F^{(x)}, F^{(y)})$ returns correspondences W based on the best matches in the joint feature space, i.e., the nearest points in $F^{(x)}$ for each $F_i^{(y)}$ are set as the correspondences.

Algorithm 1: Iterative Manifold Alignment using RANSAC

Input: X and Y: two datasets

Output:

$F^{(x)}$ and $F^{(y)}$: the embedding features of X and Y in low dimensional space.

W : the correspondence Matrix

begin
  $t \leftarrow 0$;
  $W_0 \leftarrow Init(X, Y)$ (Section 3.1);
  $B_x \leftarrow S_x \Theta_x \Theta_x^T S_x^T$;
  $B_y \leftarrow S_y \Theta_y \Theta_y^T S_y^T$ (Section 3.2.3);
  repeat
    $F^{(x)}, F^{(y)} \leftarrow \text{Align}(B_x, B_y, W_t)$;
    $W_{t+1} \leftarrow \text{NN}(F^{(x)}, F^{(y)})$;
    $t \leftarrow t + 1$;
  until $J_t - J_{t-1} < \epsilon$;
  $W \leftarrow W_t$
end

3.3 Linear Manifold Alignment

In Section 3.2, we assume that the projection functions f_x and f_y are unknown. According to this assumption, we can directly obtain the low dimensional representations $F^{(x)}$ and $F^{(y)}$. However we cannot easily project the low dimensional representation back to the original high dimensional space. Actually if we assume the projection is a linear one, we can obtain the projection matrix, that is $F^{(x)} = P_x^T X$ and $F^{(y)} = P_y^T Y$, and

$$F = P^T Z = \begin{bmatrix} P_x \\ P_y \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} \tag{3.25}$$

Next we can just slightly modify the optimization process (3.23) to obtain the projection P:

$$\begin{aligned} \arg\min_P J &= \arg\min_P tr(P^T Z L Z^T P) \\ s.t. \quad & P^T Z Z^T P = I \ . \end{aligned} \tag{3.26}$$

It can be seen that, this is still a general eigenvalue problem, and P can be computed as the eigenvectors corresponding to the smallest eigenvalues of ZLZ.

Normally, we apply linear manifold alignment in the final step after the whole iterative process and the optimal matches are obtained. The main advance for linear mapping

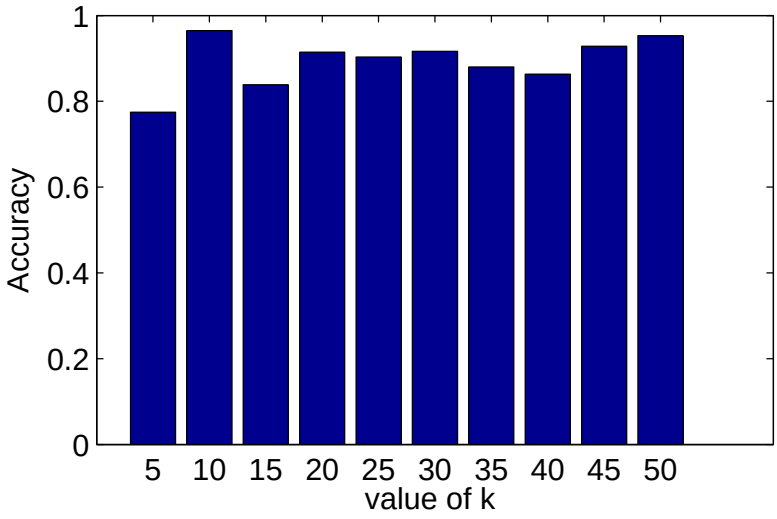


Figure 3.3: Matching accuracy with different values of k in the initial correspondence estimation.

is that we can transfer information between manifolds. We can project samples from original space, find correspondence in another manifold and project it back. However, linear manifold alignment will have larger error than the original nonlinear version.

3.4 Experiments and Discussion

Several experiments are conducted to demonstrate the effectiveness of the proposed manifold alignment. The datasets we used include protein bioinformatics data Wang and Mahadevan (2009b), face images with pose variations, and Kinect camera data comprising RGB and depth images. Wang and Mahadevan (2009b) apply to those datasets for performance comparisons. For the third dataset, the results of the proposed method are also compared with the ground truth provided inherently by Kinect i.e. the RGB images and depth images have a one-to-one correspondence.

3.4.1 Protein Data

Protein structure estimation is an important step in Nuclear Magnetic Resonance (NMR). The structure is estimated from partial pairwise distances with some constraints and human experience Wang and Mahadevan (2009b). Models related to the same pro-

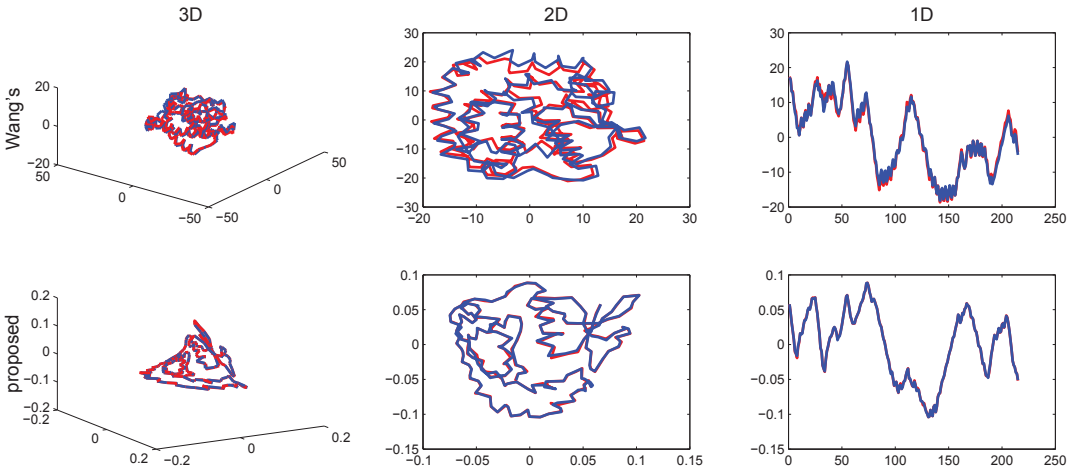


Figure 3.4: Alignment of protein structures using Wang and Mahadevan (2009b) method (top row) and the proposed method (lower row) shown in 3D, 2D, and 1D space.

tein usually have similar structure to each other but are not exactly the same. Wang and Mahadevan (2009b) test their manifold alignment algorithm using protein data. For comparison, we apply the proposed method to the same database acquired from RCSB protein Data Bank Berman *et al.* (2000). The first, 10th and 21st structures of Glutaredoxin protein PDB-1G7O are picked for testing similar to the setup used by Wang and Mahadevan (2009b). Each protein molecule has 215 amino acids, which are represented as 3D points. In this experiment, we directly align two protein structures without dimensionality reduction.

The choice of k-neighbors determines the local area used for matching. We use the intersecting set of correspondences for k and $k+1$ neighbors. Figure 3.3 shows the matching accuracy with different k values. We can see that the accuracy is not sensitive to the value of k. However, the larger the value of k, the more number of features need to be calculated. Approximately setting k at 5% to 10% of the total number of samples is usually sufficient. In this experiment, we set $k = 10$ given the total number of 215 samples. The 3D, 2D, and 1D aligned protein structures are illustrated in Figure 3.4. It can be clearly seen that our method achieves more accurate alignment compared to Wang and Mahadevan (2009b).

3.4.2 FacePix Dataset

The FacePix database Little *et al.* (2005) consists of face images of 40 subjects with pose variations in yaw. For each subject, there are 181 images representing yaw angles

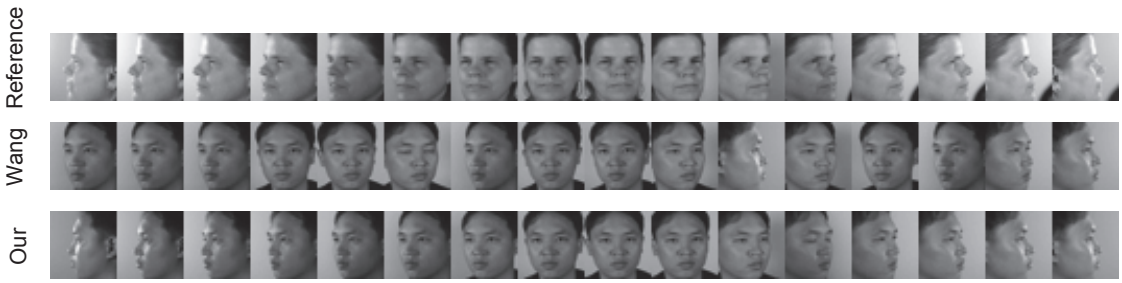


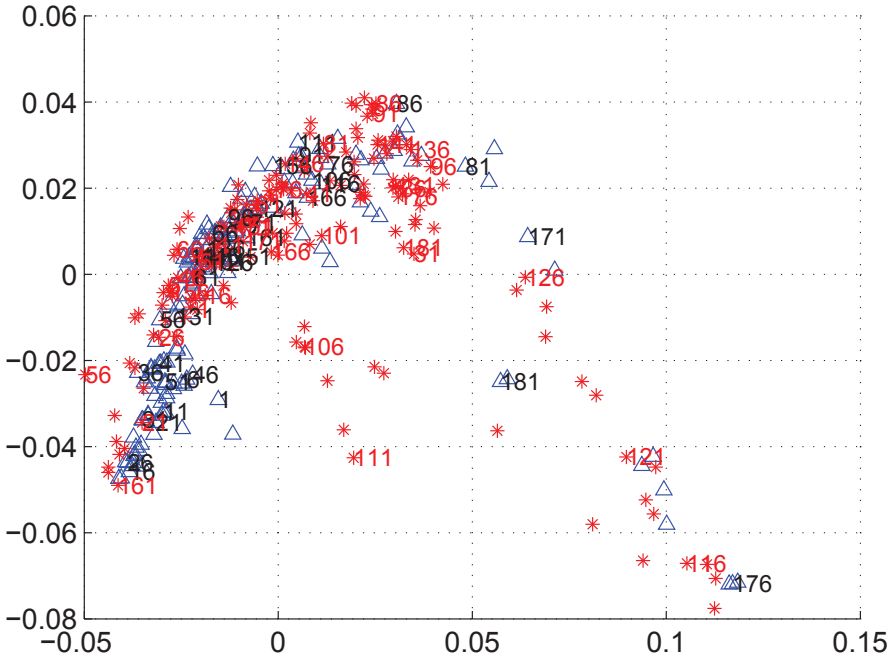
Figure 3.5: Alignment results for subject 10 and subject 11 in the FacePix database. The first row are selected images (with 10 degrees pose increment) from the reference image set (subject 10). The second row are the corresponding images for subject 11 found by Wang and Mahadevan's method Wang and Mahadevan (2009b) and the last row are the corresponding images of subject 11 found after the proposed manifold alignment. Note that our method finds better visually correct corresponding poses.

form -90° to $+90^\circ$ at 1 degree increments. Therefore, each subject has an underlying manifold of pose variations. The images are downsampled from 128×128 to 32×32 in this experiment. Five subjects (number 10 to 14) are chosen for testing manifold alignment between them. During each alignment, a test subject's images (manifold) are aligned with that of a reference subject with different identity. The nearest images, in the embedded manifold space, of the two subjects are then taken as the corresponding poses. Figure 4.8 compares alignment results of our method to Wang and Mahadevan's method Wang and Mahadevan (2009b). We can see that our method performs better visual alignment of the poses. Moreover, the proposed method obtains a more meaningful aligned joint manifold as shown in Figure 3.6b compared to Wang and Mahadevan's method Wang and Mahadevan (2009b) (see Figure 3.6a).

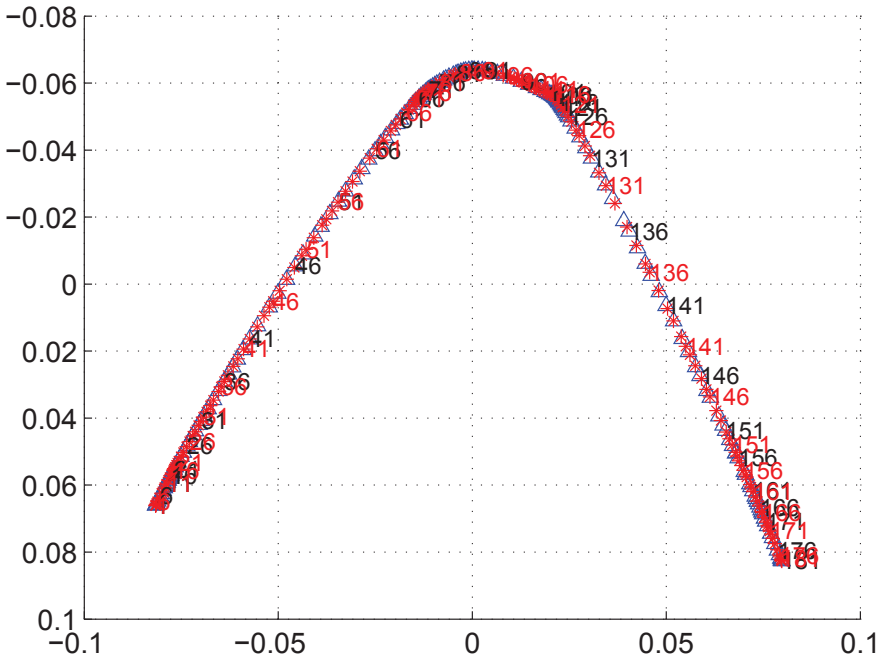
To quantify the accuracy of the proposed manifold alignment, we plot the cumulative percentage of correct poses versus the pose error. More precisely, a match is considered correct if it falls within $\pm r$ degrees pose error. Figure 3.7 compares the alignment accuracy of the proposed method with Wang and Mahadevan's method Wang and Mahadevan (2009b). Note that our method achieves significantly higher accuracy.

3.4.3 Kinect Data

The proposed manifold alignment is also tested on a dataset obtained with the Kinect sensor. Kinect simultaneously captures color (RGB) images and Depth (D) images with



(a) Result for Wang and Mahadevan (2009b)



(b) Proposed manifold alignment

Figure 3.6: Alignment of FacePix images of different subjects using (a) Wang and Mahadevan (2009b) and (b) the proposed method.(c)

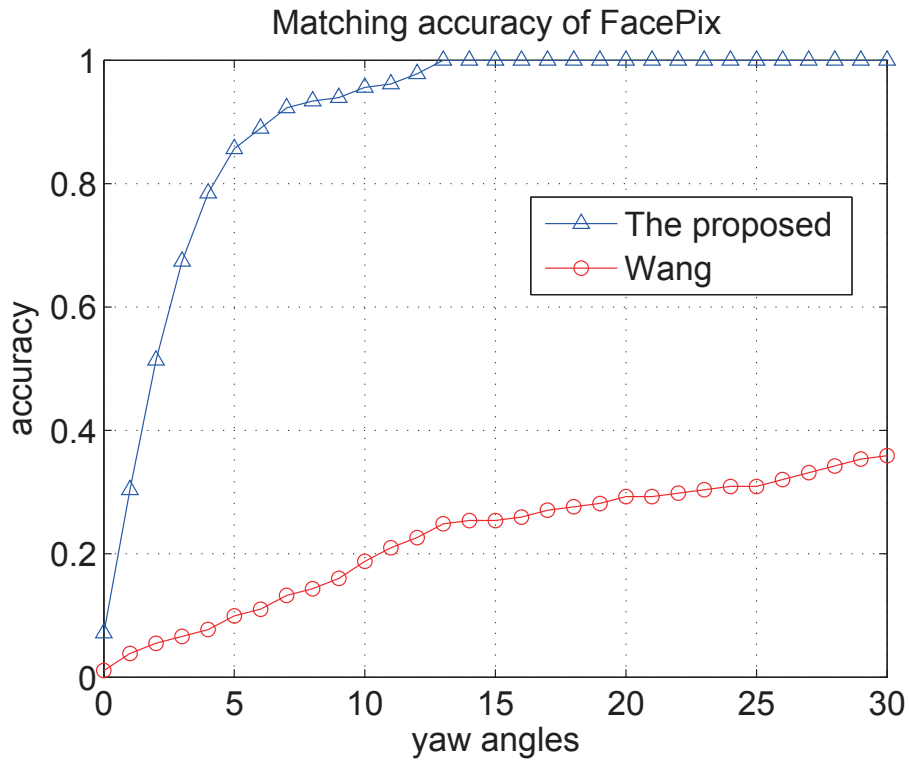


Figure 3.7: Accuracy curve for the average pose accuracy of aligning all possible combination pairs of image sets.

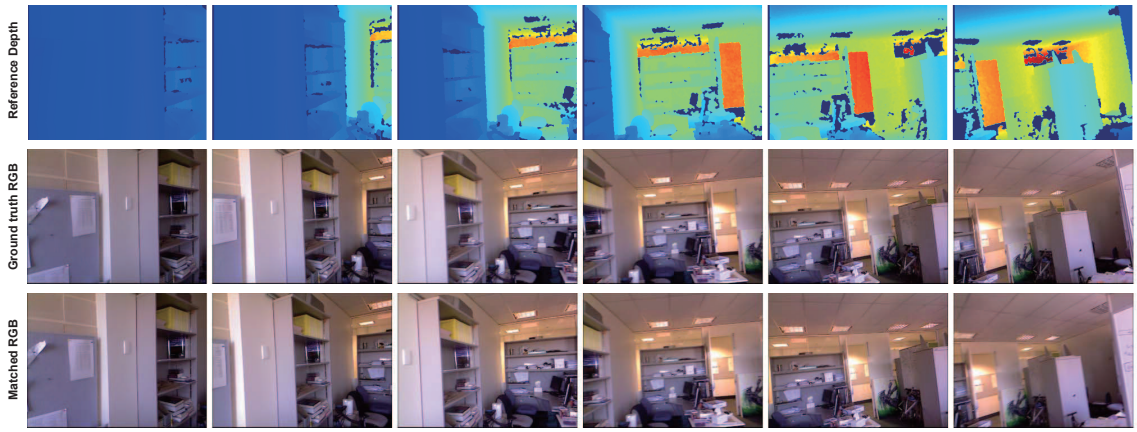


Figure 3.8: The alignment of color image and depth data from Kinect. The depth data is set as reference set , and color image is matched to depth data using the proposed method.

inherent correspondences which is used as ground truth. The two kinds of data (RGB and Depth) are totally different but represent the same scene. RGB-D video data is recorded in an office environment with a moving Kinect camera and the RGB and depth images are resized to 60×80. The depth dataset is set as the reference and the colour images are aligned to them using the proposed method. In this experiment, the parameter for the nonlinear dimensionality reduction is set to $k_n = 10$ and in correspondence estimation k is set to 15. It should be emphasized here that manifold alignment is now performed on completely different datasets since RGB represents the reflectance of the scene and depth represents the shape of the scene. The alignment results are compared against the ground truth. Some sample aligned frames are shown in Figure 3.8. Given the challenging nature of this alignment problem, highly satisfactory results have been achieved by the proposed method with only minor visual differences. Figure 3.9 demonstrates the quantitative accuracy of the proposed method on Kinect data. As can be seen, the proposed approach can achieve the matching accuracy exceeds 95%, when tolerance time equals to 300ms. Figure 3.10a and Figure 3.10b show the alignments after one iteration and after 10 iterations. It is clear that the iterative process significantly improves the final alignment.

3.4.4 Computational Time Cost

Table 3.1 lists the time required for the alignment of two manifolds from the three datasets. Timing is reported for Matlab implementations on a 3.2GHz machine with 4GB RAM. Notice that the proposed method outperforms Wang and Mahadevan's method by a significant margin.

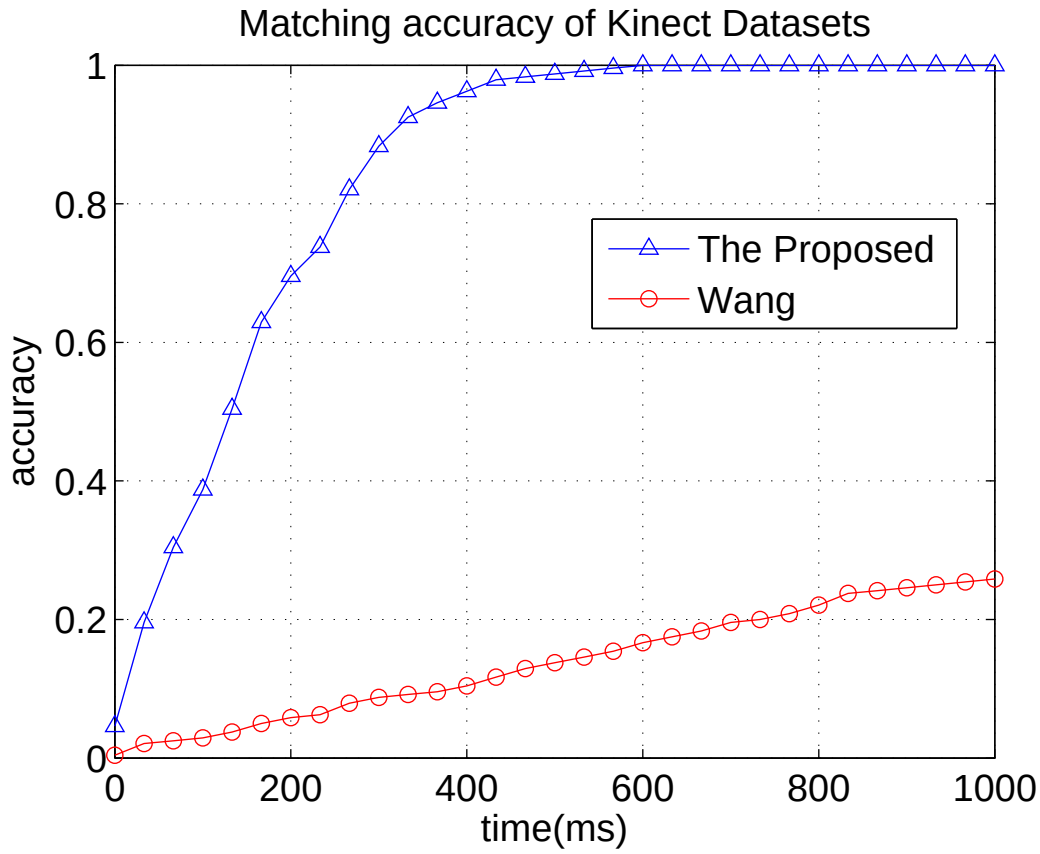
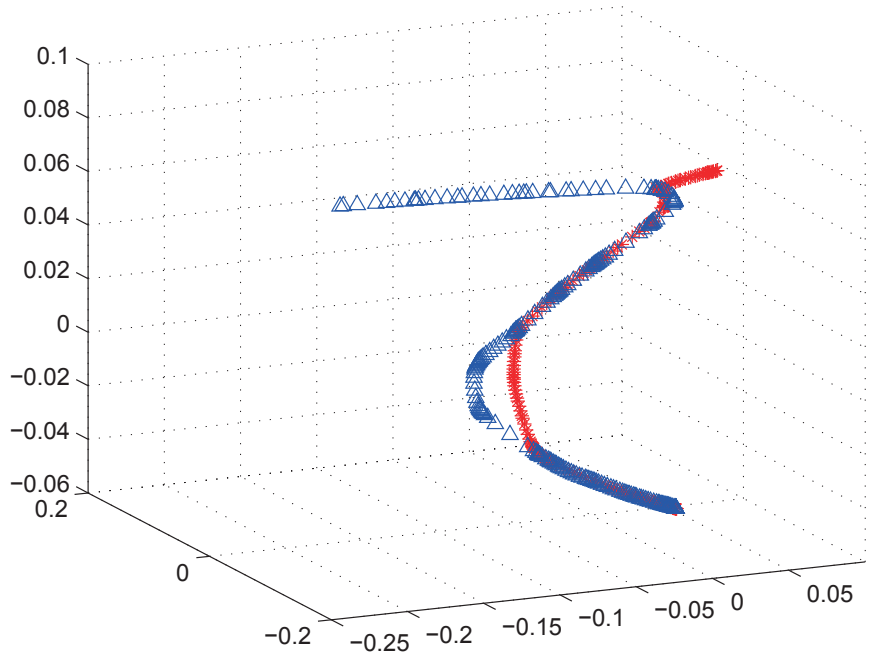


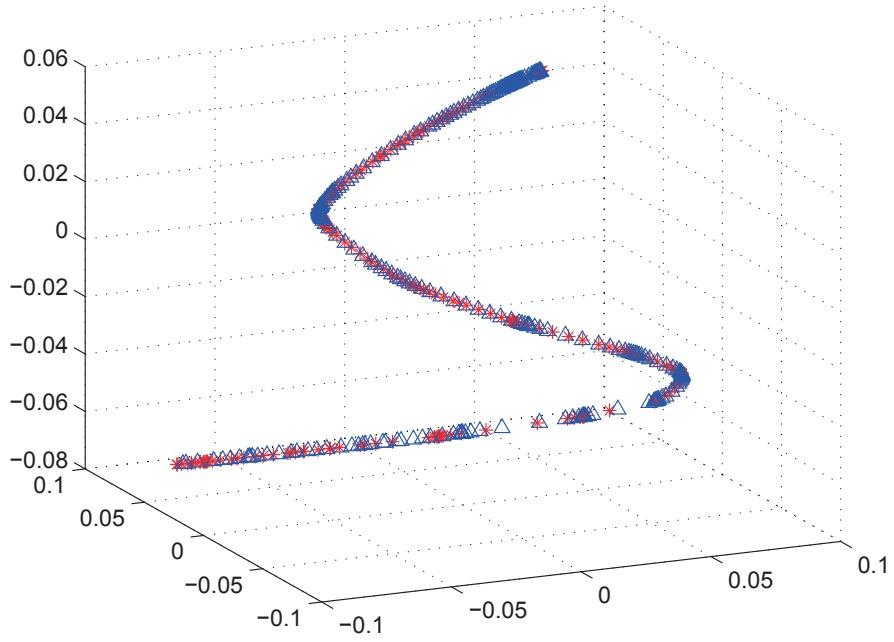
Figure 3.9: Cumulative accuracy for RGB image to Depth image alignment with respect to time.

Table 3.1: Alignment time comparison using a Matlab implementation on a 3.2GHz machine with 4GB RAM.

Database	Dim.	# of samples	Times (sec)	
			Wang and Mahadevan (2009b)	proposed
Protein	3	215	581	23
FacePix	1024	181	439	7
Kinect	4800*(3+1)(RGB+Depth)	300	523	15



(a) First iteration



(b) After 10 iterations

Figure 3.10: The joint manifolds of Kinect Data in different period of iteration. After iterations, the alignment results are perform better.

3.5 Chapter Summary

This chapter first reviews the main idea of manifold alignment and some related algorithms. Then we propose a novel method for unsupervised manifold alignment. Using feature histograms for characterizing the local manifold geometry and a robust matching algorithm, accurate correspondences are estimated between two manifolds. These initial correspondences are refined under an iterative optimization framework under manifold structure preserving constraints. A joint manifold is achieved with this algorithm. The proposed manifold alignment algorithm was applied to three different types of datasets and achieved significantly superior results when compared to existing state-of-the-art algorithms. The common assumption that the reference set is a complete set that must include all the variations of testing sets is followed here.

In the next chapter, an approach is proposed to remove the assumption that the reference set is necessary to be complete. It can achieve better alignment accuracy and more freedom of applicability. The main idea is still following the criteria of manifold alignment with some improvement of correspondence modeling.

Chapter 4

Robust Image Set Manifold Alignment

In Chapter 3, we discussed the manifold alignment technique using RANSAC and iterative binary optimization. Similar to Wang and Mahadevan (2009b); Cui *et al.* (2012), all these approaches aim to formulate the correspondence and embedding into feature space as a single optimization problem. Due to the non-convex nature of the optimization function, it is iteratively solved by first initializing with some rough correspondences, for example, using prior knowledge. All of these methods directly use a binary matrix to model the correspondence matrix and they are sensitive to the initial binary correspondences. The two step alignment structure is hard to integrate the third criteria into a single objective function. To convert the numerical feature similarity to the initial binary matrix will also lose some useful information about the correspondence. On the other hand, in Chapter 3 and previous works of manifold alignment Cui *et al.* (2012); Pei *et al.* (2012), there is a critical assumption that all of test data has a correspondence in reference set. That means the reference set should be a complete set including all variations appeared in query set. To overcome these two limitations, we further improve the manifold alignment in this chapter using the Softassign technique.

In this chapter we aim to remove the tight constraint, and allow the correspondence to exist for part of both datasets. The advantages in this chapter are summarized as follows:

- We do not need to estimate an initial correspondence. Also a more elegant way to use features is proposed and can simplify the whole procedure by avoiding usage of Sample Consensus method.
- Initial correspondence assumption in last chapter is removed.
- Higher alignment accuracy is achieved comparing to the state of the art methods.

In detail, the technique contributions are summarized. An initial feature matching is automatically established based on the local structures of the manifolds that are invariant

to transformations and robust to sampling. A local reconstruction constraint is imposed to preserve intra-manifold structures. To remove the constraint on the correspondence assumption, we model an extended correspondence matrix with extra space to handle non-correspondence data. This matrix is employed to measure the inter-manifold distance, which is supposed to be minimized. An iterative optimization is employed to refine the alignment. Additionally, a joint manifold structure with a correspondence matrix is learned which is useful for further information transfer between the datasets. The effectiveness of the proposed algorithm is demonstrated on facial images of different subjects. Comparison with the manifold alignment proposed by Wang and Mahadevan (2009b) and method in Chapter 3 shows this method is more accurate.

4.1 Local Shape Descriptors

For unsupervised manifold alignment, there is no prior information on correspondences. It is only possible to estimate relationship between manifolds based on structure and certain feature similarities in manifolds. One straightforward way is to directly match their features in the original space. However, this is very hard if the two datasets are from different sources or their dimensions are not equal. An alternate way is to find the similarity between the local structures of the two manifolds. In this section, we analyze the Local Histogram Feature proposed in Section 3.1.1, discuss how to simplify and improve the features, and further improve the procedure for accurate alignment. A manifold learning step is required to project data down to a feature space before feature extraction in the last chapter. Actually this feature can be directly applied on the original data, if the dataset contains the same type of data.

4.1.1 Improved Local Histogram Feature

The main idea of Local Histogram Feature is to represent the manifold structure and define features based on measurements that can encode its neighborhood's geometric properties.

Now consider a point p_i in P_x. As shown in Figure 4.1, all of its k-neighbors p_k^i in the Euclidean space are selected and organized in $R_i = [p_i, p_1^i, \cdots, p_k^i]$, termed as the *$k$-neighbors*. The normal vector n_i at point p_i can be approximated by the normal of the best fit plane to k-neighbors using PCA. It is obtained by selecting the eigenvector corresponding to the smallest eigenvalue.

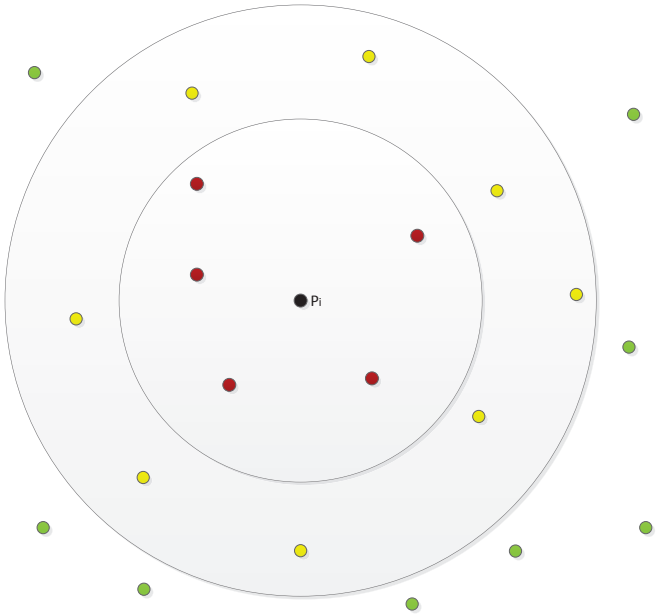


Figure 4.1: Example of a query point p_i and its neighbors. The query point and its neighbors are fully connected in three different levels.

For each pair of points p_s and p_t in the k-neighbors R of a point p, and their corresponding normals n_s and n_t (calculated from their respective k-neighbors R_s and R_t), we define a unit difference vector between them:

$$v = \frac{p_t - p_s}{\|p_t - p_s\|} \ . \tag{4.1}$$

Similarly, we can define some angular features for a pair as

$$f_1 = \max(n_t \cdot n_s, -n_t \cdot n_s) \tag{4.2}$$

$$f_2 = |\arccos(\max(n_t \cdot v, -n_t \cdot v)) - \arccos(\max(n_s \cdot v, -n_s \cdot v))| \tag{4.3}$$

$$f_3 = \|p_t - p_s\| \tag{4.4}$$

In Section 3.1.1, besides the three mentioned features, there is another feature representing the angles between n_s and n_t. Since such operation has used the cross product of vectors, in such case data needs to project into a 3D feature space before feature extraction. Moreover, since such features have a similar function with f_1 and f_2, its exclusion in this chapter would cause no significant performance decrease in terms of alignment performance. More importantly, removal of this feature can simplify the operation.

The three features measure the curvature and the angles between the normals and the different vectors. f_1 is dot product between two unit vectors. They are in fact the cosine of the angles between these vectors. The maximum operation is performed to select only the acute angles. Similarly, f_2 shows the difference of the two angles between the two normals and f_3 is the difference vector. We quantize each feature into two bins to make a 9 dimensional histogram of the curvature features. Since the features are based on mutual angles, it can be proven that they are rotation invariant. Moreover, these features are still comparable when we choose different values of k for the two manifolds.

4.1.2 Multi-scales

It can be seen in discussions from Section 4.1.1 that the three features are important to describe local structures and the choice of the radius is important. Though we can use the fixed radius sphere strategy to choose the k-neighbors for local features, it is somehow hard to choose the best radius for extracted features. Take 3D objects for example, the features for small corners show better discrimination with a small radius rather than with a large radius, as these festers only have large gradient magnitude difference in a small local area. While small local range cannot distinguish points on a nearly flat surface. Therefore, a fixed radius for selecting neighborhood cannot suit for every case. In order to guarantee the robustness and reliability of these features, we apply the multi-scale radius to select features in our experiments. Three different radiuses are chosen, and each radius generates a 9 dimensional histogram. Eventually, we combine all histograms into a 27 dimensional feature to build the initial correspondence matrix for iterative alignment.

4.2 Robust Manifold Alignment Approach

Based on the improved features in the last section, we start to design the robust manifold alignment algorithm. With the extended row and column for the correspondence matrix, we still need to formulate the objective function following the criteria of manifold alignment. The key point is to convert this function into a solvable optimization problem. In this section, the objective function is effectively solved using the Softassign technique. Next we will describe the detail.

Given a reference dataset $X = [x_1, x_2, \ldots, x_{N_x}] \in R^{d_x \times N_x}$ and a test dataset $Y = [y_1, y_2, \ldots, y_{N_y}] \in R^{d_y \times N_y}$, where the columns x_i and y_i are the samples, d_x and d_y are dimensions of datasets, and N_x and N_y are the number of samples in X and Y respectively.

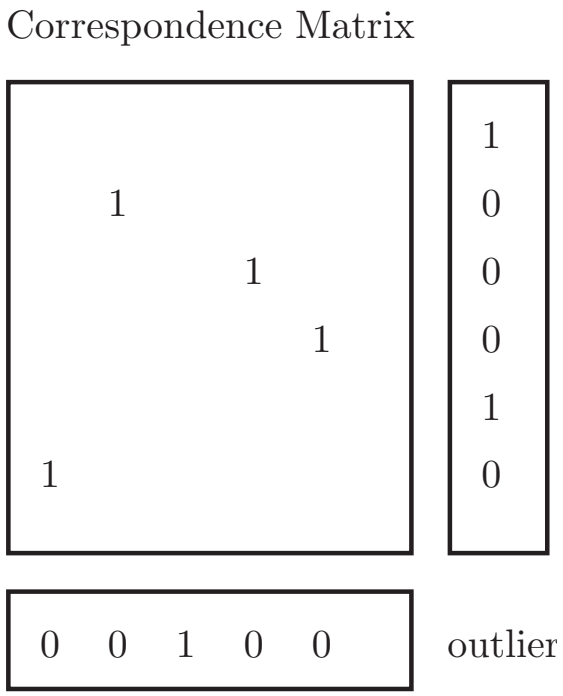


Figure 4.2: An example of the binary correspondence matrix. The inner sub-matrix defines the correspondence. the extra row and column represent the possible outliers.

To measure the distance between two manifolds, a correspondence between their elements should be established first. For such purpose, we can define an $(N_x + 1) \times (N_y + 1)$ binary correspondence matrix $\hat{W}^{xy}$ such that $\sum_{i=1}^{N_x+1} \hat{W}_{i,j}^{x,y} = 1$ for $j \in 1, 2, \cdots, N_y$, $\sum_{j=1}^{N_y+1} \hat{W}_{i,j}^{x,y} = 1$ for $i \in 1, 2, \cdots, N_x$, and $\hat{W}_{i,j}^{x,y} \in \{0, 1\}$. The summation constraints guarantee that each row and column of $\hat{W}^{(x,y)}$ can contain the value 1 at only one location. This means the correspondence is one to one for two elements between the two sets. The extra $(N_x + 1)$th row and $(N_y + 1)$th column of $\hat{W}^{x,y}$ makes the constraints always being satisfied even if there exist outliers. An example of the correspondence matrix is given in Figure 4.2. If some elements in the extra row and column are one that means those elements are outliers with no correspondence in other set. The remaining $N_x \times N_y$ partial correspondence matrix are defined as $W^{x,y}$. In comparison with the correspondence matrix defined in Section 3.1, the extra row and column give the ability to handle the outliers and missing correspondence.

4.2.1 The Energy Function

Assume that there are two projection functions f_x and f_y that map X, Y to $F^{(x)} \in R^{d \times N_x}$, and $F^{(y)} \in R^{d \times N_y}$ respectively, in an intrinsic space with dimension d. Note that $F^{(x)}$

and $F^{(y)}$ are the low dimensional representations of X and Y, and elements x_i and y_i are mapped to $F_i^{(x)}$ and $F_i^{(y)}$ respectively.

We now need to define an energy function for manifold alignment. As our aim is to find projection functions f_x and f_y that can project the data into a joint low dimensional space that corresponding data points of two manifolds are matched as closely as possible while a reasonable number of points can be rejected as outliers. Follow the objective function (3.9) we formulate a new energy function that with two more terms, which represent feature matching cost and the outliers controller.

$$\begin{aligned}\min_{W^{xy},F} E(W^{xy},F) = \quad & \min_{W^{xy},F} \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} \|F_i^{(x)} - F_j^{(y)}\|^2 W_{i,j}^{xy} \\ & +\lambda_1[J_W(F^{(x)}) + J_W(F^{(y)})] \\ & +\lambda_2 tr(D^T W^{xy}) + \lambda_3 |\hat{W}^{xy}|_1 \end{aligned} \tag{4.5}$$

subject to

$$\begin{gathered}\textstyle\sum_{i=1}^{N_x+1} \hat{W}_{i,j}^{x,y} = 1, j \in 1,2,\cdots,N_y, \\ \textstyle\sum_{j=1}^{N_y+1} \hat{W}_{i,j}^{x,y} = 1, i \in 1,2,\cdots,N_x, \\ W_{i,j}^{x,y} \in \{0,1\}.\end{gathered}$$

The first term represents the distance between two manifolds in the low dimensional space which is already explained in Section 3.2.2. The second term represents the reconstruction error of the locally generated models for each dataset, which is already defined in Section 3.2.3. The third term is the feature matching cost depending on D. If two points are corresponded, there is only one 1 appears in middle part of W; If two points are misaligned, there are two 1 in extra row and column respectively. The last term is the controller term preventing rejection of too many points as outliers. The parameters λ_1, λ_2 and λ_3 are weightings which are used to balance these terms. The following subsections provide more details.

4.2.1.1 Feature Matching Error

In order to measure the reliability of correspondences in term of feature matching criteria, we introduce the third term of (4.5) as the feature matching cost. Unlike the first term which just aims to keep two manifold close, this term constraints the matching with limited effort. If the two manifold are properly aligned, the structure feature of corresponding points should be similar. This term is manage to minimize the sum of feature distance. This is another improvement to obtain the robustness. Usually, the feature distance matrix

D is used on the features defined in Section 4.1. Of course, other features can be imported into this optimization. More than one feature can be easily combined together by using different weightings on each feature, i.e.,

$$D = \theta_1 D_1 + \theta_2 D_2 + ... + \theta_n D_n$$

For face images the specific features are discussed in Experiment Section 4.4.4.

4.3 Efficient Alignment Algorithm

In Section 4.2.1, we defined the problem clearly and now we need to develop an efficient algorithm to solve it. One can see that the alignment objective function in (4.5) consists of two alternative optimization problems: a linear assignment discrete problem on the correspondence W^{xy} and a least-squares problem on the low dimensional features F^x and F^y. Both problems are solvable when we consider them separately. However, their combination makes the nonlinear manifold alignment problem difficult. To solve the whole problem, it is natural to take an alternating algorithm approach.

At each step of the alternating approach, solving the binary one-to-one correspondences is very complex (which is NP-Complete). Consequently, we adopt the softassign Rangarajan *et al.* (1997) technique here and relax the binary correspondence $\hat{W}^{xy}$ to be a continuous valued matrix in an interval, while still keeping the row and column summation constraints. From an optimization point of view, this fuzziness makes the energy function easier to solve because the correspondences in relaxed form are able to improve gradually and continuously during the optimization process without jumping around the binary points.

Deterministic Annealing Rose (1998) is another useful technique for combinational optimization problem, by adding an entropy term $\beta \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} w_{ij} \log w_{ij}$ in (4.5). The parameter β acts as the temperature and models some thermal agitation. The higher temperatures, the entropy term would force the correspondence to be more fuzzy.

In consideration of these techniques, we then transform the original binary problem (4.5) into the problem of minimizing the following fuzzy energy function:

$$\begin{aligned} E(W,F) = & \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \|F_i^{(x)} - F_j^{(y)}\|^2 W_{i,j}^{xy} + \lambda_1 [J_W(F^{(x)}) + J_W(F^{(y)})] \\ & + \lambda_2 tr(D^T W^{xy}) + \lambda_3 |W^{xy}|_1 + \beta \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} w_{ij} \log w_{ij} \end{aligned} \tag{4.6}$$

where w_{ij} still satisfies $\sum_{i=1}^{N_x+1} \hat{W}_{i,j}^{(x,y)} = 1, j \in \{1, 2, \cdots, N_y\}$ and $\sum_{j=1}^{N_y+1} \hat{W}_{i,j}^{(x,y)} = 1, i \in \{1, 2, \cdots, N_x\}$, but $w_{ij} \in [0, 1]$. The parameter λ_1 weighs the reconstruct cost for each manifold, and can be seen as the deformation of manifolds. The parameter λ_2 balances the prior feature and the alignment cost between manifolds. The parameter β controls the fuzziness of the correspondence matrix. When β reaches zero, the fuzzy correspondence $\hat{W}^{xy}$ becomes binary.

4.3.1 Feature Updates

Following the above discussion, we start discussing about the two subproblems. The first subproblem updates the low dimensional feature F when we fix the correspondence W^{xy}. To build the subproblem, the terms independent on F are dropped. Now the function can be formulated as

$$\begin{aligned} \arg\min_F J &= \arg\min_F tr(FLF^T) \\ s.t. \quad & FF^T = I \ . \end{aligned} \tag{4.7}$$

Keep in mind that we impose the unit variance $FF^T = I$ constraint to guarantee a unique solution. It follows that F can be computed as the eigenvectors corresponding to the second to $(d+1)$th smallest eigenvalues of the sparse matrix

$$L = \begin{pmatrix} B_x + \Omega^x & -W^{xy} \\ -W^{yx} & B_y + \Omega^y \end{pmatrix} \tag{4.8}$$

where the B_x and B_y are defined as the reconstruction term in (3.22). Ω^x and Ω^y are diagonal matrix derived from (3.10).

4.3.2 Correspondence Updates

After estimating the aligned manifold F, The second subproblem is to update the correspondence matrix W^{xy}. To solve it, we apply the Deterministic Annealing technique and the update strategy is chosen as follows:

$$w_{ij} = \frac{1}{\beta} e^{-\frac{\|F_i^{(x)} - F_j^{(y)}\|^2 + \lambda_2 D_{ij}^2 - \lambda_3}{2\beta}} \tag{4.9}$$

Algorithm 2: Robust Manifold Alignment Algorithm using Softassign

Input: X and Y: two datasets

Output:

$F^{(x)}$ and $F^{(y)}$: the embedding features of X and Y in low dimensional space.

W : the correspondence Matrix

```
W_0 ← Init(X, Y);
begin
    repeat
        Update the Feature F using (4.7). Update the correspondence matrix W using
         (4.9).
        repeat
            Normalize W using (4.10) and (4.11).
        until W converge;
    until E(W, F) converge;
    Decrease β and λ.
end
```

for the points $i = 1, 2, \cdots, N_x$ and $j = 1, 2, \cdots, N_y$. We run iterative row and column normalization to make sure the sum constraints are satisfied:

$$w_{ij} = \frac{w_{ij}}{\sum_{i=1}^{N_x+1} w_{ij}}, j = 1, 2, \cdots, N_y, \tag{4.10}$$

$$w_{ij} = \frac{w_{ij}}{\sum_{j=1}^{N_y+1} w_{ij}}, i = 1, 2, \cdots, N_x. \tag{4.11}$$

When the temperature parameter β approaches zero, the distances between different w_{ij} are increasing. The iterative normalization (4.10) and (4.11) will bring the sparseness into the correspondence matrix W^{xy} with temperature decreasing.

4.3.3 Robust Manifold Alignment Algorithm

After previous discussion, we can summarize and present Algorithm 2. First, the initial W_0 is obtained from defined feature from Section 4.1. Each element W_{ij}^0 indicates the feature distance between of X_i and Y_j. In iterative alignment, the whole update process is controlled by the temperature parameter β. According to a linear annealing schedule, we update β using a certain annealing rate r that is $\beta^{new} = \beta^{old} \cdot r$. The higher temperature in the beginning makes the correspondence fuzzy, so that the alignment is roughly matching the overall structure. When temperature decrease closes to zero the correspondence matrix will converge to binary, that can been seen as a refinement of the previous alignment in high

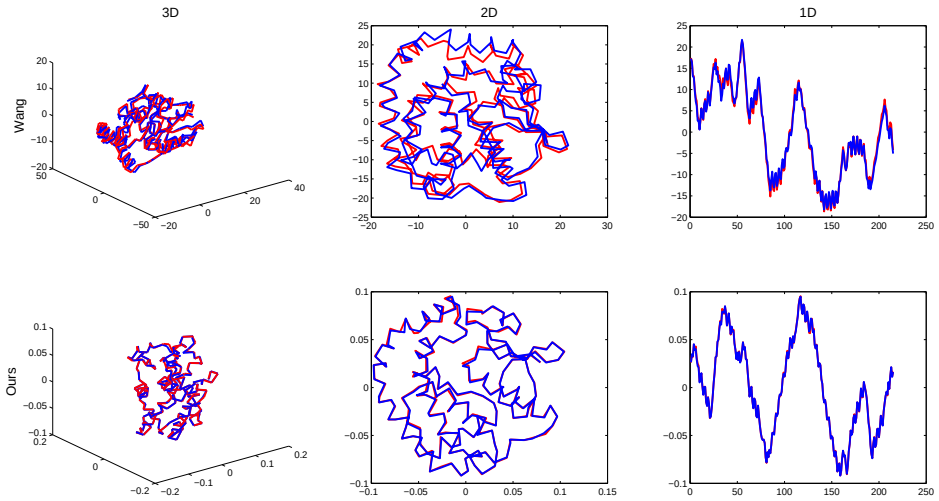
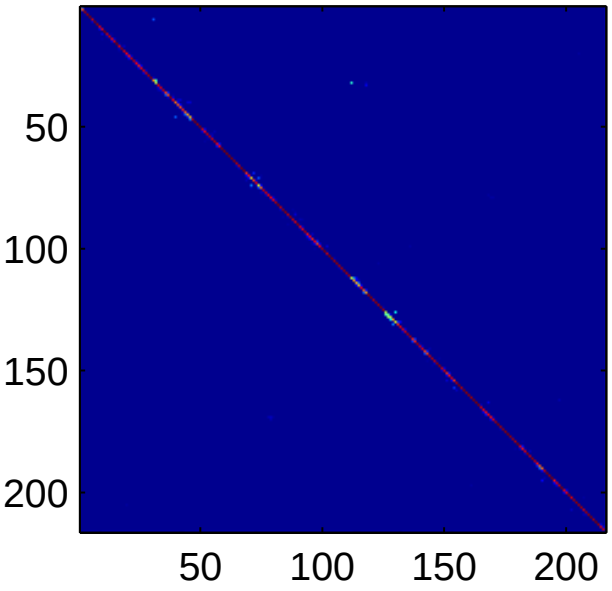


Figure 4.3: Alignment of protein structures using Wang and Mahadevan's method Wang and Mahadevan (2009b) (top row) and the proposed method (lower row) shown in 3D, 2D, and 1D space.

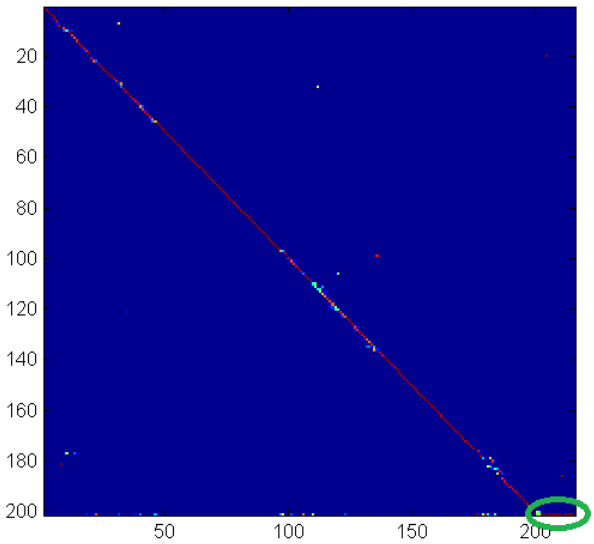
temperature. Large value of λ will limit the range of deformation of manifold structure. Similarly, we apply annealing scheme on λ. The basic idea is that more global and rigid alignment should be done first and then refine the local accuracy.

4.4 Experiments and Discussion

In this section, several experiments are conducted to demonstrate the effectiveness of the proposed manifold alignment. The datasets are used including protein bioinformatics data Wang and Mahadevan (2009b), 3D face data and face images with different variations. Wang and Mahadevan's method Wang and Mahadevan (2009b) and our previous work Manifold Alignment using RANSAC in Chapter 3 are applied to these datasets for performance comparisons. For face image datasets, the results of the proposed method are compared with the ground truth.



(a) visualization for Correspondence matrix of Protein alignment result.



(b) visualization for Correspondence matrix of Protein alignment with missing alignment. the green area is the outliers indicator in the extend row.

Figure 4.4: The Correspondence matrix of Protein alignment. a) result for fully matched alignment. b) incomplete reference set.

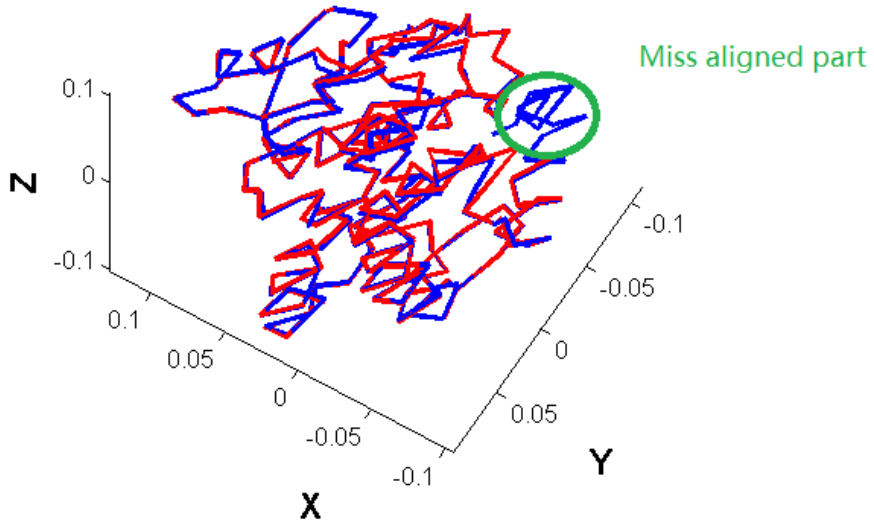


Figure 4.5: Alignment of protein structures with the missing aligned part.

4.4.1 Protein Data

Protein structure has been used for a simple test in Section 3.4.1. The experiment set-up is same as previous one. We apply the robust manifold alignment method to the database acquired from RCSB protein Data Bank Berman *et al.* (2000). The first, 10th and 21st structures of Glutaredoxin protein PDB-1G7O are picked for testing. Each protein molecule has 215 amino acids, which is represented as 3D points.

A correspondence matrix of two proteins's alignment result is shown in Figure 4.4a. Actually, 100% matching accuracy is achieved in this alignment with 90.7% accuracy initial direct feature matching. The 3D, 2D, and 1D aligned protein structures are illustrated in Figure 4.3. It can be seen clearly that our method achieves more accurate alignments compared to Wang and Mahadevan (2009b). Comparing to result in Section 3.4.1 our two methods can achieve 100% accuracy rate.

To show the ability to handle outliers, we remove 15 amino acids in reference set. Then some part of these two proteins will be misaligned, and our method can identify this part. The alignment result is shown in Figure 4.5. We can easily find the outliers between two proteins from the correspondence matrix (shown in Figure 4.4b).

4.4.2 BU-3DFE Face Data

In this section, we select first five subjects to build ten pairs of sets from BU-3DFE database Yin *et al.* (2008). This data is used to verify the effectiveness of the proposed manifold alignment method on the high dimensional data with low dimensional structure. Each face data is downsized by sampling the number of points to 20 percent of the total points, in order to reduce the computational time and memory consuming in the experiment. The selected 3D face points clouds are rotated to random directions and projected into 20 dimension space using two random projection matrix. In addition, some noises are added into the high dimensional data (20 dimension). This pre-processing aims to simulate the conditions that two data cannot align directly and need to be aligned in a low dimension space. ICP Besl and McKay (1992) have no capability to apply dimensional reduction and registration simultaneously. ICP is applied on original 3D face data as a reference of 3D registration result in Figure 4.7a. The two data are in different space, but have similar intrinsic structure (3D face). Our manifold alignment technique can handle this non-comparable data. One example of aligned results of our proposed method is shown in Figure 4.7b. We quantify the align accuracy results in Table 4.1 using the normalized leat-square Euclidean distance between the nearest correspondent points. Since the transformation of our method is not rigid, our method can achieve better quantity result even the task is harder in comparison with ICP.

Table 4.1: Normalized least-square distance between nearest neighbor between two aligned models.

Methods	ICP	The proposed approach
Normalized least-square distance	102.86	43.66

4.4.3 FacePix Database

The FacePix database Little *et al.* (2005) consists of face images of 40 subjects with pose variations in yaw. Its small part has been used in Section 3.4.2. We use all the 40 subjects in this experiment. For each subject, there are 181 images representing yaw angles from -90° to $+90^\circ$ at 1 degree increments. Therefore, each subject has an underlying manifold of pose variations. The images are downsampled from 128×128 to 32×32 in this experiment. 200 pairs of sets are randomly chosen for testing manifold alignment between them. During each alignment, the image in testing set is aligned with a reference set with different identities. Figure 4.8 compares alignment results of our method to Wang and

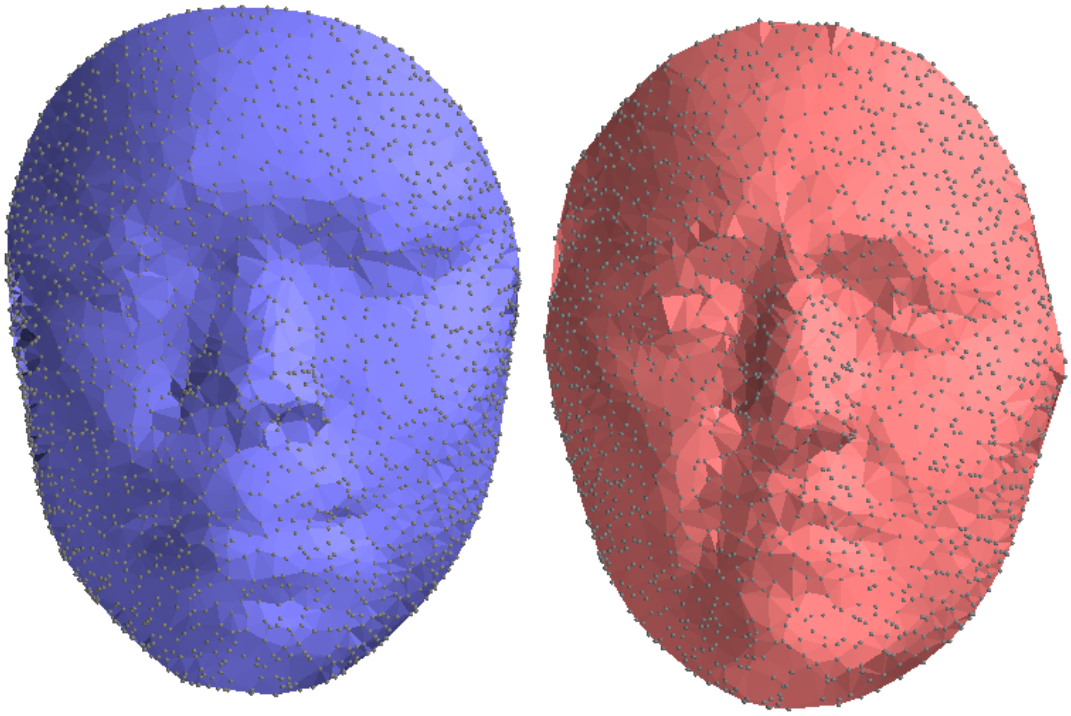
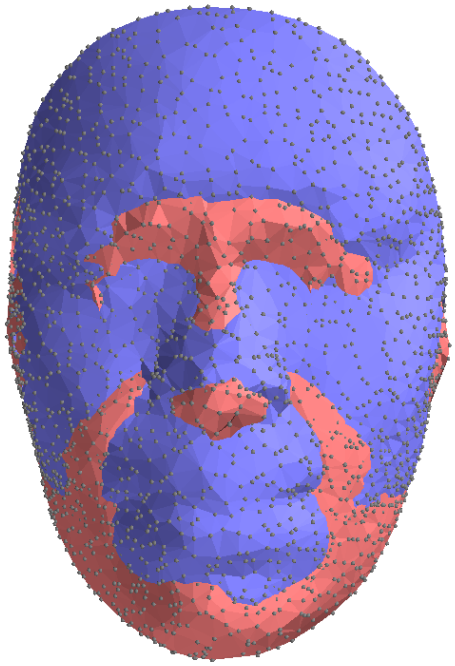
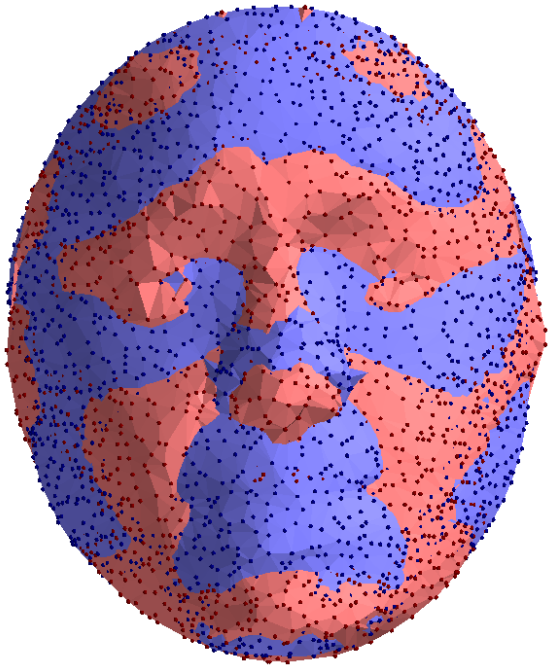


Figure 4.6: Example from BU-3DFE database.



(a) Aligned result using ICP



(b) Aligned result using Algorithm 1

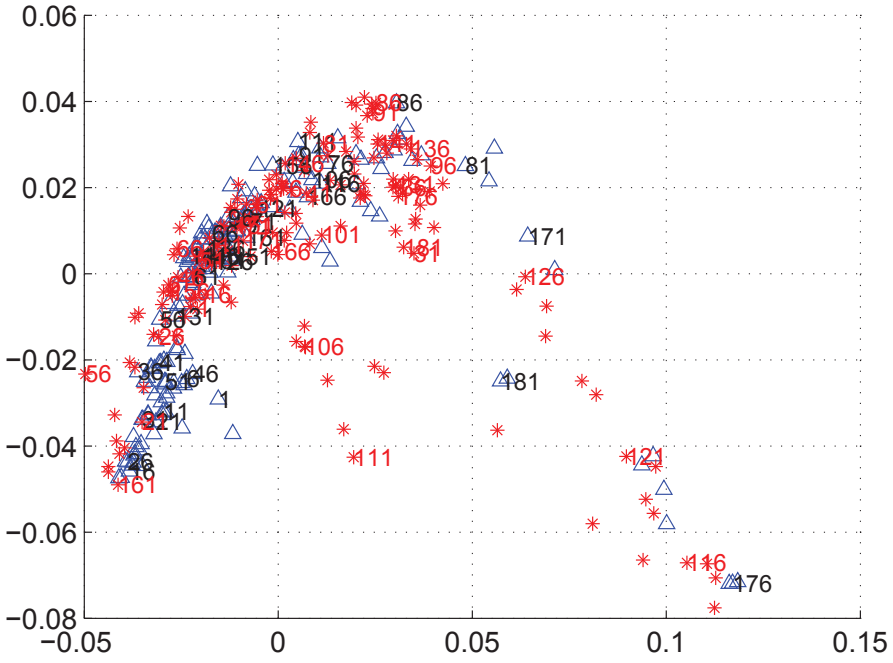
Figure 4.7: Comparison of aligned methods on a pair of 3D face.

Figure 4.8: Alignment results for subject 10 and subject 11 in the FacePix database. The first row are selected images (with 10 degrees pose increment) from the reference image set (subject 10). The last row are the corresponding images of subject 11 found after the proposed manifold alignment. Note that the MA-S method finds better visually correct corresponding poses.

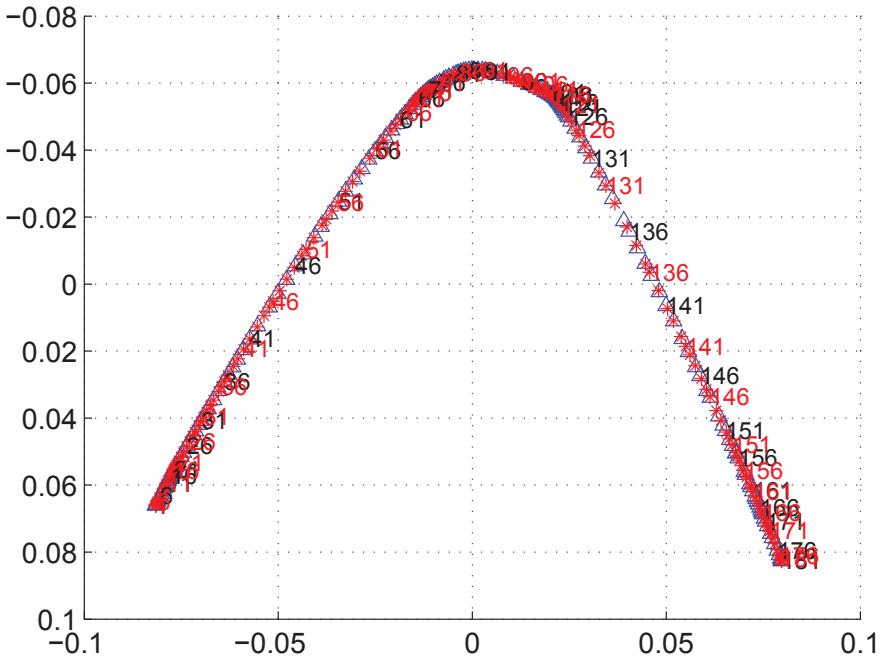
Mahadevan (2009b). We also compare to the matching method Manifold Alignment using RANSAC (MA-R) in Chapter 3, HoG feature (DM-HoG) Dalal and Triggs (2005) and Canonical Correlation Analysis (DM-CC) Kim *et al.* (2007). HoG and Canonical Correlation feature are often used for pose estimation Huang *et al.* (2011). In this experiment, directly matching using these two features is used for comparison.We can see that our method achieves better visual alignment of the poses. Moreover, the proposed method obtains a more meaningful aligned joint manifold as shown in Figure 4.9b compared to Wang and Mahadevan (2009b) (see Figure 4.9a).

To quantify the accuracy of the proposed manifold alignment, we plot the cumulative percentage of correct poses versus the pose error. More precisely, a match is considered correct if it falls within $\pm r$ degrees pose error. Figure 4.10 compares the alignment accuracy of the proposed method with Wang's method Wang and Mahadevan (2009b) and MA-R method. Note that our method achieves significantly higher accuracy.

Figure 4.9: Alignment of FacePix images of different subjects using (a) Wang and Mahadevan's method Wang and Mahadevan (2009b) and (b) the proposed method.(c) Accuracy curve for the average pose accuracy of aligning C_2^5 pairs of image sets.



(a) Result for Wang & MahadevanWang and Mahadevan (2009b)



(b) Manifold Alignment using Softassign (MA-S)

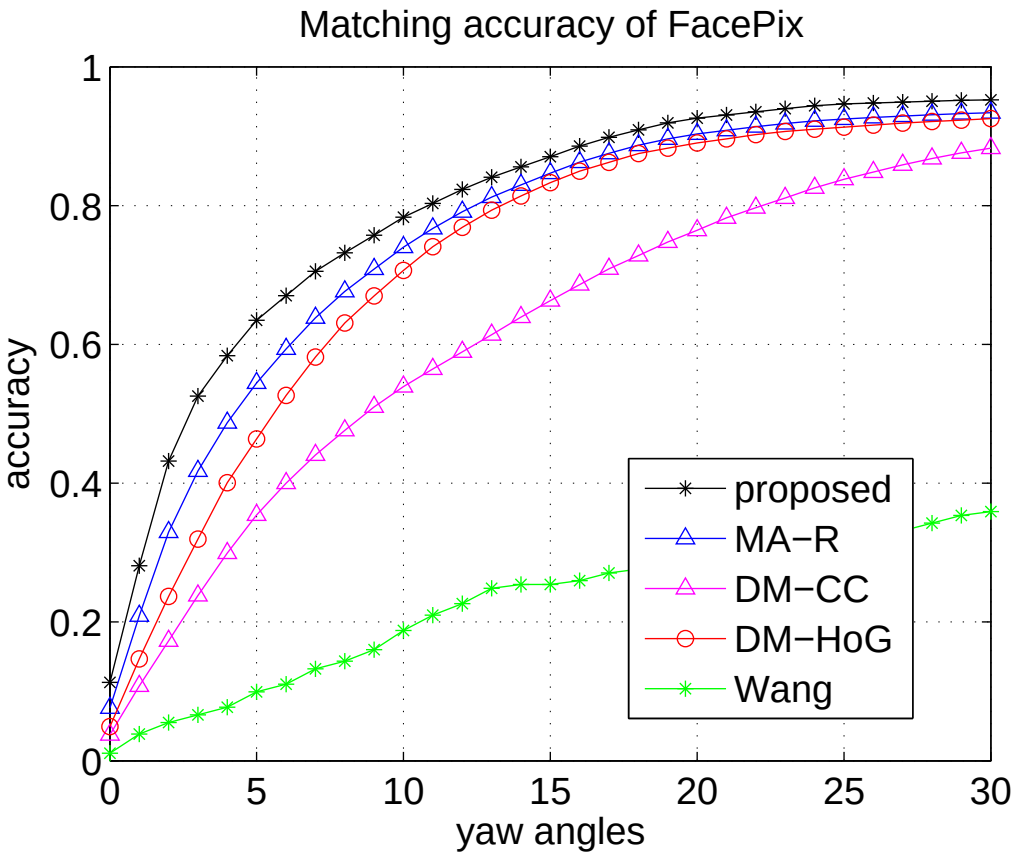


Figure 4.10: Alignment accuracy on FacePix Database

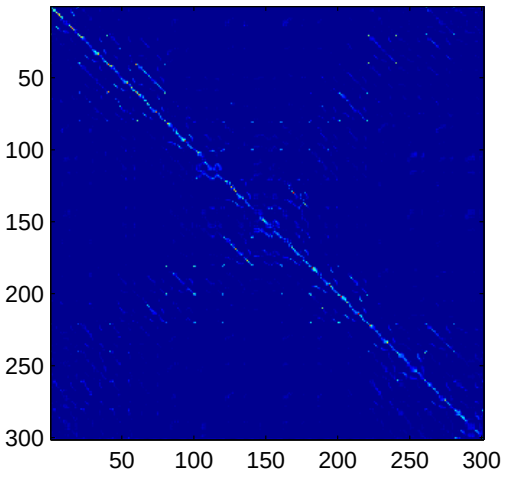
Figure 4.11: Alignment results for subject 10 and subject 12 in the multipie database. The first row are random selected images from the reference image set (subject 10). The second row are the corresponding images for subject 12 found by Wang and Mahadevan's method Wang and Mahadevan (2009b) and the third row are the corresponding images of subject 12 found after the proposed manifold alignment. The last row is the ground truth image has same lighting and pose condition. Note that our method finds better visually correct corresponding poses.

4.4.4 Multi-PIE Database

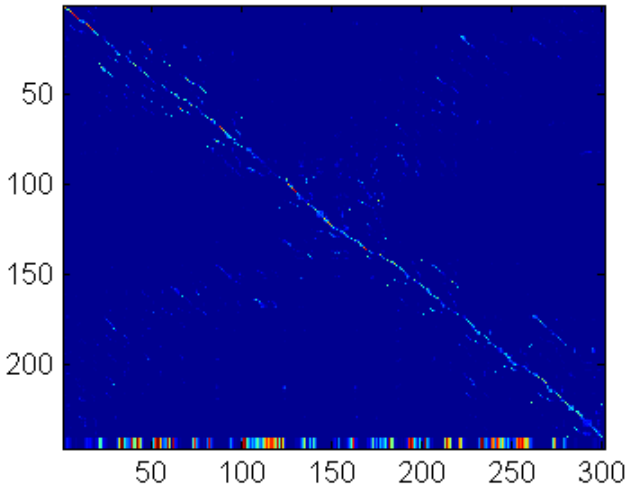
In Multi-PIE dataset in Gross *et al.* (2010), 300 images are captured for each subject under 15 view points and 20 illumination conditions. These images are cropped and resampled to 40×40 pixels for computational convenience. We demonstrate two experiments on this database in a more complicated situation with variations of illumination and pose. In first experiment, We randomly collect 200 pairs of subjects and use the entire 300 images to evaluate our method. For each pair of sets, the variation of two sets are same, i.e., the testing set is covered by the reference set. We compare the matching performance using HoG feature (DM-HoG) Dalal and Triggs (2005), Canonical Correlation Analysis (DM-CC) Kim *et al.* (2007), Manifold Alignment using RANSAC (MA-R) and Wang and Mahadevan (2009b). Due to Multi-PIE database include different type of variations, the structure feature is not enough for the alignment. The HoG feature are imported into the feature matching error in Section 4.2.1.1. We can see that our method achieves better visual alignment in Figure 4.11. Moreover, the proposed method obtains a meaningful aligned joint manifold and nearly binary correspondence matrix result as show in Figure 4.14. In Multi-PIE database, the interval of different pose is 15 degrees. To quantify the accuracy of the proposed manifold alignment, we present the percentage of correct poses versus the pose error in Table 4.2. Note that our method achieves significantly higher accuracy. Since this database has larger gap between poses and the lighting changes are included in dataset, the test results are lower than FacePix database.

In the second experiment, we use the same 200 pairs of image sets. However, we only randomly select 240 out of 300 images to build incomplete reference set. In this case, the testing set has larger range of variation than the reference set, and some samples in the testing set should be misaligned. Figure 4.12b is an example of obtained corresponding matrix of aligning a testing set (300 images) to a reference set (240 images). The extended row is enhanced for better visualization. Only our method can detect the outliers during the alignment process. The comparison of joint manifold structure between two experiments in Figure 4.14 shows that the incomplete reference set have no negative effect on our method. The experiment results in Table 4.3 show our method achieved the best alignment accuracy. Although the alignment performance is lower than that in the first experiment, 78.8% of outliers are detected. That means if we only consider the aligned samples, our method can still achieve the accuracy of 76.19%. Discovering outliers is a very useful feature and an important advantage of our approach. As outliers can be further processed in other step this can be our further improvement.

Figure 4.12: A example of Corresponding Matrix for results for Multi-PIE database.



(a) Correspondence matrix result for complete reference set case.



(b) Correspondence matrix result for incomplete reference set case. We enhance the extend row to show the misaligned samples. Only our proposed method has the capability to detect the outliers.

Table 4.2: The pose alignment accuracy (%) of MultiPIE database via different methods for complete reference set

Method	DM-CC	DM-HoG	Wang and Mahadevan (2009b)	MA-R	MA-S
Accuracy	58.19	65.62	32.96	70.14	75.45

Table 4.3: The pose alignment accuracy (%) of MultiPIE database via different methods for incomplete reference set (with outliers). The number between brackets is the accuracy after removing outliers.

Method	DM-CC	DM-HoG	Wang and Mahadevan (2009b)	MA-R	MA-S
Accuracy	53.26	58.25	24.58	56.75	64.18 (76.19)

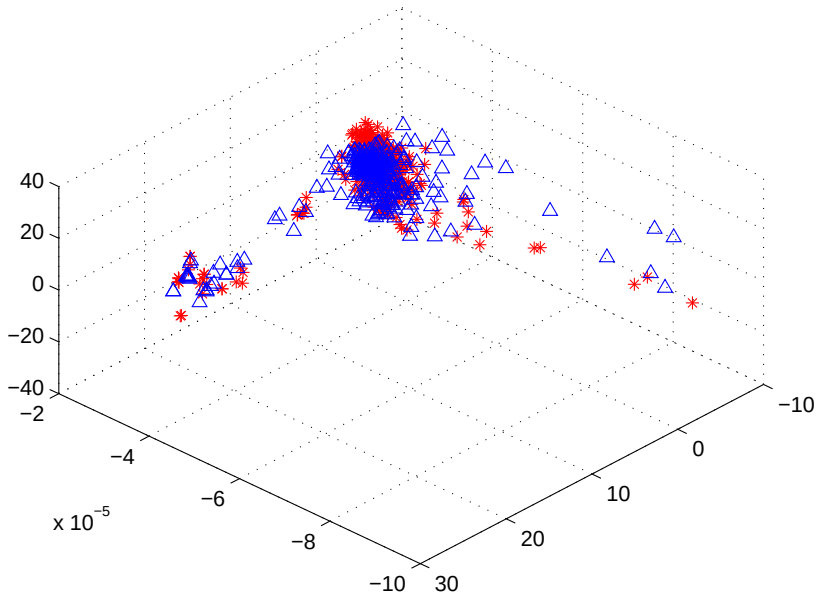
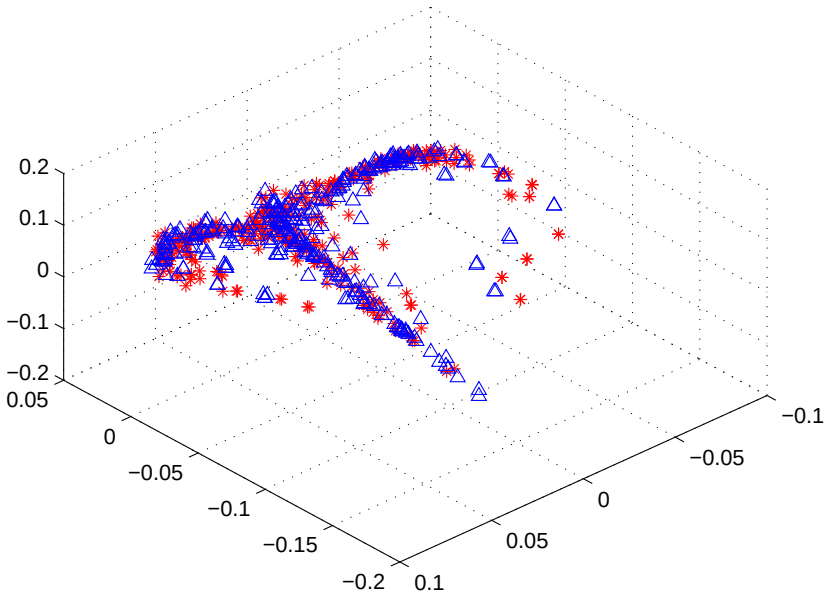
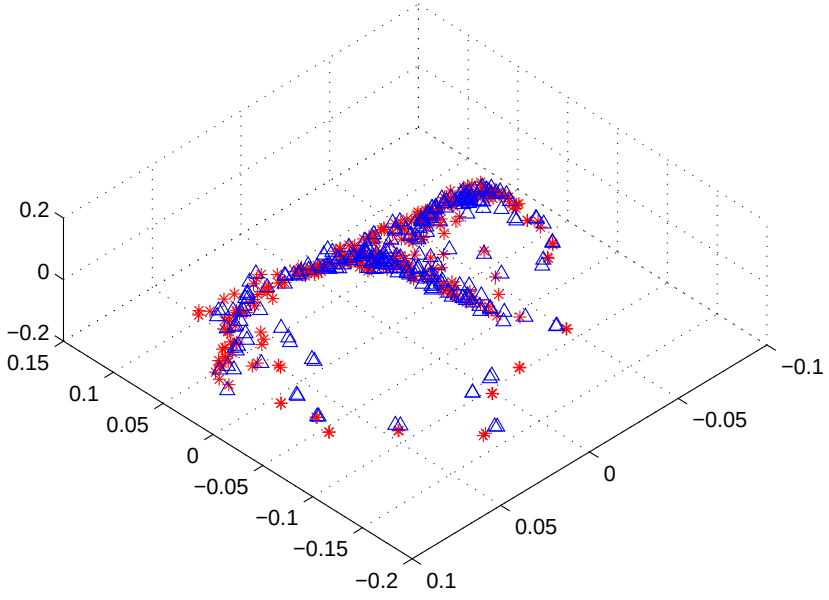


Figure 4.13: Visualization of aligned joint manifold of MulitPIE images of different subjects using Wang and Mahadevan (2009b) Wang and Mahadevan (2009b)

Figure 4.14: Visualization of aligned joint manifold of MulitPIE images of different subjects using the proposed method in (a) complete reference set case. (b) incomplete reference set case. Our method can achieve similar result in different corresponding conditions.



(a) Result for Manifold Alignment using Softassign (MA-S) using complete reference set



(b) Result for Manifold Alignment using Softassign (MA-S) using incomplete reference set

4.5 Chapter Summary

In this Chapter, we proposed a novel unsupervised method for robust manifold alignment. We improved the Local Histogram Features for characterizing the local manifold geometry and then developed a robust matching algorithm, in which accurate correspondences are estimated between two manifolds. These extracted features are set as initials of an iterative optimization framework under manifold structure preserving constraints. A joint manifold and a correspondence matrix are approximated within this framework. The main advantages are improvement of the features and removal of the assumption for reference set, which work together in a more elegant way to handle manifold alignment problem. The proposed manifold alignment algorithm was applied to four different datasets and has achieved significantly superior results when compared to existing state-of-the-art approaches.

The techniques in Chapter 3 and 4 can be used for group pose estimation on face image sets. In the next chapter, we will discuss normalize faces with different poses into the standard frontal faces.

Chapter 5

Face Normalization using Gaussian Processes Regression

As discussed in Chapter 2, there are already some good solutions to the illumination and expression variation problems, but pose variation remains the most challenging problem for face recognition. Among many techniques dealing with pose issues in Section 2.3, we adopt the 2D transformation method to handle this problem due to its simplicity and effectiveness. The existing methods are mainly using the linear assumption with different noise modeling. They cannot achieve satisfactory performance, when the pose difference becomes large.

In this chapter, we propose a novel regression approach based on Gaussian Processes Regression (GPR) technique. Unlike conventional linear regression method, the output of GPR is a Gaussian distribution. The mean of distribution is used as output, and additionally the variance can include more information about the reliability of the regression result. We choose a combined kernel function which can adaptively handle the linear and nonlinear cases. For face recognition scenario, we first train the GPR models for different poses with an aim to transforming one non-frontal pose image into a frontal view. When an unknown image comes, we should estimate the pose and assign it to the corresponding GPR model. In this chapter, a LDA based classifier is applied to find the most related 3 possible pose classes. Then we use the corresponding GPR models to predict the frontal view distribution respectively. The joint distribution is also a Gaussian distribution which is the product of the three output distributions. Face recognition is implemented on the normalized faces. The effectiveness of the proposed method is demonstrated on Multi-PIE databases Gross *et al.* (2010). Comparison with another regression based method Sparse Representation-based Regression (SRR) Zhang *et al.* (2013) shows that our algorithm is more accurate even when using a smaller size pre-training set. In this chapter we focus on describing the technique itself, its usage in our FRIS system will be investigated in Chapter 7.

The rest of the chapter is organized as follows. In Section 5.1, we introduce the Gaussian Processes Regression technique. Then we present how to implement regression technique

for a face pose normalization problem in Section 5.2. Section 5.3 presents our experiments and results emphasizing the power of the proposed method. Section 5.4 concludes this chapter and briefly introduces the application in FRIS system.

5.1 Gaussian Processes Regression

In this section, we will briefly review the Gaussian Process Regression model for learning a mapping. Regression aims to estimate the relationship between input (independent) and output (dependent) variables. Rather than defining an exact function f in normal regression techniques, we can assume $f(x)$ as a collection of random variables.

5.1.1 Gaussian Process

The definition of a Gaussian process is "a collection of random variables, any finite number of which have a joint Gaussian distribution." see Rasmussen (2006). A real process $f(x)$ is following a Gaussian distribution which is specified by the mean function $m(x)$ and covariance function $k(x, x')$ as

$$m(x) = \mathbb{E}[f(x)] \tag{5.1}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{5.2}$$

and the Gaussian process can be written as

$$f(x) \sim GP(m(x), k(x, x')). \tag{5.3}$$

That means for a given x the random variables represent the distribution of the function $f(x)$. Originally the Gaussian process is defined over time, that is, x in the formulation represents time. In our case, x represents input data and the d dimensional face images $X = \{x_1, ..., x_n\} \in \mathbb{R}^d$ are used as input. For finite high dimensional input data, the output $\mathbf{f}$ follows a joint Gaussian distribution:

$$\mathbf{f} \sim \mathcal{N}(0, K(X, X)) \tag{5.4}$$

where $K(X, X)$ is the $n \times n$ covariance matrix and can be calculated for specific covariance function $k(x, x')$, i.e., $K_{ij} = k(x_i, x_j)$. Usually, the mean function is set to be zero for notational simplicity.

5.1.2 Finding Hyperparameters

In this thesis, we adopt a widely used covariance function

$$k(x_i, x_j) = \theta_0 exp(-\frac{1}{2}(x_i - x_j)^\top M(x_i - x_j)) + \theta_1 x_i^\top x_j + \theta_2 \tag{5.5}$$

where θ_1 and θ_2 are weight variances for each radial basis function and linear kernel, $M = diag(m_1, ..., m_d)$ are parameters of length scales for different input dimension (in our case they are for each pixel), θ_2 is the model bias. Since this covariance function is the combination of linear and Gaussian kernel, it has the capability to handle both linear and non-linear data structures.

In practice, image noise should be considered. A white Gaussian noise ε with variance θ_3^2 is used for this purpose. The covariance function becomes

$$k(x_i, x_j) = \theta_0 exp(-\frac{1}{2}(x_i - x_j)^\top M(x_i - x_j)) + \theta_1 x_i^\top x_j + \theta_2 + \theta_3^2 \varepsilon_{ij} \tag{5.6}$$

where $\varepsilon_{ij} = 1$ when $i = j$, otherwise $\varepsilon_{ij} = 0$. Now we can formulate our problem as below. With given training data set $\{X : \mathbf{f}\}$, the hyperparameters $\Theta = \{\theta_0, M, \theta_1, \theta_2, \theta_3\}$ can be found by maximizing the following marginal likelihood via using the Scaled Conjugate Gradient algorithm (SCG) Møller (1993).

$$\log p(\mathbf{f}|X) = \frac{1}{2}\mathbf{f}^T K(X, X)^{-1}\mathbf{f} - \frac{1}{2}\log|K(X, X)| - \frac{n}{2}\log 2\pi \tag{5.7}$$

Generally, GPR models are designed for multi-input but only one output dimension, i.e., GPR is an univariate regression model. For our purpose, to have a d dimensional output, we need to employ d GPR models for each output dimension. That means there are $d^2 + 4d$ hyperparameters in one pose mapping function, which is far too complicated and the whole model will overfit easily. To simplify the whole model, we assume that the output dimensions (each pixel in image x) are a priori identically distributed Rasmussen (2006). This allows us to employ the same $d + 4$ hyperparameters for each output.

5.1.3 Regression

Given a test points x_*, it is not too hard to find the regressing results which are corresponding to the conditional joint Gaussian prior distribution on the observations Rasmussen (2006). We obtain the predictive mean and corresponding variance as follows:

$$\bar{\mathbf{f}}_* = K(x_*, X)[K(X, X) + \theta_3^2 I]^{-1}\mathbf{f} \tag{5.8}$$

$$var(\mathbf{f}_*) = K(x_*, x_*) - K(x_*, X)[K(X, X) + \theta_3^2 I]^{-1}K(X, x_*) \tag{5.9}$$

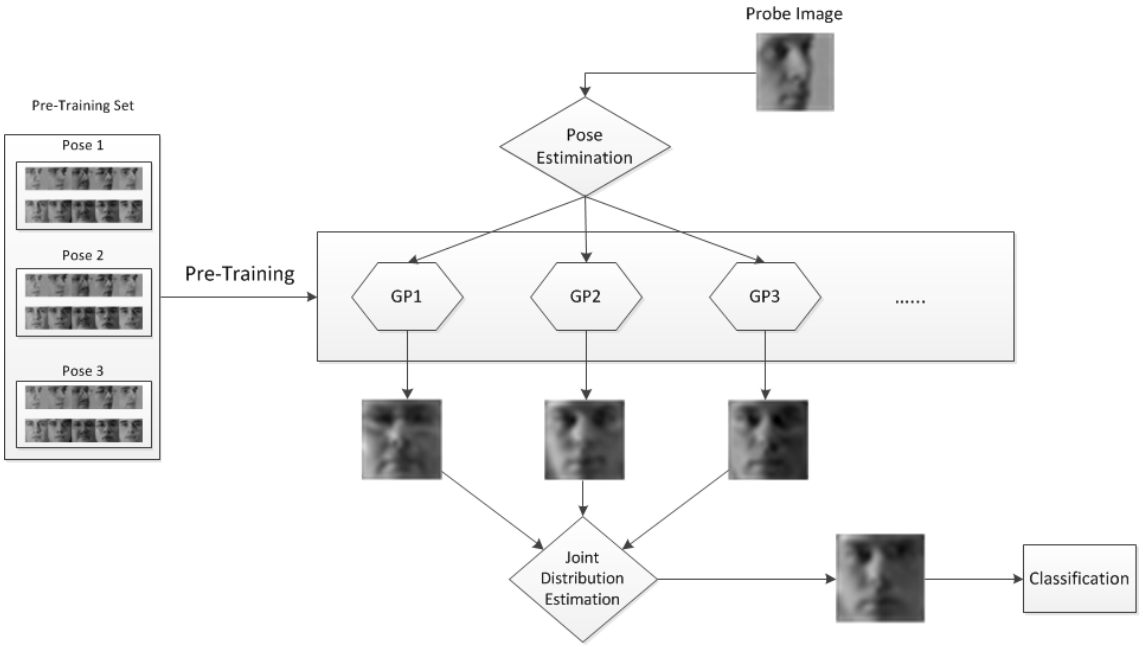


Figure 5.1: The process of normalize one face image without pose label. For each image, we use 3 Gaussian Processes Regression model to do the normalization.

Since each dimension is assumed to be identically distributed, variances of each dimension are the same as others.

5.2 Face Normalization

In this section, we propose a novel face normalization approach to assist in face recognition with different poses. Firstly, we estimate the head pose using a rough pose classifier. Next a face with different pose will be normalized into a frontal standard face using Gaussian Processes Regression (GPR). Face recognition is then performed on the normalized front pose faces with probability metric. These processes are described in detail in the following sections and shown in Figure 5.1 with Algorithm 4.

Assuming we have the pre-training data X already labeled with P discrete poses which is $X = [X^0, ..., X^k, ..., X^P]$. For each pose k the data set $X^k = [x_1^k, ..., x_{N_k}^k]$ contains N_k images. We denote X^0 to be the frontal pose, which we set as the regression target. Then we organize the data into pairs as $T = [\{X^1, X^0\}, ..., \{X^k, X^0\}, ..., \{X^P, X^0\}]$ and each subset $T^k = [\{x_1^k, x_1^0\}, ..., \{x_{N_k}^k, x_{N_k}^0\}]$, where we put the side face images with corresponding frontal face with the same identity.

Algorithm 3: LDA based Pose Estimation Algorithm

Input: Face image x_* in an unknown pose

Output: 3 possible facial pose label l_1, l_2, l_3

1 Find LDA projection matrix W using grouped training data $X_1, ..., X_P$.

2 Project $X_1, ..., X_P$ and x_* into LDA feature space.

$$y_* = W^T x_*$$
$$Y^k = W^T X^k$$

3 Calculate mean vector $\mu^k = \frac{1}{N_k} \sum(X^k)$ and the covariance matrix Σ^k of each pose set.

4 Compute the Mahalanobis distance

$$d_k(y_*) = \sqrt{(y_* - \mu^k)^T (\Sigma^k)^{-1} (y_* - \mu^k)}$$

and choose the 3 smallest pose set as the output.

5.2.1 Training GPR model for Pose Normalization

The most straightforward idea to solve the cross-pose problem is to learn one single GPR model for all P pose pairs T at once. However, the size of the covariance matrix $K(X, X)$ in the learning stage is unreasonably huge due to the oversize training samples. A practical way is to learn one GPR model GP^k for each pose pair T^k independently. More specifically, each GP^k is defined by its hyperparameters Θ^k and achieved by solving the optimization derived from (5.7) as follow:

$$\arg\max_{\Theta^k} \log p(X^0 | X^k) \tag{5.10}$$

This strategy means that each GPR model only has the ability to handle small range of poses, but each model has limited data size and are easily solved. Thus, the pose estimation is required to assign an unknown input into the corresponding GPR models first. Next we will solve this problem.

5.2.2 Head Pose Estimation

Given input face image x_* with an unknown head pose, our goal is to identify pose label of each image. Firstly images are projected to a low-dimensional LDA subspace, and identify the pose label by using a normal density metric. The estimated distance of data x being in pose k is given by computing the Mahalanobis distance $d_k(x) = \sqrt{(x - \mu^k)^T (\Sigma^k)^{-1} (x - \mu^k)}$, where μ^k is the center of the projected X^k and Σ^k is the

Algorithm 4: Gaussian Processes Regression based Cross-Pose Face Recognition Algorithm

Input: Face image x_* in an unknown pose

Output: Facial identity label (l)

1 Apply the pose estimation (Section 5.2.2) to obtain pose label l_1, l_2, l_3.

2 Normalize x_* to frontal estimation m_1, m_2, m_3 using corresponding GPR model $GP^{l_1}, GP^{l_2}, GP^{l_3}$.

3 Synthesis the joint final output m_* using (5.12)

4 Face recognition classification with the normalized frontal pose.

covariance of pose k. LDA is used since it is a simple linear transformation which can preserve the pose variations while reducing the impact of other variations. Although this simple pre-process only gives a rough pose label to the test data, the uncertainty of pose estimation can be predicted through the GPR's output covariance. In order to improve pose prediction accuracy, we chose the top 3 likely poses l_1, l_2, l_3 for the next step, since the correct pose estimation is normally lied within this range.

5.2.3 Pose Normalization

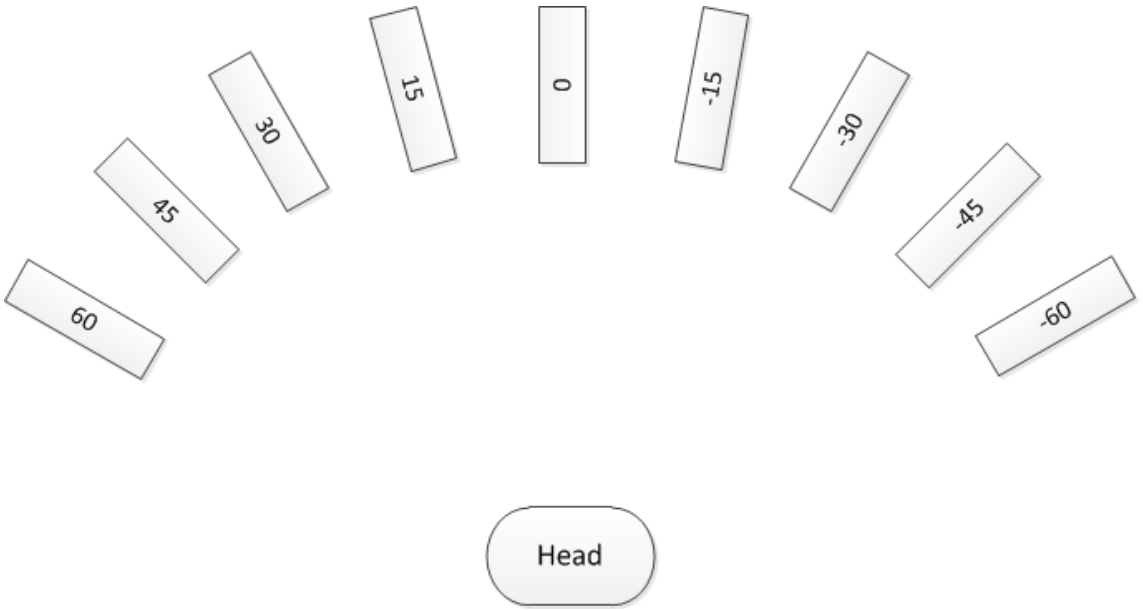
With 3 pose labels, one test face image x_* can be transformed to the frontal view through three corresponding GPR models $GP^{l_1}, GP^{l_2}, GP^{l_3}$. Applying (5.8) and (5.9) we can obtain the output of three mean vectors m_1, m_2, m_3 and the variances v_1, v_2, v_3. The product of these three Gaussian distributions is still a Gaussian distribution which can represent the synthetic frontal image. The mean m_* and the variance v_* of final joint distribution can be computed as follows:

$$v_* = \left(\sum_{i=1}^{3} v_i^{-1} \right)^{-1} \tag{5.11}$$

$$m_* = \left[\sum_{i=1}^{3} \frac{m_i}{v_i} \right] v_* \tag{5.12}$$

Then m_* is the final estimation of the normalized frontal face.

(a) Sample Images for different subjects



(b) 9 choosen camera views in MultiPIE database.

Figure 5.2: MultiPIE database.

5.3 Experiments

Several experiments are conducted to demonstrate the effectiveness of the proposed method. The dataset we used is MultiPIE database containing multi-view face data with different illumination and expression. In Multi-PIE dataset Gross *et al.* (2010), 300 images are captured for each subject under 15 view points and 20 illumination conditions.

9 poses are chosen from 60° to −60° with 15° apart and 20 illumination conditions for each pose (shown in Figure 5.2). Since we focus on cross-pose issue in this chapter, only the neutral expression faces in MultiPIE database are used in experiments. These images are

Table 5.1: Normalization error (RMSE of pixel difference) under different pose (smaller is better)

Pose	15°	30°	45°	60°	−15°	−30°	−45°	−60°
SRR	58.07	64.28	78.18	85.04	58.60	71.01	79.57	86.89
GPR	41.43	47.73	55.00	60.47	41.80	48.44	54.01	60.37

Table 5.2: Recognition rate(%) under different training size

subjects number	15°	30°	45°	60°	−15°	−30°	−45°	−60°
50	92.1	75.7	58.5	36.95	92.65	75.05	65.15	37.55
100	94.7	80.6	59.35	46.65	94.15	79.95	69.25	49.8
150	95.4	86.5	68.45	48.4	96.65	85.3	74.65	52
200	97.4	91.7	72.45	55.2	96.8	87.9	76.6	61.9

cropped and re-sampled to 20×20 pixels for computational convenience. We compare the proposed method with conventional pose-robust face recognition Sparse Representation Regression (SRR) Zhang *et al.* (2013).

Following the experiment setup in Zhang *et al.* (2013), we select 300 subjects in all 4 sessions of multiPIE and split the data into three parts:

- Pre-training set: This set consists of images of the first 200 subjects with all 9 poses. Each side pose is used jointly with a frontal face for dictionary learning in SRR Zhang *et al.* (2013) and for training face normalized GPR models.
- Gallery set: The frontal faces of the remaining 100 subjects are assigned to be the training set.
- probe set: Testing sets consists of the side faces of the remaining 100 subjects.

Note that the three parts have no overlap between each other. The aim of the experiments is to evaluate face recognition in scenarios that testing and training images are in different poses. The pre-training set is used for view-dictionary learning in Zhang *et al.* (2013) and training Gaussian processes model to obtain hyperparameters. We set the dictionary size as 800 which is the best performance size reported in Zhang *et al.* (2013).

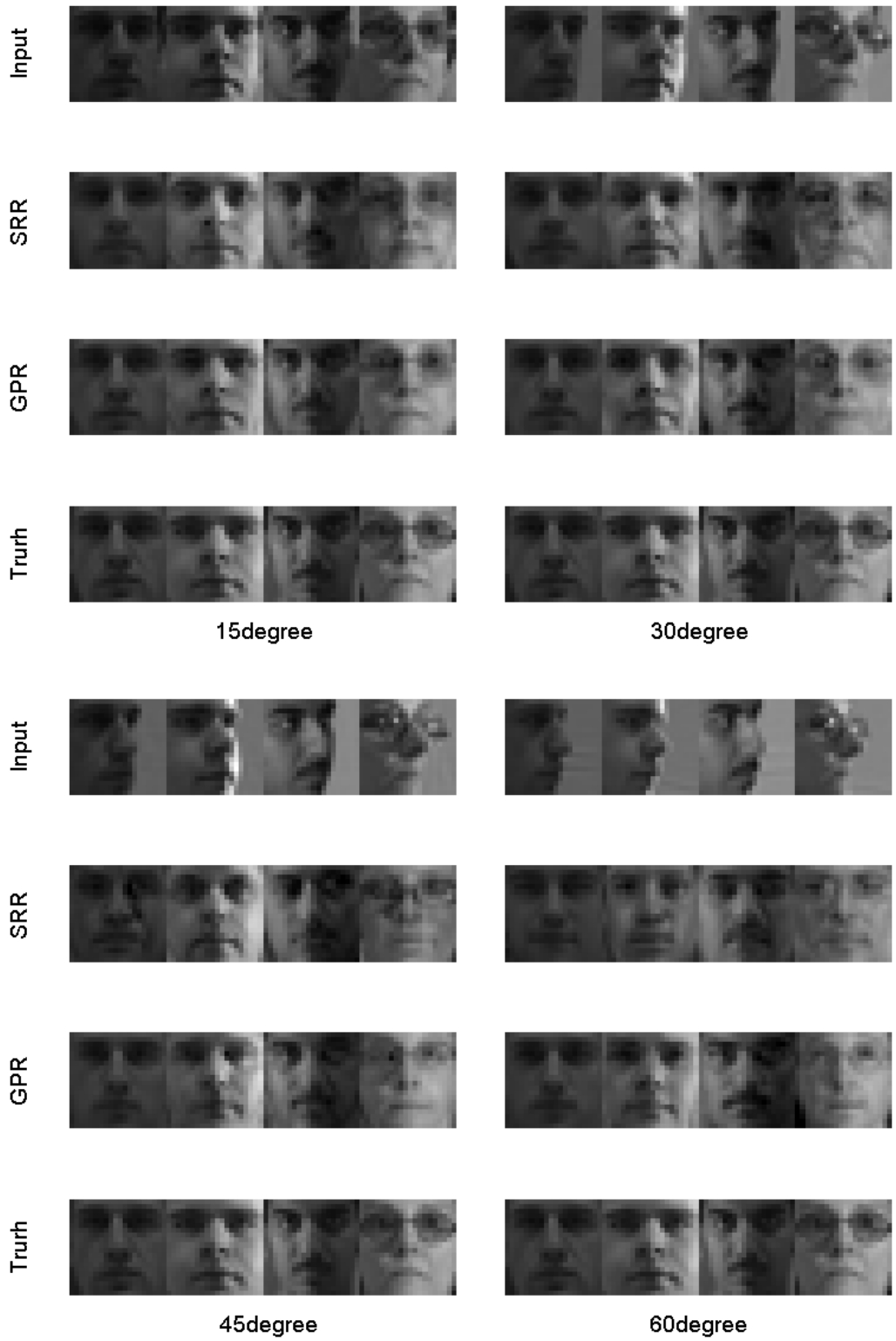


Figure 5.3: Comparison the results from different pose normalization methods. The first row is the testing images from pose $15°, 30°, 45°, 60°$. The third subject has a mustache. The last subject wears glasses.

Table 5.3: Recognition rate(%) using different classifier

Pose	15°	30°	45°	60°	−15°	−30°	−45°	−60°
GPR+NN	97.4	91.7	72.45	55.2	96.8	87.9	76.6	61.9
GPR+SRC	97.25	82.85	65.35	43.3	98	76.05	57.2	40.35
GPR+PCA	98.4	94.45	82.25	69.85	98.8	93.95	86.25	66.1
GPR+LPP	98	94.6	81.55	63.5	99.8	92.8	85.1	67.95
GPR+LDA	99	95.05	87.5	64.9	100	96.95	91.45	68.35

Table 5.4: Recognition rate(%) under different pose

Pose	15°	30°	45°	60°	−15°	−30°	−45°	−60°
SRR	89.1	69.5	44.2	27.7	88.7	67.35	45.8	31.45
GPR	97.4	91.7	72.45	55.2	96.8	87.9	76.6	61.9
GPR(without label)	95.5	87	70.4	30.3	94.6	85.2	70.3	35.2

5.3.1 Face Pose-normalization

In the first experiment, we present the normalization quality of different methods. The Root Mean Squared Error (RMSE) between the normalized face and the ground truth (Pose 0 in database with same lighting condition) is adopted as the criteria to measure the difference between images. For convenience of comparison, we use the corresponding GPR model to normalize face based on the known pose label for testing image. The complete test will be provided in the next section.

As summarized in Table 5.1, the recognition errors using the normalized faces using the proposed method compared with the ground truth are always smaller under all poses. In general, the errors of all methods will become larger when pose differences increase. That may be caused by the large missing part of the face when pose difference is rising.

Moreover, we present the pose-normalized results of 4 selected subjects from 4 different poses in Fig. 5.3. The first two subjects are normal face with different light conditions; the third subjects has mustache and the last one wears glasses. It can be seen that the proposed GPR method can synthesize better visual frontal images than the SRR method, which is more similar to the ground truth images. In details, for the first two simple cases, the proposed GPR performs in a stable and robust manner when the pose difference are

large, while the SRR method performs worse visually. In fact, when faces have little pose difference, one may have a nearly linear mapping, but when the pose difference increases, linear mapping cannot handle these large changes anymore. That may be the reason why the proposed GPR can perform better. For people with glasses, both methods perform worse than for cases without glasses, but the proposed GPR shows perceptibly better visual results.

5.3.2 Pose Robust Face recognition

From above discussions, it is demonstrated that the proposed GPR approach can produce better performance in pose normalization with regard to ground truth. In fact, the lower error of pose normalization cannot guarantee the better accuracy rate on face recognition. In this section, we evaluate two different methods for face recognition.

We divide the testing set into groups of different poses to evaluate the face recognition performance. The pose differences indicate the degree of difficulty for face recognition. We first present one experiment using different pre-training size to evaluate the generalization ability of the proposed GPR model. We test face recognition performance on different size of pre-training set and report it in Table 5.2. The recognition accuracy is reasonably good even with a smallest training set size and the performance is increasing when the pre-training set expands. We notice that the GPR model has achieved higher performance than SRR even with just a quarter of pre-training samples used in SRR.

We also test the GPR with different classification methods and report its performances in Table 5.3. The results show that this normalization method can well cooperate with most of state of the art techniques and further improve the performance even to 100% in small pose difference using LDA.

We finally compare the proposed GPR model with the state of the art algorithm SRR. We conducted three experiments on SRR, and the proposed GPR with and without knowing the pose labels for testing images, and the recognition results are shown in Table 5.4. Note that the results of SRR are slightly lower than results report in Zhang *et al.* (2013), since we double the size of the gallery set, thus the recognition difficulty is increased. The proposed GPR method can achieve superb performances that are much higher than SRR. It is just a slightly worse when comparing with the ideal case (with known pose labels). This indicates that the GPR model can perfectly handle the unknown pose issue, which is not possible for traditional regression methods in current literatures.

5.4 Chapter Summary

In this chapter, a novel facial pose normalization approach is proposed based on Gaussian Processes Regression. Technically, in the training phase, instead of learning one single complete GPR model for the whole pre-training set, we train each pose subset separately for each GPR model. In the testing phase, the probe faces are firstly assigned three possible pose labels by a LDA based pose estimation. Three prediction distributions are obtained by the corresponding GPR models and then these models are integrated into a joint distribution, which can represent the final synthetic result. Experimental results demonstrate the advantages of the proposed method in comparison with the SRR method both in face pose normalization and face recognition. Next, the face normalization technique will be tested on FRIS problem in Chapter 7.

Chapter 6

FRIS using Margin Preserving Projection

A state-of-the-art work Cevikalp and Triggs (2010) characterizes each image set as a convex region and proposes a new metric for comparison of image sets, which is defined as the minimum distance between points in convex sets. The experiments in Cevikalp and Triggs (2010) show that this approach can achieve better performance than previous works Hadid and Pietikainen (2004); Yamaguchi *et al.* (1998); Fan and Yeung (2006). Technically, Cevikalp and Triggs (2010) mainly concerns how to measure the similarity of two image sets, but pays less attention to discriminant learning with such similarity. Further, the classification in Cevikalp and Triggs (2010) is done in original dataset without dimensionality reduction. The main technique contribution of this chapter is that we use the idea of dimensional reduction into the FRIS system to improve the performance. And state the relations between SVM technique and convex hull distance in detail.

In this Chapter, a novel linear dimensionality reduction algorithm designed for convex hull model is proposed. We intend to compute the convex hull distance by definition. However, when two image sets are inseparable, direct computation is not suitable. In this case, SVM is implemented to handle the problem. The margin obtained from SVM is used to approximate the convex hull distance. Using PCA on directions derived from SVM, one can find a subspace spanned by the dominant directions. Finally, classification is implemented in the reduced feature space based on the convex hull distance. Comparing to state-of-the-art methods, the proposed method achieve better results in terms of accuracy and computational time.

6.1 Preliminaries

Let N_c be the number of classes and n_c $(c = 1, \ldots, N_c)$ be the number of data samples belonging to the cth class $(\sum_c n_c = N)$. The input data set $X_c = \{x_{ck} \in R^d, k = 1, \ldots, n_c\}$ is sampled from the cth class. The proposed method aims to perform dimensionality

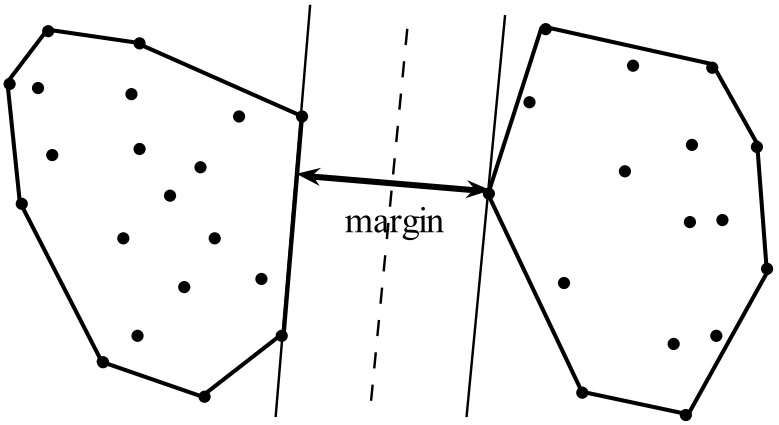


Figure 6.1: Two convex hulls and the distance between them.

reduction from the input space data points $X \in R^{d \times N}$ to a lower dimensional feature space $Z = [z_1, \ldots, z_N] \in R^{m \times N} (m \ll d)$ for FRIS.

6.1.1 Convex Model

An intuitive idea for FRIS is to approximate each image set with a convex model Cevikalp and Triggs (2010). For an unknown set we try to find its class label by using distance between convex models of testing and training image sets. The label is assigned to the training set which is closest to the testing set.

There are two major convex models, affine hull and convex hull. In the original pixel space, the dimension of affine hulls is less than d, and this necessarily holds for $n_c \ll d$. The affine space is a subset of R^d. In low-dimension feature space, the affine hulls have dimension nearly or exactly the same as the feature space dimension m. The comparability for two sets will be lost, because the affine spaces of two image sets may easily overlap even though they are separable. The restricted linear combination coefficients of a convex hull make a tighter convex model and reduce the chance of overlap in low dimension space. In this paper, we focus on the convex hull model where each image set is modeled by:

$$H_c = \{x = \sum_{k=1}^{N_c} \alpha_k x_{ck} | \sum_{k=1}^{N_c} \alpha_k = 1, \alpha_k \geq 0\} \tag{6.1}$$

Suppose we have two image sets X_i and X_j. As illustrated in Figure 6.1, the distance between them is defined as the minimum distance between any point in convex hull H_i

and any point in convex hull H_j:

$$d^c(X_i, X_j) = D(H_i, H_j) = \min_{x \in H_i, y \in H_j} \|x - y\| \tag{6.2}$$

where H_i and H_j include X_i and X_j respectively. In fact, the distance can be computed by solving the following optimization problem:

$$\begin{aligned} (\alpha^*, \beta^*) \qquad &= \arg\min_{\alpha,\beta} \|X_i\alpha - X_j\beta\|^2 \\ s.t. \qquad &\textstyle\sum_{k=1}^{n_i} \alpha_k = \sum_{k'=1}^{n_j} \beta_{k'} = 1, \quad \alpha_k, \beta_{k'} \geq 0 \end{aligned} \tag{6.3}$$

For convenience, we denote the distance by

$$d^c(X_i, X_j) = \|X_i\alpha^* - X_j\beta^*\| \tag{6.4}$$

This problem can be written more concisely in a constrained least square problem

$$\min \|\Theta\gamma\|^2$$

where $\Theta = \begin{pmatrix} X_i & -X_j \end{pmatrix}$ and $\gamma = \binom{\alpha}{\beta}$. However, the constraints of (6.3) are not standard and sometimes the solution is not unique. The possibility of solving (6.3) is under the assumption that two image sets are linearly separable. If two image sets are not linearly sparable, the defined distance will be zero. In this case, we propose to use SVM to find the support vectors, which can be used to approximate the similarity between two image sets.

6.1.2 Support Vector Machine Approximation

In an ideal case when two sets are linear separable, SVM aims to find a decision hyperplane through maximizing its margin. When the optimization margin was found, it is actually equal to the smallest distance between any samples between two sets Bishop (2006), which is exactly the definition of convex hull distance. However in practice, soft margin SVM ((6.5)) is normally applied to avoid noises caused problem. Soft margin will enlarge the distance comparing to the hard one, but will have the ability to handle the misgrouped and noisy samples. Therefore, we use the soft margin to approximate the convex hull distance in this Chapter. Suppose we have some training data x_k with corresponding class label $y_k \in \{-1, 1\}$. Usually, we solve the following soft margin SVM optimization problem

$$\arg\min_{w} \quad \frac{1}{2}\|w\|^2 + C\sum_k \xi_k \qquad s.t. \quad y_k(w^T x_k + b) \geq 1 - \xi_k, \quad \xi_k \geq 0 \tag{6.5}$$

to find the vector w perpendicular to the decision hyperplane. w represents the direction of margin, and related to the length of margin by $2/\|w\|$. When we consider the direction

w as a projection vector, the margin in the reduced subspace remains the same as in the original space as shown in Figure 6.1. Approximately, SVM problem (6.5) can be seen as the following convex hull distance problem:

$$\arg\max_{w_{ij}} d^c(w_{ij}^T X_i, w_{ij}^T X_j). \tag{6.6}$$

Here w can be found by solving the problem (6.5). w here represent not only the optimal direction of projection and the approximated convex hull distance between two image sets by

$$d^c(X_i, X_j) = \frac{2}{\|w_{ij}\|}. \tag{6.7}$$

In our experiments, all convex hull distances are calculated using (6.7).

6.2 Margin Preserving Projection

In this section, we introduce a new dimensionality reduction approach for FRIS. The process is described as follows. Given training sets $[X_1, \cdots, X_c, \cdots, X_C]$, we intend to find an optimal projection A, which maps all sets to a bunch of low dimensional sets Y_c, with $Y_c = A^T X_c$. We expect this projection to keep sufficient discriminant information through preserving margins between any two sets in lower dimensional spaces.

6.2.1 The Proposed Algorithm

The proposed algorithm is named as Margin Preserve Projection (MPP) and it includes the following three major steps:

1. **Finding the maximum margin directions**: Let w_{ij} be the maximum margin direction between sets X_i and X_j. We can obtain w_{ij} by solving SVM (6.5). (Only for $i < j$)

2. **Choosing the weights**: In order to preserve the local structure between two image sets, we calculate the relationship between them. This is motivated by Locality Preserving Projections (LPP) He and Niyogi (2003). A possible choice of S_{ij} is $S_{ij} = \exp(-\frac{d^c(X_i,X_j)^2}{\sigma^2})$ where σ is a suitable constant. $d^c()$ is computed using the margin of SVM. One simple way of selecting parameter σ is $\sigma = mean(d^c(X_i, X_j))$.

3. **Solving eigenvector problem**: Compute the eigenvectors and eigenvalues of the scatter matrix: $P = \sum \left(\frac{w_{ij}}{\|w_{ij}\|}\right)\left(\frac{w_{ij}}{\|w_{ij}\|}\right)^T S_{ij}$, where $\frac{w_{ij}}{\|w_{ij}\|}$ is a normalized direction of w_{ij}. Let the column vectors $a_1, \cdots, a_m$ be the eigenvector of P, ordered according to their corresponding eigenvalues $\lambda_1 > \cdots > \lambda_m$. The projection matrix is computed as $A = [a_1, \cdots, a_m]$ with size $d \times m$, and dimensionality reduction can be easily implemented by $Y_c = A^T X_c$ where the dimension of Y_c is m.

6.2.2 Intuition of MPP

The principal idea of this approach is underlying Principal Component Analysis (PCA) Turk and Pentland (1991). The aim of PCA is to project the data onto a low dimensional space which maximizes the variance of the projected data. The objective function of a general PCA is stated as below:

$$\arg\max_{A} \quad \sum_{ij} d(A^T x_i, A^T x_j)^2 \qquad s.t \quad A^T A = I \tag{6.8}$$

where $d()$ is the distance between two points, which is generally chosen to be the Euclidean distance. Following this idea, we expect to find a projection matrix A which can maximize the convex hull distances among image sets. The objective function can be intuitively modified as follows:

$$\arg\max_{A} \quad \sum_{i<j \quad i,j\in\{1,\cdots,C\}} d^c(A^T X_i, A^T X_j)^2 S_{ij} \qquad s.t \quad A^T A = I \tag{6.9}$$

Here we weigh each distance by S_{ij}. Since the closest two sets are the most difficult to classify, their nearest neighbors involve most important information for discrimination. According to the above analysis, we put larger weights on smaller distances.

Since A is unknown, we can hardly compute the distance $d^c()$ with variable A. In fact solving (6.9) directly is very difficult. Instead, we find A through the maximum-margin directions w_{ij} ($i < j$ and $i, j \in \{1, \cdots, N_c\}$), which are obtained by solving the SVM optimization problem (6.5).

Let W be the space spanned by the $\frac{N_c(N_c-1)}{2}$ direction vectors w_{ij}. The intrinsic dimension M_W of this space satisfies $0 \leq M_W \leq \frac{N_c(N_c-1)}{2}$, and the dimension of projection A should be in the range $0 \leq M_A \leq M_W$.

Information contained in the subspace A may not be capable to maximize all the pairwise

distance, therefore we have:

$$\max \sum d^c(A^T X_i, A^T X_j)^2 S_{ij} \leq \max \sum d^c(W^T X_i, W^T X_j)^2 S_{ij} \tag{6.10}$$

Let $\tilde{W} \in R^{M_A}$ be a subspace of W. $\tilde{W}$ can be defined as M_A dominating directions of w_{ij}. The training error

$$\tau = \max \sum d^c(W^T X_i, W^T X_j)^2 S_{ij} - \max \sum d^c(\tilde{W}^T X_i, \tilde{W}^T X_j)^2 S_{ij} \tag{6.11}$$

should be very small, when unimportant directions are ignored. In this case, $\tilde{W}$ can be seen as an approximation of the optimal projection A.

Base on above analysis, the approximation of optimum projection A can be found as eigenvectors corresponding to the largest M_A eigenvalues for the pairwise scatter matrix P. Choosing enough eigenvectors can guarantee that the margins are mainly preserved. Abandoning eigenvectors corresponding to small eigenvalues will reduce noise.

In general, we in fact expand PCA for the case of image sets by replacing point-to-point distances with set-to-set convex hull distances. Our approach is to find a subspace spanned by dominant projection directions of all w_{ij}. This subspace should provide enough information for a convex distance classifier. After projection, we expect improvements on classification performance and reduction on computational cost.

In comparison to LDA, our proposed method relaxes the dimensionality bound, the maximum rank of A is $\min(\frac{N_c(N_c-1)}{2}, N_{sv})$, where N_{sv} is the number of all support vectors, instead of $N_c - 1$.

6.3 Experiments

We tested the proposed method on two benchmark databases: Honda/UCSD Lee *et al.* (2005) and CMU MoBo Gross and Shi (2001). These two sets contain several videos each recording one subject's movement. We use a Viola-Jones face detector Viola and Jones (2004) to find all facial images used for training and testing. Before experiments all detected images were histogram normalized to eliminate some lighting effects Russ (2002).

(a) Honda/UCSD Database

(b) CMU MoBo Database

Figure 6.2: Facial images detected from Honda/UCSD and CMU-MoBo database.

6.3.1 Databases

The Honda/UCSD Video Database was collected for video-based face recognition. It contains 62 video sequences (including videos with partial occlusion) of 20 different people. It is divided into two subsets: 20 videos for training and the remaining 42 videos for testing. Each cropped facial image was normalized to 40×40 gray scale image. Figure 6.2a presents some images from this database that belong to the same subject. From each training and testing set in this database, we build a randomly selected corresponding subset which contains 50% quantity of images and perform experiment on those subsets. The experiments are repeated for 10 times and the average performances are recorded. [1]

The CMU MoBo database contains video of 24 individuals walking on a treadmill in an indoor environment. There are totally 96 sequences for 24 subjects, and each person has 4 sets of images. Each detected image was resized to 40×40 gray scale image. Figure 6.2b shows some examples of the detected faces from one subject. In this experiment, we randomly select one set of four for each subject as training and the remaining 3 as testing.

[1] We did not use the setup for Honda/UCSD database in Cevikalp and Triggs (2010), because the number of testing sets is too small.

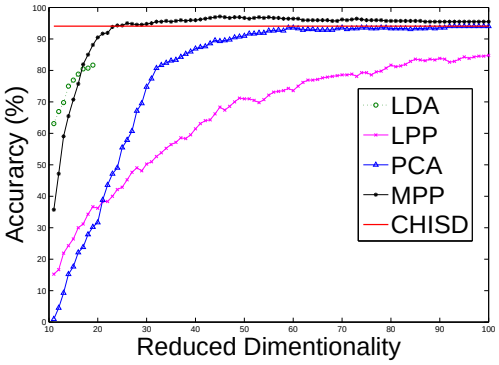
Again, all the experiments are repeated for 10 times and we report the average results.

The methods compared here include: Manifold-Manifold Distance (MMD) Wang *et al.* (2008), Convex Hull based Image Set Distance (CHISD) Cevikalp and Triggs (2010), Locality Preserving Projections (LPP) He and Niyogi (2003), Linear Discriminant Analysis (LDA) Belhumeur *et al.* (1997), Principal Component Analysis (PCA) Turk and Pentland (1991) and the proposed method MPP. We tested CHISD and MMD in the original pixel feature space as baselines. For all other methods, we perform dimensionality reduction first and then implement CHISD in these corresponding reduced feature spaces. For MMD, we use the same setup of parameters in Wang *et al.* (2008). For simplicity, we set $k = 10$ the number of neighbors in LPP. The penalty parameter C in SVM varies from 10 to 100 to explore its effects.

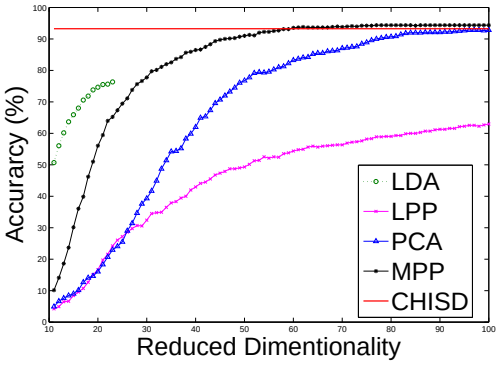
6.3.2 Experimental Results and Discussions

Figure 6.3 shows the average recognition accuracy of LDA, PCA, LPP and MPP under different reduced dimensions ($m = 11, \cdots, 100$). One exception is LDA which only can extract at most $N_c - 1$ meaningful dimensions. The best recognition rates and the averaged running time are shown in Table 6.1. Some interesting observations are provided on the performance of the evaluated algorithms.

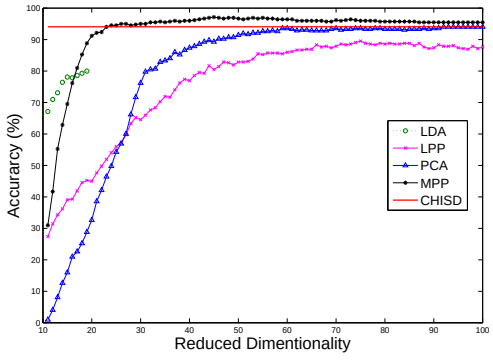
Firstly, for most methods, recognition rates increase consistently when the reduction dimensions increase. It can be seen that the two traditional methods - LDA and LPP yield poor performances. The performance of PCA is better than LDA and LPP, but not overtakes the baseline CHISD. Moreover, in a single frame based recognition problem, PCA can improve performance by preserving principal data variances, but here it is just similar to baseline CHISD. This result is different from some previous frame based experiments He and Niyogi (2003). Though it may be not fair to compare them here, the experiments suggest that traditional dimensionality reduction methods may be unsuitable for set based classification. This is due to the fact that PCA, LDA and LPP are performed on data points, but the final classifier is based on image sets. Secondly, the proposed MPP gives superior results than other methods with the best performance. Unlike other methods, the proposed MPP method preserves the margins, especially the smaller ones which contain more discriminant information. By focusing on set based information, the MPP method provides significant performance benefits. Note that the best result does not come with the highest dimension, which means that abandoned eigenvectors contain noises and have no help in recognition. Finally, CHISD, MPP and PCA are very stable when C changes. However, the performance of LDA and LPP are susceptible to the variations of C. In fact,



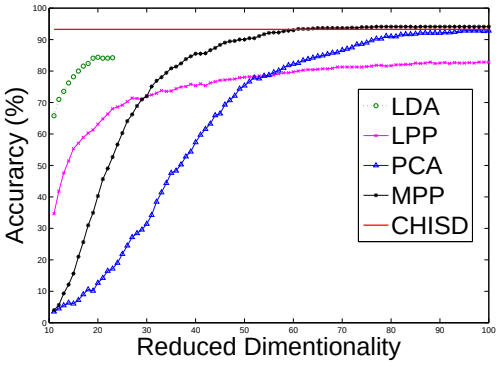
(a) Honda/UCSD $C = 10$



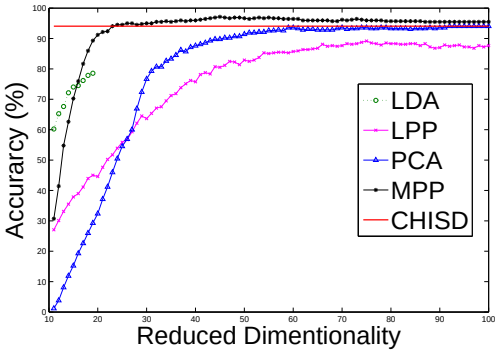
(b) CMU-MoBo $C = 10$



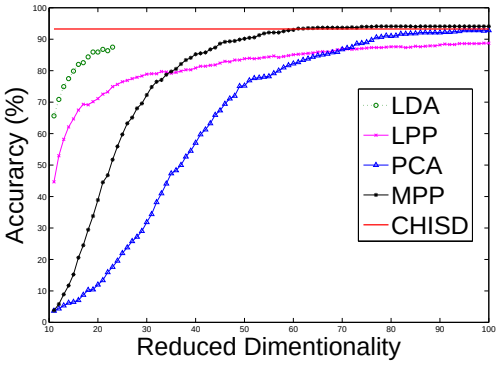
(c) Honda/UCSD $C = 50$



(d) CMU-MoBo $C = 50$



(e) Honda/UCSD $C = 100$



(f) CMU-MoBo $C = 100$

Figure 6.3: Comparison of the averaged accuracy versus the reduced dimension of LDA, LPP, PCA and MPP on the Honda/UCSD and CMU-MoBo database.

Table 6.1: Comparison of the related algorithms to MPP on the Honda/UCSD and CMU-MoBo database. In accuracy, the first number is the highest recognition rates through different reduced dimensions; the following number is the corresponding dimension. The running time in seconds is the average time consumed on testing one set with the best dimension.

Algorithms	Honda/UCSD		CMU MoBo	
	accuracy	time(s)	accuracy	time(s)
MMD	79.76%	1.13	82.54%	182.20
CHISD	94.05%	6.27	93.23%	18.21
LDA+CHISD	80.00%,19	0.04	84.37%,20	0.38
LPP+CHISD	89.52%,75	0.14	82.82%,87	1.46
PCA+CHISD	94.05%,92	0.12	92.81%,94	2.74
MPP+CHISD	**97.14%,44**	0.10	**94.09%,76**	1.76

C represents the training error of SVM. Less sensitive to different C implies that these approaches have good generalization capability.

6.4 Chapter Summary

In this chapter, we proposed a new linear dimensionality reduction algorithm called Margin Preserving Projections (MPP). It is based on the metric of convex hulls for FRIS. The most interesting feature of this method is that it focuses on the relations between image sets rather than single images. This allows the algorithm to retain more important information for set based classification problems. Experiments on face image databases show that the proposed method produces better recognition accuracy and is less time consuming than some related algorithms.

Chapter 7

Pose Robust FRIS Systems

As we discussed in Section 2.1.2, methods using local sample based similarity can only apply on data set with similar variation conditions. The existing structure based approaches need model training data to cover complete variations in order to span the manifold structure. The current methods are limited for certain variation conditions of image set. Actually, these are a significant limitations since the image sets in real applications are hard to satisfy either of these above assumptions. Since such problems are caused by the unpredictable variations appeared in the image set, how to reduce influences of pose variations is the key issue of solving this problem.

To overcome the limitations, the basic idea is to transform the uncontrolled input image set to satisfy the assumption of existing methods. There are two choices for satisfying the two assumptions respectively. The first one is to generate different conditions to build manifold structure for every person. This requires complete training set for each person, which is not always possible in practice. Converting an unknown condition image into another condition is a difficult task, since it require huge amount of training images to construct the possible transformations. On the other hand, normalizing unrestricted face into a standard one is a more reasonable solution. It is to reduce the variation which will be less noisy and more robust. Also it only requires building reasonable size training set for normalization which is much more achievable.

In this chapter, we propose a new framework to follow the second choice. We will use techniques introduced in the previous Chapters to build an automatic FRIS system for unrestricted input image sets. Manifold Alignment technique is used to assign a possible pose label to each input image. Then faces with uncontrolled variations are transformed into the standard frontal ones. Finally the dimensional reduction technique is applied to further extract the discriminant features for face recognition. We will introduce the details in the following sections.

7.1 System Overview

We introduced the overall system in Section 1.3. Figure 1.2 shows an overview of our FRIS system. The system consists of three major components:

- Manifold Alignment and Pose Estimation. We proposed manifold alignment technique in Chapter 3 and Chapter 4. The approaches find correspondence between image sets using structure and sample features. This correspondence can be utilized to estimate the pose for face images.
- Face Pose Normalization. We discussed the face normalization in Chapter 5. In this system, the LDA based pose estimation is only used for missing aligned samples. Most of the pose label is assigned based on the alignment results.
- Discriminate Feature Extraction. In Chapter 6, we proposed a discriminant feature learning algorithm for affine hull model based similarity. It can reduce the dimension of data and improve the time consuming and accuracy.
- Classification. In our system, the Sparse Approximated Nearest Points (SANP) Hu *et al.* (2012) is used for classification.

In the previous chapters, we only introduced the technique separately. There are some details about the integration of the whole system (shown in Figure 7.1). The training and testing stage have some common process including pose estimation and face normalization. A projection matrix is learned using the algorithm in Section 6.2.1 to generate the discriminant feature for classification. Testing images are then projected to this feature space, and classification is applied on the projected data.

7.1.1 Manifold Alignment and Pose Estimation

We introduced the manifold alignment technique in Chapter 3 and Chapter 4. We use the robust version of Manifold alignment (from Chapter 4) in our system, since it has higher alignment performance. The output of alignment is the correspondence between two image sets. It can be used for face image set pose estimation, if we have a reference set in which all the sample labels are already known. Then we can transfer the reference sample label to the corresponding testing image set. In our case, the testing image set will have the pose label for aligned images.

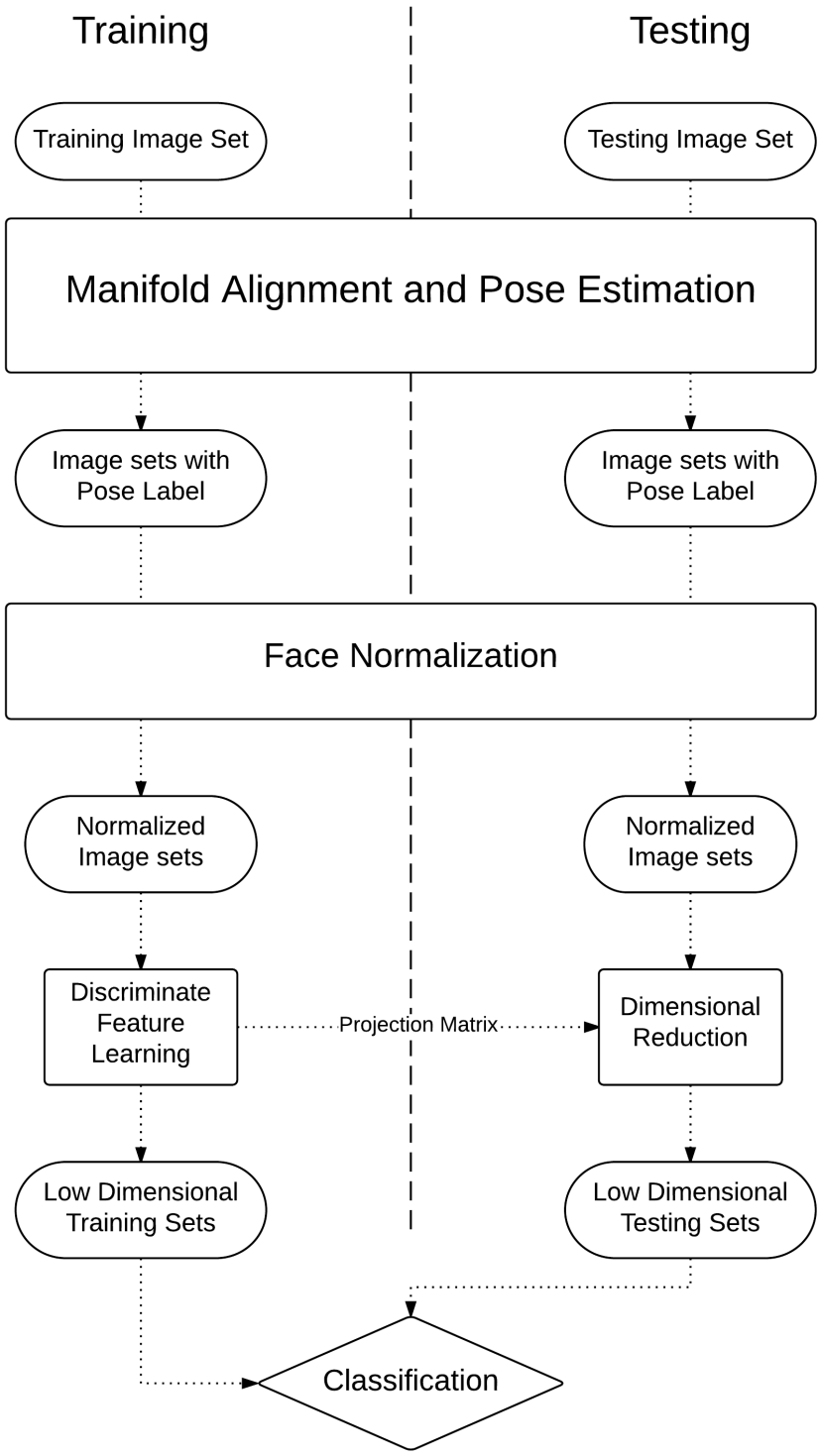


Figure 7.1: The flowchart of training and testing stage

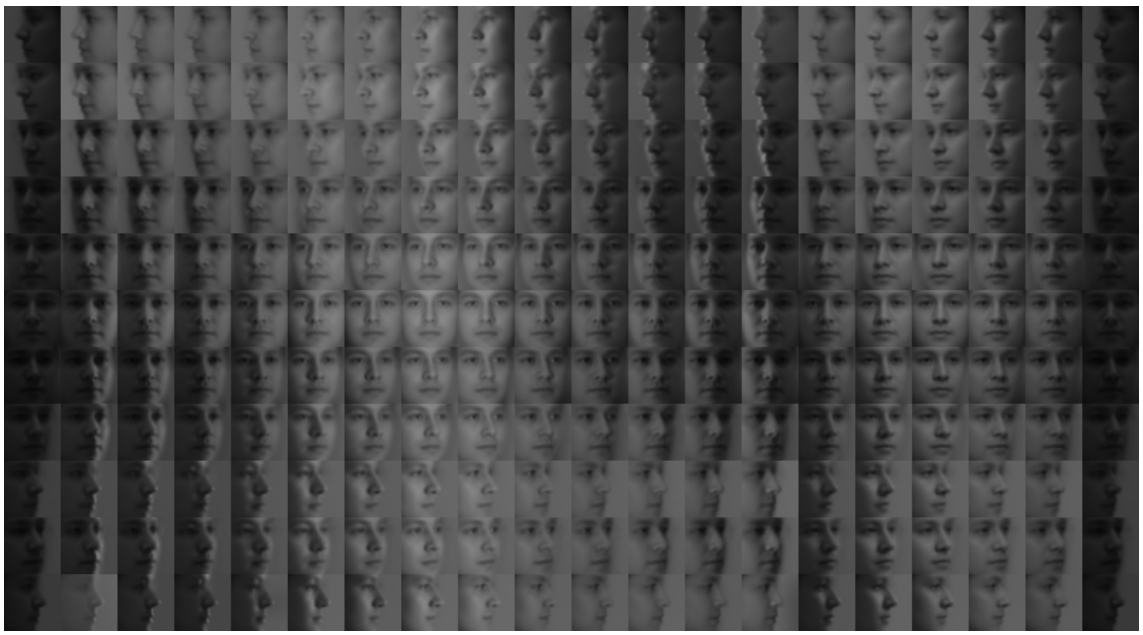


Figure 7.2: An illustration of the reference mean face image set. The illumination and pose are different in this image set.

We build a series of experiments on the Multi-PIE database. The reference set is built using the mean faces of 200 subjects in the Multi-PIE database. The reference set has the 11 poses and 20 illuminations conditions (shown in Figure 7.2). Every image set will be firstly aligned to this reference set and assigned the pose labels. In manifold alignment, it is possible that a few samples are outliers which have no assigned pose label. Next we use the LDA based pose estimation in Section 5.2.2 to detect the pose labels. Note that the alignment algorithm only assigns one possible label to one image, but LDA based estimation output three candidate poses for one image. This is because LDA pose estimation has lower accuracy and the Gaussian Processes Regression has the ability to handle uncertainty labels.

7.1.2 Pose Normalization

We use the Gaussian Processes Regression (GPR) in Chapter 5 as the transformation method. The aim of pose normalization in our FRIS system is slightly different, in the sense that our output of the normalization is an image set. The training process of GPRs is the same as we described in Section 5.2. For each pose, there is one GPR model to transform the corresponding pose image into the standard one. Images are normalized according to the pose obtained from alignment and pose estimation. For missing aligned

samples, we still use LDA based pose estimation. The only difference is that we do not use the product in Section 5.2.3 to integrate the different GPR results. Those outputs are considered as an individual sample of normalized image set. Since the variance v of each normalization result indicates the reliability of the output mean. We set a threshold to eliminate the implausible normalized output. To obtain an effective threshold, we use the experiment results in Section 5.3.1. In that experiment, we used labeled data to evaluate each GPR model that means all the results are based on correct pose. The experiment results show the confidence interval of the believable variance v. For each pose p the threshold v_M^p is set as a value that can include 95% of the known variances. We believe that if any variance of output is smaller than the threshold v_M^p, the normalization is believable and applied on correct estimated pose.

After the above process, all training and testing sets can be normalized into the standard pose. Then the normalized sets satisfy the assumption of local sample based similarity: all image sets are under similar conditions. Thus any the state of the art algorithm can be used. In our system, the state of the art local sample based algorithm Sparse Approximated Nearest Points (SANP) Hu *et al.* (2012) is used.

7.2 Experiments

In this section, the Multi-PIE database Gross *et al.* (2010)is again used to evaluate the effectiveness of our proposed framework. Since the Multi-PIE database has different labels for identity, pose and lighting conditions, it is suitable to simulate different conditions of FRIS. Our experiment will apply on different setup of pose condition, i.e., image set with or without similar pose. Following the setup in Section 5.3, the whole database is separated into three parts: pre-training set, training set and testing set. The pre-training set consists of the first 200 subjects with different poses which is same setup as in Section 5.3. The remaining images in database are the source of training and testing sets. We use the pre-training set to train the GPR models and generate mean reference set for alignment. The difference of each experiment is on how to choose training and testing sets.

The first experiment is to evaluate the simple case where the training and testing image set have similar poses. We randomly select different size of image sets from each subject to build the training and testing set. Images in the training set will not appear in testing set. That means there is no exactly same condition of face image in training and testing set. However, image sets have a good chance to appear in similar conditions, like the same pose with different lightings. We run the experiment for 10 times, each time the

Table 7.1: Comparison of the related FRIS algorithms to the proposed system on the Multi-PIE database in the case image set with similar pose. Average recognition rate (%) is shown in this Table.

Image Set Size	30	50	70
MSM	83.74	90.24	92.51
MMD	78.98	83.42	85.71
CHISD	92.32	96.84	98.25
SANP	93.54	97.38	99.1
The Proposed	94.32	96.68	97.5

training and testing set are re-selected randomly. The methods compared here include state of the art local sample based and structure based methods: Mutual Subspace Method (MSM) Yamaguchi *et al.* (1998), Manifold-Manifold Distance (MMD) Wang *et al.* (2008), Convex Hull based Image Set Distance (CHISD) Cevikalp and Triggs (2010) and Sparse Approximated Nearest Points (SANP) Hu *et al.* (2012). The experiment results are shown in Table 7.1. It can be seen that the performance of our system is better than the others when the image set size is small. That is because when image set are small, two image sets will have lower chance to have similar conditions. However our method can normalize face into the same pose to force images to lay in similar appearance. When the size of the image set becomes bigger, the performances of state-of-the-art methods are improving and exceed that of our system. That may be caused by the noise produced in the normalization process.

Since our aim is to solve the uncontrolled environment FRIS problem, we build another experiment in an extreme case that the training and testing set have no similar pose. We randomly select training set from left side images and testing set form right side pose (shown in Figure 7.3). Again the four methods in the last experiment are compared, and the results are shown in Table 7.2. In this experiment, there is no similar pose existing in both training and testing set, and only half partial conditions existing in all images which are not sufficient to discover manifold structure. From the results it is clearly demonstrated that both the structure based methods (MSM and MMD) and the local sample based methods (CHISD and SANP) are failed totally in this setup. On the contrary, our method still can achieve reasonably good and stable results.

From both experiments, it is evident that our system can handle both simple and extreme condition of FRIS problem. It is robust than most of state of art algorithm. And the experiments also prove our proposed system reached the aim and goal we set in Section

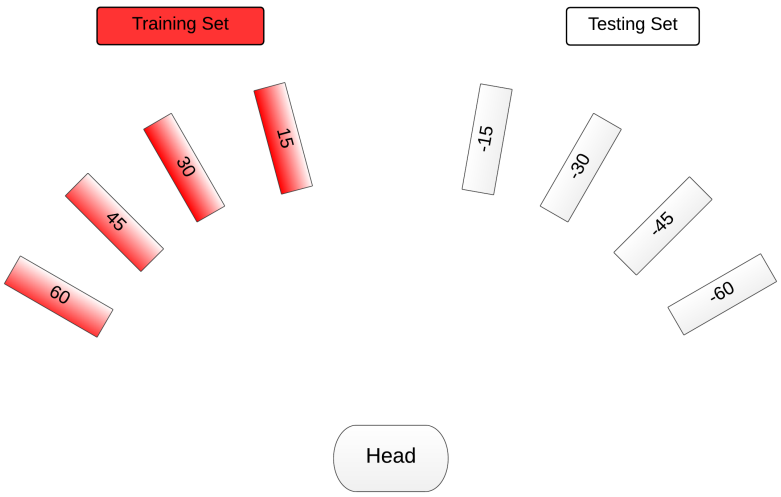


Figure 7.3: The building of training and testing image set for second experiment.

Table 7.2: Comparison of the related FRIS algorithms to the proposed system on the Multi-PIE database in second experiment that image sets without similar poses. Average recognition rate (%) is shown in this Table.

Image Set Size	30	50	70
MSM	7.01	10.65	9.69
MMD	15.51	21.78	22.15
CHISD	20.81	18.51	24.37
SANP	21.57	24.84	27.18
The Proposed	**70.42**	**74.15**	**74.82**

1.2.

7.3 Chapter Summary

In this chapter, we described the details of the our proposed complete FRIS system with the way of collaboration between each proposed component. Experiments are conducted to demonstrate the advantage of the proposed system. The capability of handling wild condition FRIS problem is proven by the convincing results. From the second experiment, we can see that our system significantly outperforms state of the art algorithms in uncontrolled environments. We will summarize the whole thesis in the next chapter and present the prospect of further improvement.

Chapter 8

Conclusions

This thesis proposed a system for the general FRIS problem. With the proposed system, there is no restriction on the image set and no assumption on relationship between training and testing set. It overcome the limitations of existing structure based and local sample based method that discussed in Chpater 2. There are three main techniques proposed in this system: manifold alignment, face pose normalization and image set based discriminant feature learning.

In Chapter 3, we proposed a novel method for unsupervised manifold alignment. It is assumed that a query set can be completely matched to a reference set. The main contribution is that we proposed a feature histograms for characterizing the local manifold geometry; and a matching algorithm to estimate accurate correspondences between two manifolds. The local histogram feature is used to establish an initial sparse correspondences. Then the initial correspondences are refined under an iterative optimization framework under manifold structure preserving constraints. A joint manifold is achieved within this framework. The proposed manifold alignment algorithm was applied to three different types of datasets and achieved significantly superior results when compared to existing state-of-the-art algorithms.

We then further improved the manifold alignment algorithm in Chapter 4. The significant contribution is that the proposed alignment algorithm is performed without the previous assumption on the correspondences between the two manifolds. The improvement is achieved by using an extended correspondence matrix with the ability to handle the outliers. We also improved the histogram-based features in Chapter 3. Based on such improvements, an embedding space is derived with the criteria that the distance between the two manifolds is minimized while maximally retaining the original structure of the manifolds. The elegance of this idea is that the extracted features can be directly applied in a generalized eigenvalue problem by using the Softassign technique. The alignment process is achieved by iteratively increasing the sparsity of the correspondence matrix until the two manifolds are correctly aligned. We demonstrate the effectiveness of our algorithm on different datasets. In comparison with state of the art algorithms and the MA-R method in Chapter 3, the results show the superiority of the proposed manifold alignment in terms

of visual effect and numerical accuracy.

In Chapter 5, a novel facial pose normalization approach is proposed based on Gaussian Processes Regression. The contribution of this work is to overcome the limitation of linear based regression methods. Unlike traditional regression technique, the output of Gaussian Processes Regression is a normal distribution with mean and variance. The variance can indicate the quality of the output. In training phase, instead of learning one single complete GPR model for whole pre-training set, we training each pose subset separately for each GPR model. This can significantly reduce the computation complexity. In testing phase, a LDA based pose estimation is used for assigning faces to the corresponding GPR model. Three prediction distributions are obtained by corresponding GPR models and then these models are integrated into a joint distribution, which can represent the final synthetic result. Experimental results demonstrate the advantages of the proposed method in comparison with the SRR method in face pose normalization and face recognition.

In Chapter 6, a new linear dimensionality reduction algorithm called margin preserving projections is proposed. We design this algorithm for convex hull model based methods, since experiments show that conventional methods are not suitable for FRIS problem. Based on the metric of convex hulls, it focuses on the relations between image sets. This is the key feature and a main advantage for the FRIS problems. Experiments on face image databases show that the proposed method produces better recognition accuracy and is less time consuming compared to related algorithms.

After discussion of each component, the whole system is demonstrated in Chapter 7. The collaboration of different components given the ability to handle extreme conditions of image sets. Experiments show that our system achieve similar high performance on simple case setup and significant improvement on extreme case, compared with existing state of art algorithms.

The new framework we proposed in this thesis is to overcome the problem that two image set have no overlap. After the whole process,image set will normalize into a variation reduced data. Image Set based recognition performance will improve greatly when using the normalized dataset. The framework can not only be suitable for FRIS problem, but also for other image sets based problem if there is an intrinsic model for each sets.

8.1 Future Study

Even though our FRIS system achieved significantly higher performance, there are still room for improvement in our system. In the manifold alignment, there are two problems worth attention. The first one is appropriately designed feature for face alignment. The local histogram feature is a general feature suitable for any kind of data; and HoG feature is not designed for face pose estimation. In an ideal case, a new feature should have the property associate with pose variation and ignore the identity information. For example, pose feature can be represented by the structure based landmark feature extracted from Active Appearance Model (AAM) Cootes *et al.* (1998). Another problem is how to handle the miss aligned images. We have not fully utilized the aligned joint manifold structure. The position of outliers in this manifold indicates the information of the intrinsic conditions of different variation. The variations should be able to be predicted based on this position.

In general, the critical point of this system is the quality of the face normalization. From experiments, Gaussian Processes Regression can be improved significantly when the pose differences are large. In that case, the transformation should be quite complicated and the GPR models may not be able to handle the complexity. Recent researches on deep learning point out a way to further improve the performance. Hierarchical networks are proved to have the capability to simulate complicated nonlinear regression and classification. Convolutional Neural Network (CNN) is a typical method using this idea to discover the intrinsic feature. The similar idea was also applied on GPR models in Deep Gaussian Processes Damianou and Lawrence (2013). We expect that the hierarchical structure of GPR models can handle the large difference pose normalization problem.

Bibliography

Arandjelović, O. and Cipolla, R. (2006). An information-theoretic approach to face recognition from face motion manifolds. *Image and Vision Computing*, **24**(6), 639–647. 10

Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., and Darrell, T. (2005). Face recognition with image sets using manifold density divergence. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 581–588. IEEE. 10

Ashraf, A. B., Lucey, S., and Chen, T. (2008). Learning patch correspondences for improved viewpoint invariant face recognition. In *CVPR*, pages 1–8. 17

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), 711–720. 13, 15, 81

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**(6), 1373–1396. 15

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, **28**(1), 235–242. 30, 49

Besl, P. and McKay, N. D. (1992). A method for registration of 3-d shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **14**(2), 239–256. 15, 17, 50

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer. 76

Blanz, V. and Vetter, T. (2003). Face recognition based on fitting a 3d morphable model. *TPAMI*, **25**(9), 1063–1074. 16

Blanz, V., Grother, P., Phillips, P. J., and Vetter, T. (2005). Face recognition based on frontal views generated from non-frontal images. In *CVPR*, volume 2, pages 454–461. 16

Castillo, C. D. and Jacobs, D. W. (2007). Using stereo matching for 2-d face recognition across pose. In *CVPR*, pages 1–8. 16

Cevikalp, H. and Triggs, B. (2010). Face recognition based on image sets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2567–2573. 11, 16, 74, 75, 80, 81, 89

Chai, X., Shan, S., Chen, X., and Gao, W. (2007). Locally linear regression for pose-invariant face recognition. *IEEE Transactions on Image Processing*, **16**(7), 1716–1725. 17

Chang, J.-M., Kirby, M., and Peterson, C. (2007). Set-to-set face recognition under variations in pose and illumination. In *Biometrics Symposium, 2007*, pages 1–6. IEEE. 10

Chang, K. I., Bowyer, W., and Flynn, P. J. (2006). Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(10), 1695–1700. 15

Chen, S., Sanderson, C., Harandi, M. T., and Lovell, B. C. (2013). Improved image set classification via joint sparse approximated nearest subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 452–459. IEEE. 12

Chen, Y. and Medioni, G. (1992). Object modelling by registration of multiple range images. *Image and vision computing*, **10**(3), 145–155. 15

Chertok, M. and Keller, Y. (2010). Efficient high order matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**(12), 2205–2215. 13

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 484–498. 93

Cui, Z., Shan, S., Zhang, H., Lao, S., and Chen, X. (2012). Image sets alignment for video-based face recognition. In *CVPR*, pages 2626–2633. 14, 15, 19, 38

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. 53, 57

Damianou, A. and Lawrence, N. (2013). Deep gaussian processes. In *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS-13)*, pages 207–215. 93

Egozi, A., Keller, Y., and Guterman, H. (2013). A probabilistic approach to spectral graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **35**(1), 18–27. 13

Fan, W. and Yeung, D.-Y. (2006). Locally linear models on face appearance manifolds with application to dual-subspace based classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1384–1390. IEEE. 10, 12, 74

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, **24**(6), 381–395. 24

Fukui, K. and Yamaguchi, O. (2005). Face recognition using multi-viewpoint patterns for robot vision. In *Robotics Research*, pages 192–201. Springer. 12

Gao, Y., Leung, M., Wang, W., and Hui, S. (2001). Fast face identification under varying pose from a single 2-d model view. *IEE Proceedings-Vision, Image and Signal Processing*, **148**(4), 248–253. 16

Gross, R. and Shi, J. (2001). The cmu motion of body (mobo) database. 79

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and Vision Computing*, **28**(5), 807–813. 57, 62, 68, 88

Hadid, A. and Pietikainen, M. (2004). From still image to video-based face recognition: an experimental analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 813–818. IEEE. 8, 74

Ham, J., Lee, D., and Saul, L. (2005). Semisupervised alignment of manifolds. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence, Z. Ghahramani and R. Cowell, Eds*, volume 10, pages 120–127. 14

Hamm, J. and Lee, D. D. (2008). Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th International conference on Machine learning*, pages 376–383. ACM. 13

Harandi, M. T., Sanderson, C., Shirazi, S., and Lovell, B. C. (2011). Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2705–2712. IEEE. 10, 12

He, X. and Niyogi, P. (2003). Locality Preserving Projections. *NIPS*. 15, 77, 81

Hoppe, H., Derose, T., Duchamp, T., McDonald, J. A., and Stuetzle, W. (1992). Surface reconstruction from unorganized points. *SIGGRAPH*, **26**, 71–78. 22

Hu, Y., Mian, A. S., and Owens, R. (2011). Sparse approximated nearest points for image set classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 121–128. IEEE. 11

Hu, Y., Mian, A. S., and Owens, R. (2012). Face recognition using sparse approximated nearest points between image sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(10), 1992–2004. 11, 16, 85, 88, 89

Huang, D., Yi, Z., and Pu, X. (2009). Manifold-based learning and synthesis. *TSMCB*, **39**(3), 592–606. 21, 26

Huang, D., Storer, M., De la Torre, F., and Bischof, H. (2011). Supervised local subspace learning for continuous head pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. 53

Huang, J., Yuen, P. C., Chen, W.-S., and Lai, J. H. (2007). Choosing parameters of kernel subspace lda for recognition of face images under pose and illumination variations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **37**(4), 847–862. 17

Kanade, T. and Yamada, A. (2003). Multi-subregion based probabilistic approach toward pose-invariant face recognition. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, volume 2, pages 954–959. 17

Kim, T. and Cipolla, R. (2009). On-line learning for maximizing orthogonality between subspaces and its application to image set-based face recognition. *IEEE Trans. Image Processing*, **19**(4), 1067–1074. 13

Kim, T.-K., Kittler, J., and Cipolla, R. (2006). Incremental learning of locally orthogonal subspaces for set-based object recognition. In *BMVC*, pages 559–568. 13

Kim, T.-K., Kittler, J., and Cipolla, R. (2007). Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**(6), 1005–1018. 12, 53, 57

Lee, K.-C., Ho, J., Yang, M.-H., and Kriegman, D. (2005). Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, **99**(3), 303–331. 79

Levine, M. D. and Yu, Y. (2006). Face recognition subject to variations in facial expression, illumination and pose using correlation filters. *Computer Vision and Image Understanding*, **104**(1), 1–15. 17

Li, A., Shan, S., Chen, X., and Gao, W. (2009). Maximizing intra-individual correlations for face recognition across pose differences. In *CVPR*, pages 605–611. 17

Li, A., Shan, S., and Gao, W. (2012). Coupled bias–variance tradeoff for cross-pose face recognition. *IEEE Transactions on Image Processing*, **21**(1), 305–315. 17

Little, D., Krishna, S., Black, J., and Panchanathan, S. (2005). A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *ICASSP*. 30, 50

Liu, X. and Chen, T. (2003). Video-based face recognition using adaptive hidden markov models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–340. 8

Maciel, J. and Costeira, J. P. (2003). A global solution to sparse correspondence problems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **25**(2), 187–199. 13

Mahmood, A., Mian, A., and Owens, R. (2014). Semi-supervised spectral clustering for image set classification. CVPR. 12, 19

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, **6**(4), 525–533. 64

Nishiyama, M., Yamaguchi, O., and Fukui, K. (2005). Face recognition with the multiple constrained mutual subspace method. In *Audio-and Video-Based Biometric Person Authentication*, pages 71–80. Springer. 10

Nishiyama, M., Yuasa, M., Shibata, T., Wakasugi, T., Kawahara, T., and Yamaguchi, O. (2007). Recognizing faces of moving people by hierarchical image-set matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 10

Pei, Y., Huang, F., Shi, F., and Zha, H. (2012). Unsupervised image matching based on manifold alignment. *TPAMI*, **34**(8), 1658–1664. 14, 15, 38

Rangarajan, A., Chui, H., and Bookstein, F. L. (1997). The softassign procrustes matching algorithm. In *Information Processing in Medical Imaging*, pages 29–42. 44

Rasmussen, C. E. (2006). Gaussian processes for machine learning. MIT Press. 63, 64

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, **86**(11), 2210–2239. 44

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500), 2323–2326. 13

Russ, J. (2002). The Image Processing Handbook, 4, Edition. 79

Sanderson, C., Bengio, S., and Gao, Y. (2006). On transforming statistical models for non-frontal face verification. *Pattern Recognition*, **39**(2), 288–302. 16

Satoh, S. (2000). Comparative evaluation of face sequence matching for content-based video access. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 163–168. IEEE. 11

Scholkopft, B. and Mullert, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*. 13

Shakhnarovich, G., Fisher, J. W., and Darrell, T. (2002). Face recognition from long-term observations. In *ECCV*, pages 851–865. 10

Sharma, A. and Jacobs, D. W. (2011). Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR*, pages 593–600. 17

Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, **19**(6), 1635–1650. 15

Tenenbaum, J., De Silva, V., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323. 13

Turk, M. A. and Pentland, A. P. (1991). Face recognition using eigenfaces. In *CVPR*, pages 586–591. 12, 15, 78, 81

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, **57**(2), 137–154. 79

Wang, C. and Mahadevan, S. (2008). Manifold alignment using procrustes analysis. In *ICML*, pages 1120–1127. 14

Wang, C. and Mahadevan, S. (2009a). A general framework for manifold alignment. In *AAAI Fall Symposium on Manifold Learning and its Applications*. 14, 25

Wang, C. and Mahadevan, S. (2009b). Manifold alignment without correspondence. In *IJCAI*, pages 1273–1278. vii, viii, 14, 15, 20, 25, 29, 30, 31, 32, 35, 38, 39, 47, 49, 50, 53, 54, 56, 57, 59

Wang, H., Li, S. Z., and Wang, Y. (2004). Face recognition under varying lighting conditions using self quotient image. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 819–824. IEEE. 15

Wang, R. and Chen, X. (2009). Manifold discriminant analysis. In *CVPR*, pages 429–436. IEEE. 13

Wang, R., Shan, S., Chen, X., and Gao, W. (2008). Manifold-manifold distance with application to face recognition based on image set. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1–8. 8, 10, 12, 81, 89

Wang, T. and Shi, P. (2009). Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, **30**(13), 1161–1165. 10

Wolf, L. and Shashua, A. (2003). Learning over sets using kernel principal angles. *The Journal of Machine Learning Research*, **4**, 913–931. 10, 12

Wolf, L., Hassner, T., and Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534. IEEE. 11

Yamaguchi, O., Fukui, K., and Maeda, K.-i. (1998). Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323. IEEE. 10, 12, 74, 89

Yang, M., Zhu, P., Van Gool, L., and Zhang, L. (2013). Face recognition based on regularized nearest points between image sets. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–7. 11

Yin, L., Chen, X., Sun, Y., Worm, T., and Reale, M. (2008). A high-resolution 3d dynamic facial expression database. In *FG 08*, pages 1–6. 50

Zhang, H., Zhang, Y., and Huang, T. S. (2013). Pose-robust face recognition via sparse representation. *Pattern Recognition*, **46**(5), 1511 – 1521. 17, 62, 69, 72

Zhang, X. and Gao, Y. (2009). Face recognition across pose: A review. *Pattern Recognition*, **42**(11), 2876–2896. 15

Zhang, Z. and Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, **26**(1), 313–338. 13

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, **35**(4), 399–458. 8