DISSERTATION

CLASSIFICATION ON THE GRASSMANNIANS: THEORY AND APPLICATIONS

Submitted by

Jen-Mei Chang

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2008

COLORADO STATE UNIVERSITY

April 8, 2008

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY JEN-MEI CHANG ENTITLED "CLASSIFICATION ON THE GRASSMANNIANS: THEORY AND APPLICATIONS" BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Dr. Gerhard Dangelmayr

Dr. Bruce Draper

Dr. Chris Peterson

Adviser: Dr. Michael Kirby

Department Head: Dr. Simon Tavener

ABSTRACT OF DISSERTATION

CLASSIFICATION ON THE GRASSMANNIANS: THEORY AND APPLICATIONS

This dissertation consists of four parts. It introduces a novel geometric framework for the general classification problem and presents empirical results obtained from applying the proposed method on some popular classification problems. An analysis of the robustness of the method is provided using matrix perturbation theory, which in turn motivates an optimization problem to improve the robustness of the classifier. Lastly, we illustrate the use of compressed data representations based on Karcher mean.

The success of this geometric framework builds upon the facts that the geometry and statistics of the Grassmannians are well-understood and families of patterns with a common characterization possesses discriminatory variations that are useful for classification. Under the right conditions, these families of patterns can be viewed as points on the Grassmannian where distances are available for classification. In this dissertation, we will make precise this connection, review various ways these metrics arise, and how to efficiently compute distances between points on this manifold.

Under this framework, we achieve excellent classification results for a variety of applications in face recognition and offer new insights to the problem in general. Attempting to break the method, we consider nonlinear data sets and images of extremely low resolutions. We are pleased to learn that the Grassmann method is robust against resolution reductions.

In order to understand how robust the Grassmann method is against perturbation, we draw support from matrix perturbation theory where we exploit the natural correspondence between linear subspaces and points on the Grassmannians. We are then led to formulate an optimization problem using these characteristics as an objective function and further connect this optimization criterion to the idea of *Fisher Linear Discriminant*

on general image sets. Numerical solutions obtained show promising improvements on the separability criterion.

The thesis is concluded by providing a novel algorithm that computes subject prototypical points using the Karcher mean on the Grassmannian. A lot of new ideas for geometric data analysis are generated through studies of old ideas. We hope that the suite of these frameworks and algorithms can collectively provide useful insights in studying geometric aspects of large data sets.

Jen-Mei Chang
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523-1874
Spring 2008

ACKNOWLEDGEMENTS

I gratefully acknowledge my advisor, Professor Michael Kirby, for his assistance, patience, and encouragement throughout the course of my work and my research at Colorado State University. His constructive suggestions helped to define the direction of this work. Under his guidance, I was able to extend my academic horizons and gain valuable industrial experiences that will unquestionably help my academic career in the long run.

My sincere appreciation is also due to Dr. Holger Kley and Dr. Chris Peterson, for influencing parts of this dissertation and contributing valuable time and ideas to this work. Their help and support not only improved the quality of this work but also boosted my confidence in wanting to do more.

I am deeply thankful to my mother, for paving the way for my education. Without whom, none of this would have been possible. It is because of her love and confidence in me, I am able to excel in every task that I encounter in life.

Lastly, I would like to thank my life partner, Ti-An Brenda Chiang, for her patience, encouragement, and moral support throughout my graduate study. It is her loving company, delicious meals, and endless smiles that helped me through the most difficult time of my graduate career.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The general techniques of pattern analysis can be seen in numerous interesting applications, such as hurricane modeling, biometrics, understanding of brain wave patterns, stock market trend modeling, landscape ecology, etc. The analysis of patterns in data has typically been a subject in statistics and engineering. Recently, however, fundamental mathematical theory in areas such as differential/algebraic geometry and topology have provided a new mathematical framework and insights for understanding large data sets residing in spaces of large ambient dimensions. Consequently, understanding the geometry of the data becomes an essential ingredient in algorithm selection.

As the technology of digital imaging grows, the task of organizing and analyzing high-dimensional data becomes increasingly important and difficult. For example, the development of high-speed digital cameras brings the need for sophisticated high-capacity memory storage. Data are often captured in high resolution but needed to be analyzed in coarse resolution. This is precisely why we are interested in both the understanding of large data sets in spaces of large dimensions and modeling of such data sets in much lower dimensions. Not only do we want to develop a way to correctly identify subject classes in a large data set, but we want to do so in the most efficient way.

An application that utilizes the geometry in patterns is the discipline of biometrics with significant emphasis on face, iris, fingerprint, and voice recognition. The federal government and industry have identified a pressing need to provide robust identity management tools and principles on how to employ these tools intelligently to meet national and international needs. This is because the existing identity management tools, such as passwords, personal identification numbers (PINs), tokens and cards, which are in use today for applications ranging from employee verification to theme park access, fail to

provide a definitive response. Hence, these traditional management tools are vulnerable to compromise and identity theft. Biometric systems present an advantage over these other tools since they are based on an individual's physiological and behavioral characteristics, making them more difficult to steal, copy, and compromise. A successful working example using biometrics is Federal Bureau of Investigation's (FBI's) Integrated Automated Fingerprint Identification System (IAFIS), which provides non-stop automated fingerprint search capabilities, latent search capability, electronic image storage and electronic exchange of fingerprints and responses in support of thousands of law enforcement organizations. The system contains biometrics records of more than 51 million criminal subjects and provides an open-set identification of submitted fingerprints. It normally returns responses within two hours of a criminal request and within 24 hours of civil fingerprint submissions [61]. On the other hand, the use of biometrics in border control and law enforcement is also abundant internationally (e.g. ePassport with iris recognition in UK, access of entry with fingerprint in Hong-Kong). In the meantime, academia is also assisting this development by implementing biometrics courses in both undergraduate and graduate curricula (e.g. West Virginia University, IEEE Tab Committee) while commercial applications of biometrics include bank surveillance and security as well as personal computer and cell phone securities (e.g. more commercial applications are currently being investigated at Korea's Biometrics Engineering Research Center (BERC)).

Because of the need for analyzing massive data sets, a lot of effort has been devoted to feature extraction and classifier building. Classification done in the reduced feature space simplifies computational complexity, and by considering an appropriate classifier, accuracy can be improved. Therefore, the heart of our quest for the ideal classification paradigm centers around these two topics. Ever since the early 1980's when automated pattern classification first became popular, it has always been a standard practice to consider the *subspace method* for pattern recognition tasks. Typically, a signal or waveform or picture contains much redundant information that may be removed by using, e.g., Karhunen-Loève (KL) transform. Each class then has its own set of representative features extracted from KL transform that forms a vector subspace (so-called *feature space*) of the original pattern space. The subspace method is a geometrically sound approach since these class subspaces can be used to classify an input sample into the best fitting class and they tell us something about the properties shared by all the items in

that category. For this reason, subspace method works extremely well when samples are selected from a uniformly distributed variation state. It is very fast to compute since the classification rule is based on a small number of inner products. On the other hand, if the classifier depends solely on a *single* input sample, the method will be sensitive to outliers and anomalies. Therefore, to improve and extend the traditional subspace method, we consider the case where multiple input (probe) samples per class are available.

In the traditional sense of the *subspace method*, class subspaces are formed for the gallery samples but not for the probe samples. Since geometry is present in the subject subspaces formed by performing KL on the gallery samples, why not consider the geometry of the subject subspaces obtained from performing KL on the probe samples? From this, we introduce the *many-to-many* or *set-to-set* subspace method.

Collections of patterns with a common characterization may be viewed as families of patterns and such characterization or variation can be modeled by subspaces. The collection of these raw patterns for a single subject can be mathematically represented by a matrix of dimension n-by-k, where k is the number of distinct patterns and n is the resolution of the patterns. The linear span of this matrix forms a k-dimensional vector subspace in $\mathbb{R}^n$, which can be realized naturally as a point on the Grassmannian $G(k, n)$. Now, performing classification of sets (of patterns) in their natural setting is equivalent to performing classification of points on the Grassmannians. Distance measures on the Grassmannians are well-established in this context and can be applied readily to this problem. Overall, classification on the Grassmannians is a mathematically simple framework that can be extended to any pattern classification problem that requires a many-to-many data comparison.

We will review and examine how different classification paradigms evolved in Chapter 2 and provide mathematical justifications to the set-to-set framework in Chapter 3 along with algorithmic details. We will formally introduce the notion of Grassmann separability in Chapter 4 and present two simple examples to illustrate the use of the Grassmann framework. More examples are shown in Chapters 5 and 6 where we present excellent classification results in a variety of face recognition applications. We further dwell on the ways to make the proposed method more efficient and applicable in Chapters 7 and 8 where two classes of linear transformations are introduced and applied to raw data. It is reasonable to ask ourselves whether or not this proposed approach is

robust against representations. Matrix perturbation theory that is relevant to studying this question is given in Chapter 9, from which we derive theories and obtain numerical solutions for improving robustness of the proposed classifier in Chapter 10. The dissertation is concluded by demonstrating how another geometric concept, Karcher mean, can be used to provide prototype representations in objective classification problems in Chapter 11. Finally, the concluding remarks are given in Chapter 12.

Chapter 2

SUBSPACE METHODS

As mentioned in the Chapter 1, classification using multiple instances of subject classes is the fundamental architecture of a classification scheme that we are interested in developing in this thesis. We will make the definition of a set-to-set classification paradigm precise in Chapter 2.4 immediately followed by the development of vector space framework for classification, single-to-single, and single-to-many classification paradigms in Chapters 2.1, 2.2, and 2.3, respectively. A formal introduction of the *Grassmann method* is given in Chapter 2.5 that will be referred to throughout the thesis. We will then review some state-of-the-art set-to-set classification techniques that have been successfully applied to face recognition problems in Chapter 2.6.

2.1 Vector Space Framework

The general approach to the pattern classification problem is to compare labeled instances of data to new, unlabeled exemplars. Implementation in practice depends on the nature of the data and the method by which features are extracted from the data and used to create a representation optimized for classification.

In general, an r-by-c gray scale digital image corresponds to an r-by-c matrix where each entry enumerates one of the 256 (on 8-bit machines) possible gray levels of the corresponding pixel. See Figure 2.1 for a graphical correspondence between an image J and its matrix representation X. Given a color image, the image can be represented in each of the three color channels, red (R), blue (B), and green (G), with a similar matrix. Different color channels contain different information regarding the image. Although gray scale (luminance) images are often used for classifications, this should not prohibit us from exploring the discriminatory information contained in the color channels.

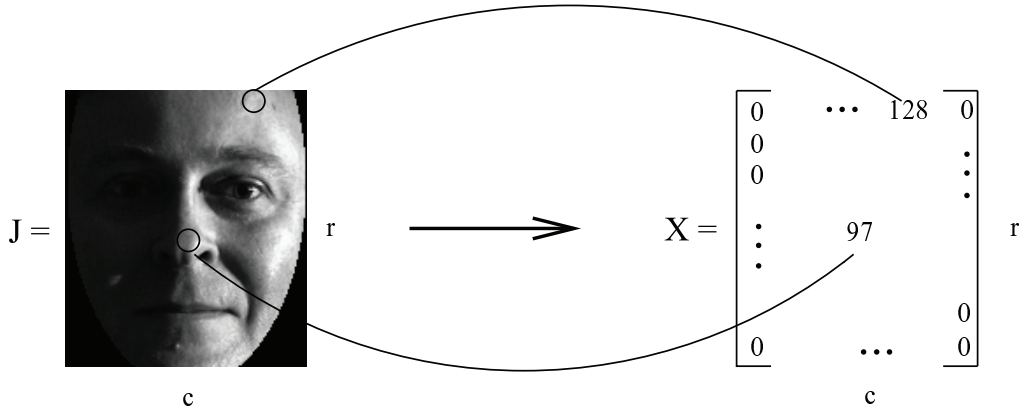


Figure 2.1: How an gray-scaled image can be realized as a mathematical matrix.

Now, partition the image matrix into columns, x_i, and concatenate the columns to obtain a long column vector as depicted below:

$$X = \begin{bmatrix} x_1 & | & x_2 & | & \cdots & | & x_c \end{bmatrix} \in \mathbb{R}^{r \times c} \quad \longrightarrow \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_c \end{bmatrix} \in \mathbb{R}^{rc \times 1}$$

Thus, an image J, whose matrix representation X, can be realized as a column vector of length equal to the product of J's resolutions. Thus, from now on, 2D images can be realized as a data point in high-dimensional space, the dimension of which is equal to the number of pixels in the image.

2.2 Single-to-Single Classification Paradigm

Given a set $G = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\right\}, x^{(i)} \in \mathbb{R}^n$, of m points and an assignment map $f : G \to C = \{1, 2, \ldots, N\}$, where $f(x^{(i)}) = j$ for some $1 \leq j \leq N$. When given a new point $p \in \mathbb{R}^n$, we can ask the question "What is the class of p among the known identities 1 through N?" A straightforward way to answer this question is to compute the pairwise distances between p and $x^{(i)} \in G$ for all $i \in 1, \ldots, N$. p is then assigned the identity of the class k such that $d(p, x^{(k)}) < d(p, x^{(i)})$ for all $i \neq k$. The set G is commonly known as the *gallery* and the point p is known as the *probe*. The choice of the distance function d can be any appropriate metric. Different metrics may provide different geometric quantities between p and the points in G. We refer to this type of comparison as the **single-to-single** classification scheme and a cartoon illustration is given in Figure 2.2.

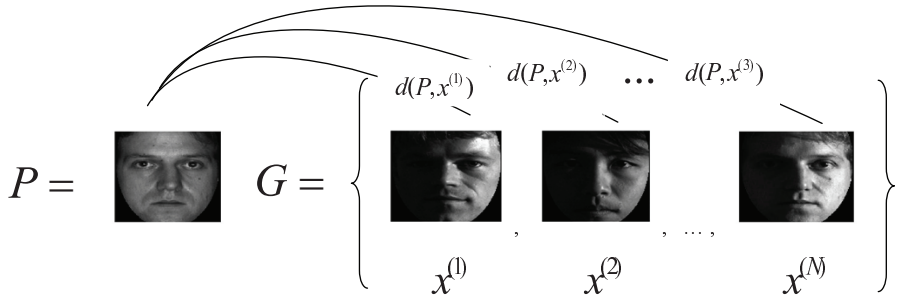


Figure 2.2: Structural illustration for a single-to-single classification paradigm, where a probe subject possesses a single instance of data and all of the gallery subjects possess a single data example as well.

2.3 Single-to-Many Classification Paradigm

Imagine now if every class in the gallery has multiple examples and the probe is still a single vector, i.e., $G = \left\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\right\}$, where $X^{(i)} = \left[x_1^{(i)}|x_2^{(i)}|\cdots|x_{i_k}^{(i)}\right]$ for all $i = 1, \ldots, m$. There are a number of ways to assign labels to a probe vector. A naive approach would be to apply the single-to-single vector comparison described in Chapter 2.2 to each of the vectors in $X^{(i)}$ for all i and assign p the identity of the class k where the minimal pairwise distance $d(p, x_j^{(k)})$ occurs for some j. This method overlooks the intrinsic variability of the sets in the gallery. For example, what if p is close to $\alpha_1 X^{(i)} + \alpha_2 X^{(2)}$? Moreover, this brute force method requires heavy computational resources when the number of examples for each class is large.

One way to get around both of these problems simultaneously is to consider the *subspace method* [65]. The method essentially assigns probe vector the class in gallery where the longest projection in terms of Euclidean norm occurs. Typically, a signal or waveform or picture contains much redundant information that may be removed by using, e.g., Karhunen-Loève (KL) transform. Each class then has its own set of representative features extracted from KL transform that forms a vector subspace of the original pattern space. Classification of a single probe vector is done by finding the best fitting class (in terms of projections) in the gallery. The low-dimensional subspace representation of faces was first proposed by Sirovich and Kirby [76] in 1987 as an application of the *Principal Component Analysis (PCA)* and later popularized by Turk and Pentland [84]. There have now been many applications of this idea in the context of face recognition. The central idea of PCA in face recognition is to find a small set of eigenfaces that best

represent points in a face data set. This statistical method is particularly suited for data with linear structure and can be extended (e.g. Kernal PCA [95]) and applied to nonlinear data. However, despite its relatively simple methodology, PCA-based methods perform poorly when images are acquired in uncontrolled settings, such as unstable light source directions and unpredictable facial movements of the subjects. This is primarily due to the fact that the recognition is based on a single-shot. For instance, a test image that is captured under a different illumination condition than the training images possesses a variability that is not inherited in the training images, therefore more likely to be wrongfully classified. This type of comparison is known as the single-to-many comparison. For an Automatic Face Recognition (AFR) scheme to be applicable in real-world situations, it needs to take into account of unforeseen variability in pose, illumination, and facial expression, etc. In fact, it is much easier to ask someone to perform random head motions under varying illumination conditions than to request the person to perform strictly defined motions under controlled lighting conditions. We refer to this classification scheme as **single-to-many** and a cartoon illustration of this type is given in Figure 2.3.

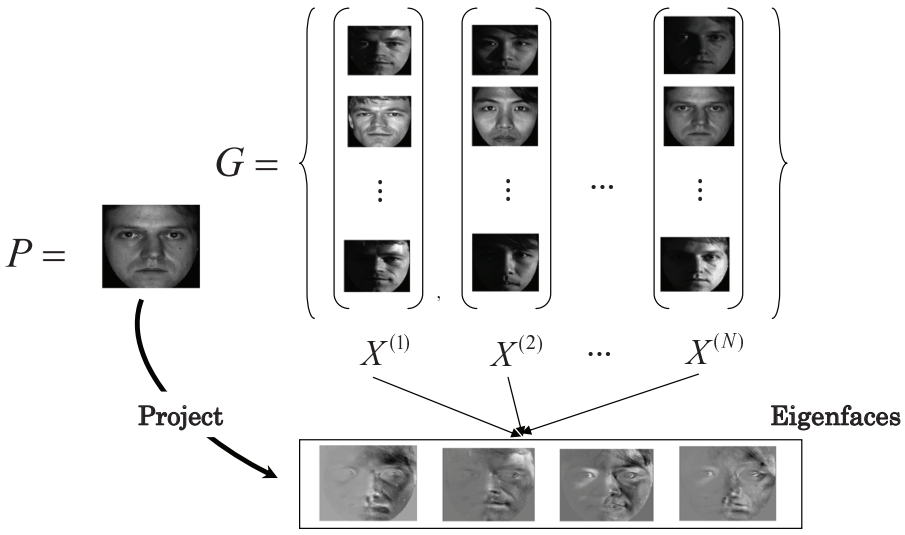


Figure 2.3: Structural illustration for a single-to-many classification paradigm, where a probe subject possesses a single instance of data while one or more of the gallery subjects possess multiple examples. In this figure, a set of basis vectors that span a feature subspace of significantly lower resolution is first obtained. Classification of a probe image is done by projecting the probe onto the feature space and assigning identity based on the Euclidean differences with the subject classes.

2.4 Many-to-Many Classification Paradigm

It is then natural to consider a set of images and the subspace it forms. Since the subspace can effectively represent the distribution of the changes, comparing subspaces is therefore more stable against the influence of illumination variations, head motions, and facial expressions. Therefore the notion of many-to-many comparison arises. In this setting, we consider collecting multiple examples of a probe class so that $p = [p_1|p_2|\cdots|p_k]$ is an n-by-k matrix with each p_i being an element of $\mathbb{R}^n$ and keeping the structure of the gallery the same as before. Now both the probe and each class in the gallery have multiple examples to be used in the process of comparison. We refer to this type of classification scheme as **many-to-many** or **set-to-set** with a cartoon illustration of this type given in Figure 2.4.

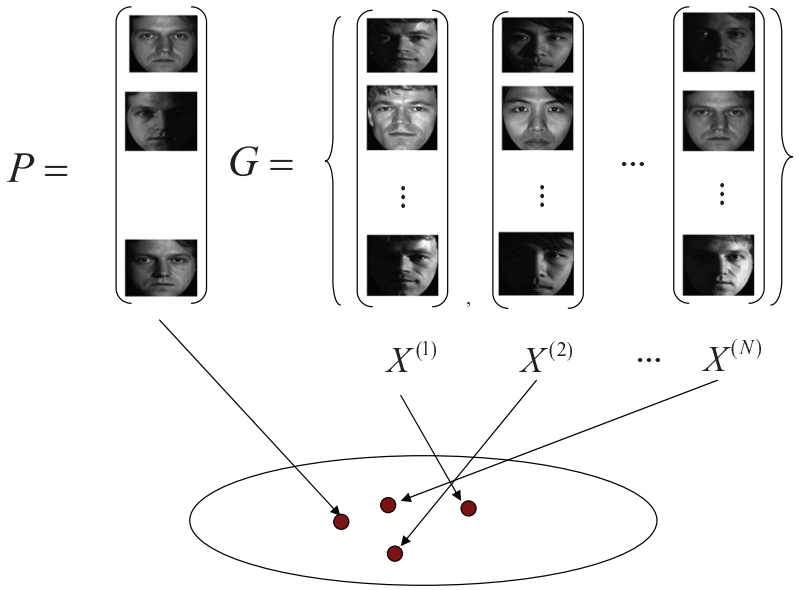


Figure 2.4: Structural illustration for a many-to-many classification paradigm, where probe and gallery subjects possess multiple instances of data. In this figure, each set of images (selected from probe or gallery) can be realized as a point on some parameter space. Therefore, classification of a probe class requires only calculation of the pairwise distances between these points.

Notice that the number of data in any probe class does not necessarily have to equal the number of data in any gallery class. For example, often times the gallery images are collected in a controlled setting whereas probe images are recorded in a much noisier environment. It is not uncommon to have fewer probe images than gallery images for

recognition. Thus, when we speak of many-to-many paradigm, we do not assume a symmetric comparison. Rather, it is particularly interesting to investigate the minimum number of images required for either the probe or gallery in order to ensure a perfect classification in a general *x-set to y-set* classification problem.

Under this paradigm, one can choose to solve the general classification problem by computing the pairwise distances between each point in the probe with each point in the gallery as mentioned in Chapter 2.3. On the contrary, if we form a subspace from the instances of the probe and for each class in the gallery, then this many-to-many comparison only requires m pairwise distance calculations where m is the number of image sets. It is reported in [29] that when given two average-length sequences, determining the match score by single-to-single matching is roughly 400 times more expansive than a single many-to-many match. Nevertheless, this still raises the concerns about the computational cost for the many-to-many matching.

Learning from multiple instances allows us to construct subspaces that embody the natural variability of the data. Incorporating these subspace representations into a classification paradigm provides geometric insight in understanding the neighboring relationships between subjects in a data set. In a world where parallelization has become a reality and where we have no upper bound on the computational resources, why should we limit ourselves to an algorithm that is relatively faster but inaccurate and less robust (single-to-many) when we have access to an algorithm that is otherwise accurate and robust (many-to-many).

2.5 The Grassmann Method

The geometry of the data sets affects the fundamental design of a classification algorithm. For example, it is reasonable to quantify the distances between points on the xy-plane with Euclidean metric but rather foolish to do so among a set of points on the 2-sphere, S^2, using the same metric. In any case, the optimal choice for the metric is the appropriate "geodesic" on that space. In this section, we introduce a novel geometric framework, so-called *Grassmann method*, that is suitable for the many-to-many classification paradigm.

Let K be a field (such as the field of real numbers), and let V be a vector space over K. As usual, we call elements of V vectors and call elements of K scalars. Suppose

that W is a subset of V. If W is a vector space itself, with the same vector space operations as V has, then it is a *subspace* of V. Now, let $V \in \mathbb{R}^n$, then a collection of k distinct data points give rise to a matrix X of size n-by-k where each column of X is a distinct data point. Further denote $\mathcal{R}(X)$ to be the column space (or range) of X. We can always associate a basis (e.g., obtained by QR-decomposition or Singular Value Decomposition) to the column space of X, which is a k-dimensional linear subspace of $\mathbb{R}^n$. Let $G(k, n)$ denote the Grassmann manifold (Grassmannian) parameterizing k-dimensional real vector subspaces of the n-dimensional vector space $\mathbb{R}^n$, then the many-to-many classification problem can be transformed to a problem on $G(k, n)$ if we realize the linear span of a set of k images as a k-dimensional vector subspace of the space of all possible images at a given resolution. Our objective is to match an unlabeled *set* of images by comparing its associated point with a collection of given points on $G(k, n)$. As a consequence of the encoding of sets of images as points on a Grassmann manifold we may avail ourselves of a variety of well-known distance measures between points on the manifold.

The Grassmann method is a linear method since the Grassmannian is a parameter space for linear subspaces. Once we have established a concrete paradigm for classification on linear data sets, we can attempt to solve the nonlinear cases by exploring their linear structures. For example, nonlinear data manifolds exhibit local linear structures characterized by the tangent spaces. By associating tangent spaces with points on the Grassmann manifold, the method of the Grassmann can proceed as depicted before.

In summary, the general approach to transforming a conventional classification problem into one on the Grassmannians requires two basic steps:

1. Find linear subspaces by forming bases for the data and consequently realize these subspaces as points on the Grassmannians.

2. Assign class label of probe points by determining neighborhood relationship using the distances of corresponding points on the Grassmannians.

2.6 Face Recognition Using Many-to-Many Framework

Traditionally, work on face recognition has focused on comparisons between single still images. Recently, however, Experiment 2 of the Face Recognition Grand Challenge

considered many-to-many comparisons [66]. Sets of four images were compared based on four by four (sixteen total) single image comparisons. This was enough to raise the recognition performance of the baseline algorithm from about 66% on single still images to about 88% on the four-way comparison. Works of [29, 90, 2, 50, 49], with images selected from video sequences, and [13, 14, 15, 16] with 2-dimensional still images have promising results supporting the use of set-to-set image comparisons.

The earliest work on using set-to-set image comparison in object recognitions can be found in [94] and [75]. A crude partition of the contemporary approaches in multi-set image comparison yields two branches: model-based and sample-based methods [50]. Three well-known representatives in the model-based approaches utilize the concepts of *Tangent Distance* [75, 12], *Manifold Density Divergence* [73, 2], and *Canonical Correlation Analysis (Principal Angles)* [94, 13]. Sample-based methods, such as using image similarities, generally require heavy computational capacities as they require a comparison of every pairwise samples of any two sets. Furthermore, it does not take into account the natural variability of the data due to the 3D nature of the observed objects and it generally performs worse than the subspace methods. Thus, we will concentrate on the discussions of model-based methods that will eventually lead us to discover fundamental properties of a data set.

When the manifold is made up of images obtained via affine transformations (e.g., rotations, translations and scaling), the manifold has the differential topology of a Lie Group. Thus it is possible to calculate an optimal linear approximation of an image p on the manifold that captures the relevant linear effects of deformation. This subspace, called the tangent space, typically offers a low-dimensional characterization of the image and contains nearly the same information as the original manifold for small transformations, since tangent space $T_p(\mathcal{M})$ is the best linear approximation of a manifold $\mathcal{M}$ at the point $p \in \mathcal{M}$. Measuring the distance between two images can then easily be done by forming their respective deformation manifold (multi-set) and tangent space followed by finding the gap distance between the two tangent spaces. This distance is called the two-sided tangent distance [75]. One known drawback of the method is that the accuracy of the tangent distance depends on the point of tangency [28, 43, 85]. This adversely makes the tangent distance method less robust and sensitive to outliers. As an alternative, Fitzgibbon and Zisserman in [29] and Chang in [12] proposed the uses of *Joint*

Manifold Distance and *Subspace Distance*, respectively, that are essentially the subspace analogue of tangent distance. In essence, tangent spaces are replaced by subspaces in the computation of the tangent distance.

When using manifold density divergence to compare sets, each set of images is typically represented by a Gaussian distribution function and compared against other sets by the Kullback-Leibler Divergence (KLD) [73, 2]. This method works well when the density distribution of the sets is *a priori* known and when the training and testing sets have strong statistical correlations.

Yamaguchi et al. [94] used the minimal principal angle (or maximum correlation) between training and testing subspaces to capture the similarity between the two sets and named their method *Mutual Subspace Method*. Since then, the concept of canonical correlation has been widely used. For example, the central idea in *Constraint Mutual Subspace Method (CMSM)* [33] is that by projecting the probe and gallery subspaces to a constrained subspace (generated by considering the difference subspaces), the new principal angle preserves the difference between people while excluding unnecessary components for recognition, namely, undesirable variations. In *Multiple Constrained Mutual Subspace Method (MCMSM)* [63], multiple constrained subspaces are created using methods of ensemble learning (bagging and boosting) where probe and gallery subspaces are projected onto and MSM is used to classify. The combined similarity between two subspaces is given by combining the similarities calculated on each constrained subspace. In *Hierarchical Image-Set Matching (HISM)* [64], sets of face images of the same individual are acquired from multiple cameras and integrated. A distribution of each individual is created and compared using MSM. In [83, 82], authors proposed a new feature extraction method and a new feature fusion strategy based on the generalized canonical correlation analysis (GCCA). Kim et al. proposed in [50] a method that maximizes the canonical correlations of within-class sets and minimizes the canonical correlations of between-class sets that is inspired by *Linear Discriminative Analysis*.

In another extension of CCA, Wolf and Shashua [90] constructed a positive definite kernel $f(A, B) = \prod_{i=1}^{k} \cos^2 \theta_i$ that is used to compare two image sets A and B where θ_i's are the principal angles between the column spaces of A and B. Kim et al. [49] addressed one of the major shortcomings of MSM-based methods: *ad-hoc* fusion of information contained in different principal angles. They proposed using principal angles to build

simple weak classifiers which are then combined using the AdaBoost algorithm [32]. The algorithm learns a weighting of decisions cast by weak learners and by examining the magnitude of these weights, it is easy to see which principal angles are more significant and should be used in building a similarity measure. They go on and extend this idea to build a similarity measure that would capture the nonlinearity in the data. Table 2.1 gives a comprehensive and comparative summary of these current state-of-the-art face recognition algorithms that are implemented with principal angles. It goes without saying that the use of principal angles has been widely spread and future research on how to further employ CCA is highly anticipated. Future research on face recognition is likely to take much more seriously the question of how to best compare sets of images.

Two major distinctions between the prior works and the proposed Grassmann framework are worth pointing out. First, with the exception of [94] and [90], all of the work presented in Table 2.1 require some form of training prior to classification. This could be subspace training [33, 63, 64], feature extraction [82, 50], or classifier training [49]. On the contrary, when applying the Grassmann method to face recognition problems, it can be implemented without training while obtaining excellent classification outcomes. Secondly, these authors have not put their work in the context of Grassmann manifolds, therefore limiting the geometric scope of the ideas. By introducing the idea of the Grassmannian and pre-existing tools to quantify its geometry, we are able to come up with many new and useful tools, such as Karcher mean on the Grassmannians, to study the geometry of the data sets.

Table 2.1: A comprehensive and comparative summary of current state-of-the-art algorithms that are implemented with principal angles.

Method	P.A. Used	Type(s) of Variation	Illumination	Pose	Resolution	Apparatus	Training
MSM [94]	min. p.a.	facial expr., m-ments	single	yes	30×30	camera	no
CMSM [33]	min. p.a.	head dir./position	8 (3 lamps on/off)	minor	15×15	camera	yes
MCMSM 1 [63]	min. p.a.	arb. facial pose	10 (7 lgt srcs)	minor	30×30	video (5 f/s)	yes
MCMSM 2 [63]	min. p.a.	arb. facial pose	uniform	minor	30×30	video (5 f/s)	yes
HISM 1 [64]	min. p.a.	pose, illum.	normalized	corrected	64×64	video (15 f/s)	yes
HISM 2 [64]	min. p.a.	occlu. pose, illum.	normalized	corrected	16×16	≥ 1 cam.	yes
Kernal P.A. [90][a]	first 20 p.a.	head and facial	yes (n/details)	yes (n/details)	35×47	video	no
DCC [50]	sum of c.c.[b]	arb. motion	histogram equal.	yes (n/details)	20×20	video	yes
GCCA 1[c] [82]	all	time, facial expr.	no	tilt, rot., scale	112×92	camera	yes
GCCA 2[d] [82]	all	pose, illum., expr.	no	tilt, rot., scale	120×91[e]	camera	yes
BoMPA [49]	wgted all[f]	arb. motion[g]	histogram equal.	yes	50×50	video	yes

[a]For the face recognition experiment, instead of using the positive definite measure proposed in the paper, the mean of the smallest 20 principal angles are used as a similarity measure.

[b]c.c. denotes canonical correlations.

[c]ORL Database.

[d]Yale Database.

[e]There is a mention of using a cubic wavelet transformation on the original images using Daubechies orthonormal wavelet to create 12×15 sub-images.

[f]The weights are determined by AdaBoost algorithm.

[g]Subjects were instructed not to perform extreme facial expressions, but many users talked or smiled during the acquisition.

Chapter 3

DISTANCE MEASURES ON THE GRASSMANNIANS

In a many-to-many classification paradigm, we are interested in knowing how far apart given image sets are. This question can be cast into a question regarding the subspaces formed by considering the linear spans of the image sets. What people do next is exactly where the differences occur. As described earlier in the prior works, the idea of principal angles (a.k.a. the canonical correlation analysis) is widely used as a tool to measure variations and similarities between subspaces. For example, Kim et al. used the sum of all the canonical correlations between two subspaces in [50] as a similarity measure between the two subspaces and Yamaguchi et al. used the maximum canonical correlation between two subspaces as a similarity measure in [94] as well. These two measurements are merely "similarity" functions and do not qualify as actual "distance" functions, since they fail the definition of being a metric. We propose to look at the problem from a geometric point of view. By viewing these subspaces as points on the Grassmann manifold, we can avail ourselves of the actual "distance" measures naturally available on the Grassmann manifold. Because of this connection, not only can we determine how far apart subspaces are from each other using actual distance functions, but afford new geometric insights in designing classification algorithms that incorporate the geometry of the Grassmann manifold.

In this chapter, we will first show the connection between matrix representation of images to matrix representation of points on the Grassmann manifold in Chapter 3.1, then discuss how various Grassmannian distances arise from embedding the Grassmannians into various types of spaces as well as in the context of linear algebra in Chapter 3.2. Finally, a quantitative measure for determining how well a distance function performs

in classification problems is introduced in Chapter 3.3 and numerical algorithms for calculating distances between points on the Grassmann manifold are given along with a complexity analysis in Chapter 3.4.

3.1 Matrix Representation For Points on The Grassmann Manifold

As mentioned in Chapter 2.1, an $r \times c$ gray scale digital image corresponds to an $r \times c$ matrix where each entry enumerates one of the 256 (on 8-bit machines) possible gray levels of the corresponding pixel. After concatenation by columns, an image vector of length $n = r \cdot c$ can be seen as a point in $\mathbb{R}^n$. In the original subspace method, this point will then be projected into a feature space of a much lower dimension for classification. We will, however, group k (generally independent) example images of a subject and consider the k-dimensional *feature subspace* they span in $\mathbb{R}^n$. The connection between this linear subspace to a point on the Grassmann manifold will be made precise next.

Definition 3.1.1. The Grassmannian $G(k, n)$ or the Grassmann manifold is the set of k-dimensional subspaces in an n-dimensional vector space K^n for some field K. i.e.,

$$G(k, n) = \{W \subset K^n \mid \dim(W) = k\}.$$

Let V be a vector space of dimension n with basis $\{e_1, \ldots, e_n\}$, then for $k \geq 0$ we can define a new vector space over K:

$$\Lambda^k V = \begin{cases} K, & \text{if } k = 0; \\ 0, & \text{if } k > n; \\ \text{space with basis} & e_{i_1} \wedge e_{i_2} \wedge \ldots \wedge e_{i_k}{}^1,\ 1 \leq i_1 < \ldots < i_k \leq n. \end{cases}$$

Furthermore, if we define

$$V \times V \times \cdots \times V \xrightarrow{\Phi} \Lambda^k V$$

by

$$\Phi(e_{i_1}, \cdots, e_{i_k}) = \begin{cases} 0, \text{ if } i_j = i_l \text{ for some } j \neq l; \\ sgn(\sigma)\left(e_{i_{\sigma(1)}} \wedge \cdots \wedge e_{i_{\sigma(k)}}\right), \text{ otherwise;} \end{cases}$$

where σ is the unique permutation of $\{1, 2, \ldots, k\}$ such that $i_{\sigma(1)} < \cdots < i_{\sigma(k)}$, then it is easy to see that Φ is alternating. We can then write $\Phi(v_1, v_2, \ldots, v_k) = v_1 \wedge v_2 \wedge \cdots \wedge v_k$ for all $v_i \in V$.

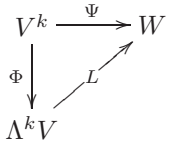
Theorem 3.1.1. *Let V be a vector space and*

$$\Psi : V \times V \times \cdots \times V = V^k \longrightarrow W$$

a multilinear map from V's to W, which is alternating. Namely,

$$\Psi(v_{\sigma(1)}, \cdots, v_{\sigma(k)}) = sgn(\sigma)\Psi(v_1, \cdots, v_k), \quad \forall \sigma \in S_k,$$

where S_k is the set of permutations of k elements. Then there exists a unique linear map $L : \Lambda^k V \to W$ such that $\Psi = L \circ \Phi$. i.e., $V^k \to \Lambda^k V$ is unique up to unique isomorphism and satisfies the following commutative diagram.



Then by the universal property of $\Lambda^k V$, any map

$$\Phi : V \times V \times \cdots \times V \longrightarrow \Lambda^k V$$

that is multilinear and alternating is unique up to isomorphism. Thus we can talk about *the kth exterior power* of V over a field K.

The Grassmannian can then be viewed as a subset of projective space, $\mathbb{P}(\Lambda^k V)$, via the Plücker embedding:

$$\begin{array}{ccc} G(k,n) & \xrightarrow{Pl} & \mathbb{P}(\Lambda^k V) \\ W & \mapsto & \Lambda^k W \end{array}$$

where $\dim(\mathbb{P}(\Lambda^k V)) = \binom{n}{k} - 1$. This map is injective. The homogeneous coordinates on $\mathbb{P}(\Lambda^k V)$ are called the Plücker coordinates on $G(k,n)$. Moreover, $Pl(G(k,n))$ = class of totally decomposable multivectors, is a subvariety of $\mathbb{P}(\Lambda^k V)$ [41].

In coordinates, we can explicitly represent a plane $W \in G(k,n)$ by a unique matrix up to a change of basis transformation. Let W be a k-dimensional vector subspace of V with basis $f_j = \sum_{i=1}^{n} b_{ij} e_i$, $j = 1, 2, \ldots, k$ and let $B = (b_{ij})$. Moreover, assume U is the standard affine open subset of $\mathbb{P}(\Lambda^k V)$ whose first $k \times k$ minor is nonzero. Then

$$B = \left[b_{ij}\right] \sim \begin{bmatrix} I_k \\ ---- \\ B'_{(n-k)\times k} \end{bmatrix}.$$

The matrix B is determined up to right multiplication by an invertible $k \times k$ change of basis matrix. B uniquely determines B', and B' uniquely determines W. Then the entries of B' give the bijection of $U \cap G(k,n)$ with $K^{k(n-k)}$, i.e., $G(k,n)$ is covered by affine space of dimension $k(n-k)$. Consequently, $\dim(G(k,n)) = k(n-k)$ when the Grassmannian is realized as a submanifold of a projective space.

It is now clear that points in the Grassmannian are equivalence classes of n-by-k orthonormal matrices, where two matrices are equivalent if their columns span the same k-dimensional linear subspace, i.e.,

$$G(k,n) = \{[p] \mid p \sim q \text{ if and only if } \quad q = Q^T p \quad \text{for some} \quad Q \in O_k\},$$

where p and q are n-by-k orthogonal matrices and O_k is the group of k-by-k orthogonal matrices.

Therefore, the Grassmann manifold $G(k,n)$ can be identified as the quotient group $O_n/(O_k \times O_{n-k})$. Despite this abstract mathematical representation of the Grassmannian, one may choose to represent a point on the Grassmannian by specifying an arbitrary orthonormal basis stored as an n-by-k matrix. Although this choice of the orthogonal matrix is not unique for points on the Grassmannian, it does give rise to a k-dimensional linear subspace that is obtained via the column space of the matrix and will serve as a representative of the equivalence class on the computer [24].

3.2 Grassmannian Distances

In this section, we will present an overview of how distances may be computed between subspaces, or points on $G(k,n)$. We will do this in a few steps. First, a characterization of the distance between subspaces by defining gap functions is given. We then go on to show that all gap functions are equivalent and induce the same gap topology. Finally, a discussion on unitarily invariant metrics and their relations to principal angles is presented. We will conclude the section by noting that ultimately any unitarily invariant norm on two subspaces can be written in terms of some symmetric gauge function of the principal angles between the subspaces [67]. In particular, we will focus our attention on characterizing (with principal angles) the *Grassmannian metrics* that arise naturally from realizing the Grassmann manifold as subsets of various spaces.

It is natural to first consider the distance between a point in $\mathbb{C}^n$ and a subspace of $\mathbb{C}^n$, then develop the notion of distance between two subspaces. Also note that the definitions and results given here can easily be given in the space of reals. Much of this presentation follows [79], including Definitions 3.2.1 — 3.2.3.

Definition 3.2.1. Let $\mathcal{X}$ be a subspace of $\mathbb{C}^n$ and $y \in \mathbb{C}^n$. If ν is a norm on $\mathbb{C}^n$, then the $\nu-$*distance* between y and $\mathcal{X}$ is the function

$$\delta_\nu(y, \mathcal{X}) := \min_{x \in \mathcal{X}} \nu(y - x). \tag{3.1}$$

Now, the distance between two subspaces of $\mathbb{C}^n$ follows naturally from Equation (3.1).

Definition 3.2.2. Let $\mathcal{X}, \mathcal{Y} \in G(m, n)$ and let ν be a norm on $\mathbb{C}^n$. Then the ν-*gap* between $\mathcal{X}$ and $\mathcal{Y}$ is the number

$$\rho_{g,\nu}(\mathcal{X}, \mathcal{Y}) := \max \left\{ \max_{\substack{x \in \mathcal{X} \\ \nu(x)=1}} \delta_\nu(x, \mathcal{Y}), \max_{\substack{y \in \mathcal{Y} \\ \nu(y)=1}} \delta_\nu(y, \mathcal{X}) \right\}. \tag{3.2}$$

Note that the gap function does not need to be a metric. At this point, the gap functions closely depend on the norm function. It would be convenient if all gap functions give rise to the same topology regardless the choice of the norm so we can speak of *the* subspace topology. In fact, it turns out that all gap functions are equivalent in the same sense that all norms are equivalent. Once we can establish that one gap function is a metric, it will then follow from their equivalence that all gap functions generate the same topology. With that being said, we will only present the results to show that the gap function $\rho_{g,2}$ is a metric, which will be used extensively to serve as a foundation in the development of perturbation theory for subspaces. As a preliminary step, we will review the notion of *principal angles* (a.k.a. *canonical angles*) between subspaces as well as some of the results about projection matrices in conjunction to the principal angles.

The concept of principal angles was introduced by Jordan in 1875 [44] and Hotelling introduced the recursive definition in 1936 [42]. During the past century, numerous researchers have developed theories and algorithms to quantify principal angles. Two well-known algorithms that are used extensively in various applications are based on *Singular Value Decomposition (SVD)* and *CS Decomposition*. Björck and Golub gave a numerically stable algorithm in 1973 in terms of the SVD of the matrices characterizing

the subspaces [9], while Stewart cast the problem in the form of CS decomposition [78] which he first introduced in 1977 [77]. We will adapt recursive definition of principal angles given in [9].

Definition 3.2.3. If x and y are two nonzero vectors in $\mathbb{C}^n$, then the angle between x and y is defined to be

$$\angle(x, y) := \cos^{-1} \frac{\left|y^H x\right|}{||x||_2||y||_2}.$$

Definition 3.2.4. [9] If $\mathcal{X}$ and $\mathcal{Y}$ are two vector subspaces of a unitary space $\mathbb{E}^n$ such that $p = \dim(\mathcal{X}) \geq \dim(\mathcal{Y}) = q \geq 1$, then the *principal angles* $\theta_k \in [0, \frac{\pi}{2}], 1 \leq k \leq q$ between $\mathcal{X}$ and $\mathcal{Y}$ are defined recursively by

$$\cos(\theta_k) = \max_{u \in \mathcal{X}} \max_{v \in \mathcal{Y}} \left|u^H v\right| = \left|u_k^H v_k\right| \tag{3.3}$$

subject to $||u||_2 = ||v||_2 = 1$, $u^H u_i = 0$ and $v^H v_i = 0$ for $i = 1, 2, \ldots, k-1$.

Henceforth, $\theta = (\theta_1, \ldots, \theta_q)$ will denote the principal angle vector while $\Theta = \text{diag}(\theta_1, \ldots, \theta_q)$ will denote the diagonal matrix with entries from the principal angle vector. A numerically stable algorithm that computes the canonical correlations (cosine of these principal angles) between subspaces $\mathcal{X}$ and $\mathcal{Y}$ is given in the following theorem. Note that this algorithm is proved to be mixed stable, and QR factorizations with the complete pivoting are recommended for computing $U_\mathcal{X}$ and $U_\mathcal{Y}$ [23]. Moreover, the algorithm is accurate for large principal angles ($> 10^{-8}$) [52] and requires about $4n(q^2 + 2p^2) + 2pq(n + q) + 12q^3$ flops [36]. The sine-based algorithm for calculating small principal angles is available in [52].

Theorem 3.2.1. *[9] Let $\mathcal{X} \in G(p, n)$ and $\mathcal{Y} \in G(q, n)$, $p \geq q$. Assume that the columns of matrices $U_\mathcal{X}$ and $U_\mathcal{Y}$ form unitary bases for the two subspaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let the SVD of the $p \times q$ matrix $U_\mathcal{X}^H U_\mathcal{Y}$ be*

$$U_\mathcal{X}^H U_\mathcal{Y} = UCV^H, \quad C = \text{diag}(c_1, c_2, \ldots, c_q), \tag{3.4}$$

where $U^H U = V^H V = VV^H = I_q$. If we assume that $c_1 \geq c_2 \geq \ldots \geq c_q$, then the principal angles $\theta_1, \ldots, \theta_q$ associated with $\mathcal{X}$ and $\mathcal{Y}$ satisfy

$$\cos\theta_k = c_k, \quad k = 1, \ldots, q.$$

Notice that in the Equation (3.4), for any complex matrix $U_{\mathcal{X}}^H U_{\mathcal{Y}}$ there always exists such a decomposition with positive singular values [36].

In many cases, we would like to represent subspaces using their unique orthogonal projections to be rid of ambiguity given by their non-unique basis representations. The following lemmas and theorems by [81] and [79] provide tools to explore the connections between principal angles of subspaces and their corresponding orthogonal projections. In the following discussions, we will assume the notations given above and the usual notations in matrix analysis to let $\sigma(A)$ denote the set of singular values of a matrix A, $\sigma_+(A)$ the set of non-zero singular values of A, $A^\dagger$ the Moore-Penrose generalized inverse of A, and $P_A = AA^\dagger$ the orthogonal projection onto the column space of A (or $\mathcal{R}(A)$).

Definition 3.2.5. Let $A \in \mathbb{R}^{m\times n}$. The range of A, denoted $\mathcal{R}(A)$, is the set of all vectors $\mathbf{v}$ such that $\mathbf{v} = A\mathbf{x}$, i.e.,

$$\mathcal{R}(A) = \{\mathbf{v} \in \mathbb{R}^m \mid \mathbf{v} = A\mathbf{x} \quad \text{for some} \quad \mathbf{x} \in \mathbb{R}^n\}.$$

Lemma 3.2.1. *[81] Suppose that* $\sigma(U_{\mathcal{X}}^H U_{\mathcal{Y}}) = \{c_k\}_{k=1}^q$, $c_k = \cos\theta_k$, $\frac{\pi}{2} \geq \theta_1 \geq \ldots \geq \theta_q \geq 0$. *If* $(U_{\mathcal{X}}, W_{\mathcal{X}})$ *forms an* $n \times n$ *unitary matrix and* $\sigma(W_{\mathcal{X}}^H U_{\mathcal{Y}}) = \{s_k\}_{k=1}^q$, $s_1 \geq \ldots \geq s_q$, *then*

$$s_k = \sin\theta_k, \quad k = 1, \ldots, q.$$

Proof. See Appendix A.2. □

Lemma 3.2.2. *[81] Assume the notations above for* $\mathcal{X}$, $\mathcal{Y}$, $U_{\mathcal{X}}$, $U_{\mathcal{Y}}$, *and* $W_{\mathcal{X}}$, *we have*

$$\sigma_+(U_{\mathcal{X}}^H U_{\mathcal{Y}}) = \sigma_+\left(P_{\mathcal{X}} P_{\mathcal{Y}}\right) \tag{3.5}$$

and

$$\sigma_+(W_{\mathcal{X}}^H U_{\mathcal{Y}}) = \sigma_+\left(\left(I - P_{\mathcal{X}}\right) P_{\mathcal{Y}}\right). \tag{3.6}$$

Proof. See Appendix A.3. □

Clearly, the lemmas above cast the problem of finding principal angles between subspaces in terms of singular values of projection matrices. This generalization is useful in showing the gap function $\rho_{g,2}$ is in fact a metric. We will review this fact that is based on a characterization of $\rho_{g,2}(\mathcal{X}, \mathcal{Y})$ in terms of the principal angles between $\mathcal{X}$ and $\mathcal{Y}$.

Theorem 3.2.2. *[79] Let $\mathcal{X}, \mathcal{Y} \in G(m,n)$, and let $\Theta = \text{diag}(\theta_1, \ldots, \theta_m)$, where $\theta_1 \leq \ldots \leq \theta_m$ are the principal angles between $\mathcal{X}$ and $\mathcal{Y}$. Then*

$$\rho_{g,2}(\mathcal{X}, \mathcal{Y}) = \sin\theta_m = ||\sin\Theta||_2 \tag{3.7}$$

One can further show with CS decomposition that the singular values of $P_{\mathcal{X}} - P_{\mathcal{Y}}$ are exactly sines of the principal angles between $\mathcal{X}$ and $\mathcal{Y}$. Thus, the following Corollary is an immediate consequence of Theorem 3.2.2.

Corollary 3.2.3. *[79] In the 2-norm,*

$$\rho_{g,2}(\mathcal{X}, \mathcal{Y}) = ||P_{\mathcal{X}} - P_{\mathcal{Y}}||_2. \tag{3.8}$$

Triangle inequality for $\rho_{g,2}$ follows immediately from Corollary 3.2.3 and consequently it is easy to see that $\rho_{g,2}$ indeed defines a metric on $G(m,n)$. To sum up, $\rho_{g,2}$ being a metric on $G(m,n)$ induces a topology on $G(m,n)$. Moreover, because all gap functions are equivalent (proof available in [79]), all gap functions induce the same topology, which we will call the *gap topology*. It is not hard to see that Equation (3.8) does not hold in general. Namely, if we replace the norm function in the definition of the gap function by some arbitrary ν, then the ν-gap between two subspaces is not necessarily ν of their orthogonal projectors. However, the right hand side of Equation (3.8) remains a metric on $G(m,n)$. Thus, the following theorem:

Theorem 3.2.4. *[79] Let ν be a matrix norm on $\mathbb{C}^{n\times n}$. Then the function*

$$\rho_{p,\nu}(\mathcal{X}, \mathcal{Y}) := \nu(P_{\mathcal{X}} - P_{\mathcal{Y}}) \tag{3.9}$$

is a metric on $G(m,n)$, which generates the gap topology.

It is also natural to ask if we can find new unitarily invariant metrics of the form $||\sin\Theta(\mathcal{X}, \mathcal{Y})||$, where $||\cdot||$ is an unitarily invariant norm. Unfortunately, these metrics are just the $\rho_{p,\nu}$ metrics in disguise.

Theorem 3.2.5. *[79] Let ν be a unitarily invariant matrix norm on $\mathbb{C}^{m\times m}$. Then there is a unitarily invariant matrix norm ν' such that*

$$\nu[\sin\Theta(\mathcal{X}, \mathcal{Y})] = \rho_{p,\nu'}(\mathcal{X}, \mathcal{Y}) \tag{3.10}$$

for all $\mathcal{X}$, $\mathcal{Y} \in G(m,n)$.

As it seems like we have found a nice way to quantify distance between subspaces, one may start to ponder over the question whether or not there are other metrics that generate the gap topology as well. A natural place to look for such a metric is in the set of unitarily invariant norm.

Definition 3.2.6. A norm $||\cdot||$ on the set of $m \times n$ matrices, $\mathcal{M}_{m,n}$, is unitarily invariant if for any $A \in \mathcal{M}_{m,n}$, $U \in O_m$, and $V \in O_n$, $||UAV|| = ||A||$.

A nice property of unitary spaces is that angles and distances between subspaces are preserved under rotations. Thus, an unitarily invariant norm is not sensitive to the choice of the bases used to represent a subspace. Consequently, it is used in a wide variety of applications. We are particularly interested in unitarily invariant metrics that generate the gap topology. For example, $\rho_{g,2}$ is one of the few unitarily invariant norms that generate the gap topology. However, this metric exhausts the class of unitarily invariant metrics that can be generated by gap functions, since up to a constant multiple the 2-norm is the only unitarily invariant *vector* norm on $\mathbb{C}^n$. Fortunately, there are many unitarily invariant *matrix* norms (e.g., spectral and Frobenius norm) that can be used in Equation (3.9) to give unitarily invariant metrics on $G(m, n)$ generating the gap topology. It is clear from the following theorem:

Theorem 3.2.6. *[79] If ν is a unitarily invariant matrix norm, then $\rho_{p,\nu}$ defined in (3.9) is a unitarily invariant metric on $G(m, n)$.*

Theorem 3.2.6 gives a nice way to generate unitarily invariant metrics on $G(m, n)$. Moreover, because of the connections between singular values and the principal angles of a pair of subspaces, all of the norms arise in such a way can be expressed in terms of principal angles. It is stated and proved in [67] that any symmetric gauge function of the principal angles is a metric, where a *symmetric gauge function* $\Phi : \mathbb{R}^r \to \mathbb{R}$ is a norm function that is symmetric and absolute. i.e., for any $x \in \mathbb{R}^r$ and any permutation matrix P,

$$\begin{aligned} \Phi(Px) &= \Phi(x) \quad \text{(symmetric)} \\ \Phi(|x|) &= \Phi(x) \quad \text{(absolute).} \end{aligned}$$

A particularly useful class of symmetric gauge functions, called Ky Fan k-function [26] is defined as the following,

$$\Phi_k(x_1, x_2, \ldots, x_r) = \max_{1 \le i_1 < \ldots < i_k \le r} \{|x_{i_1}| + |x_{i_2}| + \ldots |x_{i_k}|\}.$$

For example, if σ_i's are the singular values of a matrix $A \in \mathcal{M}_{m,n}$ listed in descending order, then Ky Fan k function simplifies to the sum of the k largest singular values, i.e.,

$$\Phi_k(\sigma_1, \sigma_2, \ldots, \sigma_r) = \sum_{i=1}^{k} \sigma_i.$$

Equipped with both the notions of symmetric gauge functions and Ky Fan functions, we are able to explicitly construct unitarily invariant norms.

Theorem 3.2.7. *[67] Let $\Phi : \mathbb{R}^m \to \mathbb{R}$ be a symmetric gauge function. Define $\rho : G(m,n) \times G(m,n) \to \mathbb{R}$ by*

$$\rho(\mathcal{X}, \mathcal{Y}) = \Phi(\theta(\mathcal{X}, \mathcal{Y})), \tag{3.11}$$

where $\theta(\mathcal{X}, \mathcal{Y})$ denotes the principal angle vector between $\mathcal{X}$ and $\mathcal{Y}$. Then ρ is an unitarily invariant metric and is called the angular metric.

In particular, if Φ is the Ky Fan k function, then Theorem 3.2.7 implies that sum of the k largest principal angles between two subspaces is an unitarily invariant metric. More generally, a theorem by Von Neumann states that any unitarily invariant norm of a matrix $A \in \mathcal{M}_{m,n}$ comes from some symmetric gauge function of the singular values of A.

Theorem 3.2.8. *[62] There is a one-to-one correspondence between unitarily invariant norms $||\cdot||$ on $\mathcal{M}_{m,n}$ and symmetric gauge function $\Phi : \mathbb{R}^r \to \mathbb{R}$, where $r = \min\{m,n\}$, given by*

$$||A||_\Phi = \Phi(\sigma_1, \ldots, \sigma_r).$$

As Theorem 3.2.7 provides an useful tool to generate unitarily invariant norms in terms of principal angles between subspaces, we will focus on the ones that arise naturally from realizing the Grassmann manifold as subsets of various spaces.

The (differential) topology on $G(k,n)$ can be described in several ways [13]: First, as a quotient (homogeneous space) of the orthogonal group,

$$G(k,n) = O_n / \left(O_k \times O_{n-k}\right). \tag{3.12}$$

The standard invariant Riemannian metric on orthogonal matrices O_n descends via (3.12) to a Riemannian metric on the homogeneous space $G(k,n)$. The resulting **geodesic** distance function d_g (also known as **arc length**) on the Grassmannian in terms of principal angles $\theta_1, \ldots, \theta_k$ between $\mathcal{X}$, $\mathcal{Y} \in G(k,n)$, is [91]

$$d_g(\mathcal{X}, \mathcal{Y}) = \left(\sum_{i=1}^{k} \theta_i^2 \right)^{1/2} = ||\theta||_2. \tag{3.13}$$

Next, $G(k,n)$ can be realized as a submanifold of projective space,

$$G(k,n) \subset \mathbb{P}(\Lambda^k \mathbb{R}^n) = \mathbb{P}^{\binom{n}{k}-1}(\mathbb{R}) \tag{3.14}$$

via the Plücker embedding. Then the Grassmannian inherits a Riemannian metric from the **Fubini-Study** metric on projective space [38], and the resulting *Fubini-Study* distance d_{FS} is given in terms of the principal angles by

$$d_{FS}(\mathcal{X}, \mathcal{Y}) = \cos^{-1} \left(\prod_{i=1}^{k} \cos \theta_i \right). \tag{3.15}$$

Finally, as a submanifold of Euclidean space,

$$G(k,n) \subset \mathbb{R}^{(n^2+n-2)/2} \tag{3.16}$$

via a projection embedding described recently in [20]. In this case, one can restrict the usual Euclidean distance function on $\mathbb{R}^{(n^2+n-2)/2}$ to the Grassmannian via (3.16) to obtain the **projection F** or **chordal** distance d_c (so called because the image of the Grassmannian under (3.16) lies in a sphere, so that the restricted distance is simply the distance along a straight-line chord connecting one point of that sphere to another; see [20]) which, in terms of the principal angles, has the expression

$$d_c(\mathcal{X}, \mathcal{Y}) = \left(\sum_{i=1}^{q} (\sin \theta_i)^2 \right)^{1/2} = \| \sin \theta \|_2.$$

This projection F distance d_c has recently been used in the context of sphere-packing/coding theory in the Grassmannian, where it has been reported to be significantly more efficient than the "standard" geodesic distance d_g [20], [3].

Let X and Y be two unitary matrices that span the range of the subspaces $\mathcal{X}$ and $\mathcal{Y} \in G(k,n)$, respectively, and let O_k denote the set of $k \times k$ unitary matrices. The **chordal 2-norm** and **chordal Frobenius-norm** are derived by embedding the Grassmann manifold in the vector space $\mathbb{R}^{kn}$, then using the operator 2-norm and Frobenius

norm, respectively [3]. In the context of linear algebra, the chordal Frobenius-norm is given by the minimization problem with Frobenius norm [24]

$$d_{cF}(\mathcal{X}, \mathcal{Y}) := \min_{U,V \in O_k} ||XU - YV||_F.$$

It can be shown that

$$d_{cF}(\mathcal{X}, \mathcal{Y}) = ||2 \sin \frac{1}{2}\theta||_2. \tag{3.17}$$

To prove this equality, one can first apply the CS-decomposition to the matrices X and Y to put them into standard form, then consider the Frobenius norm of the minimization problem [79]. See a detailed proof in Appendix A.4.

Similarly, the chordal 2-norm is also given by the same minimization problem but with the matrix 2-norm [24]

$$d_{c2}(\mathcal{X}, \mathcal{Y}) := \min_{U,V \in O_k} ||XU - YV||_2.$$

It can be shown that

$$d_{c2}(\mathcal{X}, \mathcal{Y}) = ||2 \sin \frac{1}{2}\theta||_F. \tag{3.18}$$

To prove this equality, one can first apply the CS-decomposition to the matrices X and Y to put them into standard form, then consider the spectral norm of the minimization problem [79]. See a detailed proof in Appendix A.5.

In terms of a generalization from the symmetric gauge functions, chordal 2-norm and chordal Frobenius-norm are special cases of the general *gap metric* [67]

$$\Phi(\sin \theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \sin \theta_k(\mathcal{X}, \mathcal{Y})) = \inf \left\{ ||U_{\mathcal{X}} - U_{\mathcal{Y}} Q|| : Q \in \mathbb{R}^k \right\},$$

where $U_{\mathcal{X}}$ and $U_{\mathcal{Y}}$ are orthonormal basis matrices for $\mathcal{X}$ and $\mathcal{Y}$, respectively, and when Φ is the Ky Fan m-function with $m = 1$ and $m = k$, respectively.

The **Projection 2-norm** [24] is defined by taking the spectral norm of the difference between projection matrices of $\mathcal{X}$ and $\mathcal{Y}$. With this definition, it is identical to the gap functions $\rho_{g,2}(\mathcal{X}, \mathcal{Y})$ and $\rho_{p,2}(\mathcal{X}, \mathcal{Y})$. It is straight-forward to see that it is equal to $\sin \theta_{\max}$. Wedin gave a geometric interpretation of the metric in [88]. This metric is also called the *subspace distance* in [36] and so widely adapted in the engineering and image

processing literature that it is now treated as the *Euclidean distance* between subspaces. Namely,

$$d_{p2}(\mathcal{X}, \mathcal{Y}) := ||P_{\mathcal{X}} - P_{\mathcal{Y}}||_2 = ||\sin\theta||_\infty. \tag{3.19}$$

In [24], some strict inequalities between these metrics are given, which gives us a clue about their equivalence relationships. For small principal angles, d_g, d_{FS}, d_{cF}, and d_c are all asymptotically equivalent and all but d_c are approaching $||\Theta||_F$ with d_c approaching $\sqrt{2}||\Theta||_F$. On the other hand, d_{p2} and d_{c2} are asymptotically approaching $||\Theta||_2$. Despite the difference in their asymptotic behaviors, all of these norms are unitarily invariant metrics that generate the gap topology. The moral of the story is that any reasonable attempt to construct a unitarily invariant metric will yield something that can be expressed in terms of principal angles. See Table 3.1 for a quick reference to the metrics discussed above.

Metric Name	Mathematical Expression
Fubini-Study	$d_{FS}(\mathcal{X}, \mathcal{Y}) = \cos^{-1}\left(\prod_{i=1}^{k} \cos\theta_i\right)$
Chordal 2-norm	$d_{c2}(\mathcal{X}, \mathcal{Y}) = \left\|2\sin\frac{1}{2}\theta\right\|_F$
Chordal F-norm	$d_{cF}(\mathcal{X}, \mathcal{Y}) = \left\|2\sin\frac{1}{2}\theta\right\|_2$
Geodesic (Arc Length)	$d_g(\mathcal{X}, \mathcal{Y}) = \|\theta\|_2$
Chordal (Projection F-norm)	$d_c(\mathcal{X}, \mathcal{Y}) = \|\sin\theta\|_2$
Projection 2-norm	$d_{p2}(\mathcal{X}, \mathcal{Y}) = \|\sin\theta\|_\infty$

Table 3.1: Table of Grassmannian distances.

3.3 Grassmann Separation Criterion

In the case of pattern recognition, most of the time the data set is compact and fixed. For example, in face recognition, it is a common practice to project face data down to a low-dimensional feature space first via KL-transform before classification. It is because the shape of the face of different people looks alike, so we only need a few vectors to represent the different features of the faces. Thus, if we form a subspace from thousand of images of a single person and form another subspace from a bunch of images of a different person, then the first few principal angles are enough to provide discriminatory information about the neighboring relationship between these two people.

Besides, well-established numerical algorithms for finding the smallest eigenvalues can be utilized to enhance efficiency and reduce cost in the computation of the principal angles. Thus, it is natural to consider nested subspaces of $\mathcal{X}$, $\mathcal{Y} \in G(k,n)$ by defining the *ℓ-truncated principal angle vector* $\theta^\ell := (\theta_1, \dots \theta_\ell)$ where $\theta_1 \leq \cdots \leq \theta_k$ are the principal angles between X and Y and $1 \leq \ell \leq k$. Note if $k > \dim(\mathcal{X} \cap \mathcal{Y}) \geq \ell$, then all of the ℓ-truncated distances between $\mathcal{X}$ and $\mathcal{Y}$ are zero, even though $\mathcal{X} \neq \mathcal{Y}$. Thus, strictly speaking, these are semi-distances at best. However, in practice, $\dim(\mathcal{X} \cap \mathcal{Y}) = 0$ whenever $\mathcal{X}$ and $\mathcal{Y}$ are distinct, so the ℓ-truncated distances are true distances on the discrete set of the experimental data. We then have ℓ-truncated semi-metrics.

Definition 3.3.1. Let ℓ-truncated principal angle vector be $\theta^\ell := (\theta_1, \dots, \theta_\ell)$ where $\theta_1 \leq \cdots \leq \theta_k$ are the principal angles between X and Y and $1 \leq \ell \leq k$, then, e.g., *ℓ-truncated Grassmannian semi-distances* between $\mathcal{X}$ and $\mathcal{Y}$ are defined as follows:

$$d_g^\ell(\mathcal{X},\mathcal{Y}) := \|\theta^\ell\|_2, \qquad d_{FS}^\ell(\mathcal{X},\mathcal{Y}) := \cos^{-1}\prod_{i=1}^{\ell}\cos\theta_i,$$
$$d_c^\ell(\mathcal{X},\mathcal{Y}) := \|\sin\theta^\ell\|_2 \qquad d_{cF}^\ell(\mathcal{X},\mathcal{Y}) := \|2\sin\tfrac{1}{2}\theta^\ell\|_2.$$

We are now equipped with all the necessary tools to analyze the separability of a data set with these Grassmannian distances and semi-distances that are based on calculations of principal angles. Before we proceed, we will introduce a separation measure that arises from the context of classification that will serve as the basis of the separation criterion on the Grassmannians.

Definition 3.3.2. The distance between different realizations of subspaces for the same class are called *match distances* while for different classes they are called *non-match distances*.

Definition 3.3.3. *False accept rate (FAR)* is the ratio of the number of false acceptances divided by the number of identification attempts. This is also referred to as a type II error in statistics.

Definition 3.3.4. *False reject rate (FRR)* is the ratio of the number of false rejections divided by the number of identification attempts. This is also referred to as a type I error in statistics.

Given match and non-match distances for a set of classes, the **false accept rate (FAR)** at a zero **false reject rate (FRR)**[2] (defined, e.g., in [56]) indicates how well a metric separates classes. This score is the ratio of the number of non-match distances that are smaller than the maximum of the match distances divided by the number of non-match distances. A zero FAR for a data set indicates that the classes are perfectly separable without ambiguity. As we establish the framework for classifying subjects in a data set, we will make use of the concept of FAR in building separability conditions.

3.4 Algorithms And Operation Counts

In this section, we will present the essential algorithms for calculating a Grassmannian distance between a pair of subspaces along with complexity analysis of the given algorithms. The corresponding MATLAB implementations will be given in Appendix B.

As mentioned in Chapter 3.2, any reasonable attempt to construct an unitarily invariant metric on subspaces will yield something that can be expressed in terms of principal angles. Thus, the fundamental building block on any Grassmannian distance will rely on the notion of principal angles. We will review two algorithms here that compute large and small principal angles between a pair of subspaces. See [9] (cosine of large principal angles) and [52] (sine of small principal angles) for detailed derivations. Algorithm 3.4.1 computes the large principal angles and cosine of the principal angles between two subspaces $\mathcal{R}(A)$ and $\mathcal{R}(B)$ based on the recursive algorithm given by Björck and Golub [9].

---

algorithm 3.4.1 [9] Large Principal Angles

Input: matrices A (n-by-p) and B (n-by-q).
Output: cosine of the principal angles between subspaces $\mathcal{R}(A)$ and $\mathcal{R}(B)$, C.

1. Find orthonormal bases Q_a and Q_b for A and B such that

$$Q_a^T Q_a = Q_b^T Q_b = I \quad \text{and} \quad \mathcal{R}(Q_a) = \mathcal{R}(A), \mathcal{R}(Q_b) = \mathcal{R}(B).$$

2. Compute the SVD of $Q_a^T Q_b$: $Q_a^T Q_b = UCV^T$, so that $\text{diag}(C) = \cos\theta$.

---

[2]For the sake of brevity we refer to the false accept rate (FAR) at a zero false reject rate (FAR) simply as FAR in the following discussions.

When using the standard double-precision arithmetic, this algorithm is only accurate up to angles greater than or equal to 10^{-8}. This problem is pointed out and treated in the classical paper [9]. In short, in order to avoid the round-off error occurred when using a SVD-based algorithm for calculating cosine of the principal angles, a sine-based algorithm for calculating the principal angles that is motivated by the notion of *gap* (see Chapter 3.2) is considered. The idea of the algorithm is based on a theorem from [9], which is essentially the same as that of Lemma 3.2.2. An algorithm for computing all the principal angles is given in Algorithm 3.4.2 and taken from [52]. Notice that this algorithm provides accurate result for small angles and the cosine-based algorithm for computing large angles is kept.

---

algorithm 3.4.2 [52] Small and Large Principal Angles

Input: matrices A (n-by-p) and B (n-by-q).
Output: principal angles θ between subspaces $\mathcal{R}(A)$ and $\mathcal{R}(B)$.

1. Find orthonormal bases Q_a and Q_b for A and B such that

$$Q_a^T Q_a = Q_b^T Q_b = I \quad \text{and} \quad \mathcal{R}(Q_a) = \mathcal{R}(A), \mathcal{R}(Q_b) = \mathcal{R}(B).$$

2. Compute SVD for cosine: $Q_a^T Q_b = Y \Sigma Z^T$, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_q)$.

3. Compute matrix $B = \begin{cases} Q_b - Q_a(Q_a^T Q_b) & \text{if rank}(Q_a) \geq \text{rank}(Q_b); \\ Q_a - Q_b(Q_b^T Q_a) & \text{otherwise.} \end{cases}$

4. Compute SVD for sine: $[Y, \text{diag}(\mu_1, \ldots, \mu_q), Z] = \text{svd}(B)$.

5. Compute the principal angles, for $k = 1, \ldots, q$:

$$\theta_k = \begin{cases} \arccos(\sigma_k) & \text{if} \quad \sigma_k^2 < \frac{1}{2}; \\ \arcsin(\mu_k) & \text{if} \quad \mu_k^2 \leq \frac{1}{2}. \end{cases}$$

The MATLAB qr and orth commands use the LINPACK routine zqrdc, which is based on householder reflections. For a general m-by-n matrix, QR-decomposition using houseboulder reflections costs $2n^2m - \frac{2}{3}n^3$ flops. And the MATLAB svd command uses the LINPACK routine zsvdc. The zsvdc routine calculates the singular values of an m-by-n complex matrix X in two steps. First, it reduces X to a bidiagonal matrix B by finding orthogonal matrices U_1 and V_1 such that $A = U_1 B V_1^T$ based on householder reflections. In general, if X is n-by-n, the cost of bidiagonal reduction is $\frac{8}{3}n^3 + O(n^2)$ flops. Then QR-iteration with deflation and shifting is applied to the covariance matrix

$C = B^T B$ to find its eigenvalues, which then give the squared singular values of B, thus X. Since C is symmetric and tridiagonal, the total costs to find just the eigenvalues of C is less than $\frac{4}{3}n^3 + O(n^2)$ flops. Therefore, it costs less than $4n^3 + O(n^2)$ to get the singular values of X. In general, for a m-by-n matrix, it costs $4n^2(m - \frac{1}{3}n)$ flops to reduce it to a bidiagonal form using Householder reflections and if only singular values are required, it costs just $O(n^2)$ for the rest of the operations. Therefore, a MATLAB svd routine costs $4n^2(m - \frac{1}{3}n) + O(n^2)$ flops to compute the singular values of a m-by-n matrix. Overall, Algorithm 3.4.1 makes two calls to economy version of qr and one call to economy version of svd for which it costs $\frac{2}{3}p^2n + \frac{2}{3}q^2n + 4p^3$ flops. When $p = q = k$, the cost becomes $\frac{4}{3}k^2n + 4k^3$. Now, calculating the chordal distance between a pair of subspaces based on principal angles is straightforward and given in Algorithm 3.4.3.

---

algorithm 3.4.3 Computation of Geodesic (Arc Length) Between Two Subspaces

---

Input: $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$.
Output: chordal distance, d_c, between $\mathcal{X} = \mathbb{R}(X)$ and $\mathcal{Y} = \mathbb{R}(Y)$.

1. Calculate cosine of the principal angles, C_i, $i = 1, 2, \ldots, q$, between $\mathcal{X}$ and $\mathcal{Y}$ using either Algorithm 3.4.1 or 3.4.2.
2. Calculate $d_c(\mathcal{X}, \mathcal{Y}) = ||\sin\theta||_F = \sqrt{\sum_{i=1}^{q}(1 - C_i^2)}$.

In terms of speed of the algorithms, we perform a single chordal distance calculation between two subspaces with data from the CMU-PIE database [74]. The images are of size 160-by-138 for which eye coordinates are registered and geometric normalization is performed. A single pair of chordal distance calculation on a probe point of cardinality (see Definition 4.1.8) ten and a gallery point of cardinality ten with aforementioned MATLAB routines takes 4.266 seconds on a Pentium M, 1.60 GHz processor while it takes 0.05 seconds to run a single pair of chordal distance calculation on a probe point of cardinality one and a gallery point of cardinality ten.

Chapter 4

CLASSIFICATION ON THE GRASSMANNIANS

Grassmann framework can be implemented in a variety of applications. In particular, if a data set is linear and can be captured with low-dimensional linear subspaces, then it is natural to transform the classification problem to one on the Grassmannians. Linearity is certainly advantageous, but should not be considered as a necessity. Even when a data set is not intrinsically linear, often times we are able to find a linear subspace of relatively low dimension that encapsulates the data set. Classification on the Grassmannians can be extended to any data set that has multiple examples per subject. For example, in the two class problem of gender, where we are interested in determining whether a set of images is drawn from the male or female population, we can generate a set of points on the Grassmann manifold associated with the male population and another set of points associated with the female population for the gallery subspaces. When a new set of images comes in, we can tell if it comes from a female or male population by comparing its Grassmannian distance with the gallery subspaces. As illustrated in Figure 4.1, we encode each set of images as a point on $G(k, n)$ by considering their span, where k is the number of distinct images associated to a single subject and n is the pixel resolution of the images.

Under this framework, improved classification outcomes are often observed since families of patterns with a common characterization often possesses discriminatory variations that are useful for classification. By collecting multiple images per subject, the state of this intrinsic variation is captured by a subspace, therefore less sensitive to other unwanted variations, such as noise. We will introduce in Chapter 4.1 the necessary machineries that lead to the notion of *Grassmann separability* of a data set and briefly describe a classification result using the paradigm on two two-class classification problems in Chapters 4.2 and 4.3. Readers who are interested in the details of the experiments results are referred to [14].

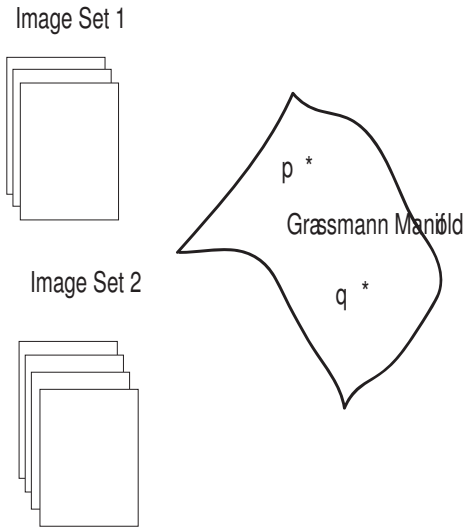


Figure 4.1: Each set of images (represented as a single stack) may be viewed as a point on the Grassmann manifold by computing the span of elements in the set.

4.1 Framework

To classify a collection of sets, we first realize them as points on $G(k, n)$, where n is the resolution of the data and k is the minimum number of data available across all identities. Then pairwise distances are calculated among the realizations of the identities. Classification can then be done based on simple comparison of these distances. See Figure 4.2 for an illustration of this classification flow. Notice that the distance between sets of distinct identities should always be larger than the distance between sets of the same subject to ensure a perfect classification.

We propose to look at the classification problem of linear subspaces with the framework for many-to-many comparisons using the metrics derived in Chapter 3.2. We will establish the notion of separability when sets of vectors are realized as points on the Grassmannian, hence the notion of *Grassmann Separability*. Namely, identities are distinct from each other if their respective subspaces are Grassmann separable.

Definition 4.1.1. Two classes are linearly separable in n-dimensional space if they can be separated by an $(n-1)$-dimensional hyperplane.

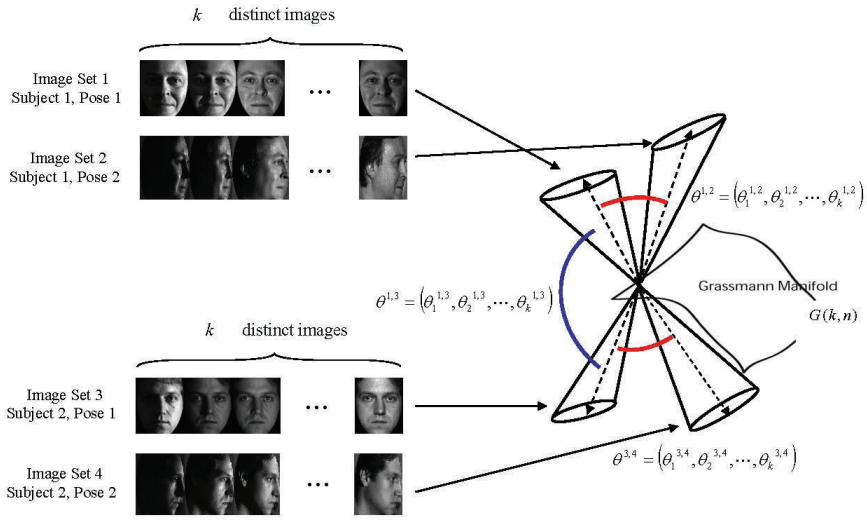


Figure 4.2: Illustration of the Grassmann method, where each set of images may be viewed as a point on the Grassmann manifold by computing an orthonormal basis associated with the set. Pairwise distances are found using the principal angles $\theta^{i,j}$'s. Distance between sets of distinct identities (largest arc) should always be larger than the distance between sets of the same subject (smaller arcs) to ensure a perfect classification.

In particular, in a one-dimensional space (such as a line), two classes are linearly separable if there exist a single point that divides the line into two rays. In general, it is not easy to find a $(n-1)$-dimensional hyperplane that linearly separates classes in n-dimensional space. Therefore, we cast the problem of separating subspaces to separation of subspaces with their associated Grassmannian distances, which form a 1-dimensional space that can be separated by a single point.

In general applications, data sets often consist of a collection of images $\{x_1, x_2, \ldots, x_m\}$ with each $x_i \in \mathbb{R}^n$ and belongs to one of the subject classes. One subject might have one image associated with it, while another subject might have multiple images associated with it. We would like to talk about the Grassmann separability of a data set using the available data for each subject without ambiguity. Thus, the following definition gives a way to discuss ways of partition on the available data for all subjects.

Definition 4.1.2. Given a collection of data $x_1, x_2, \ldots, x_N$ with each $x_i \in \mathbb{R}^n$ belonging to the same class, C, that can be grouped into partitions, $\{P_1, P_2, \ldots, P_{r+1}\}$, where for each i, $x_i \in P_j$ for some $1 \leq j \leq r+1$ and $1 \leq |P_j| \leq N$. If $|P_j| = k$ for all j but one and the span of the elements in P_j forms a k-dimensional subspace of $\mathbb{R}^n$, then we say that

the set $\{P_1, P_2, \ldots, P_r\}$ is a *dimension-k subspace configuration* of C. In the case where $P_i \cap P_j = \emptyset$ for all $1 \leq i, j \leq r$, then we say the set $\{P_1, P_2, \ldots, P_r\}$ is a *dimension-k complete subspace configuration* of C.

Example 4.1.1. Given $C = \{a, b, c, d, e, f, g, h, i, j, k\}$ where each element is in $\mathbb{R}^n$. The set $\{\{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}, \{i, j\}\}$ is a dimension-2 complete subspace configuration of C. The set $\{\{a, b, c\}, \{c, e, f\}, \{d, h, i\}, \{g, h, a\}\}$ is a dimension-3 subspace configuration of C but NOT a complete subspace configuration. Notice that we do not need to consider all of the elements in C in forming the subspaces.

Definition 4.1.3. (2-class Grassmann separability) Let $C^{(1)}$ and $C^{(2)}$ be two classes with dimension-k complete subspace configurations $C^{(1)} = \{S_1^{(1)}, S_2^{(1)}, \ldots, S_{k_1}^{(1)}\}$ and $C^{(2)} = \{S_1^{(2)}, S_2^{(2)}, \ldots, S_{k_2}^{(2)}\}$, where $S_{k_i}^{(i)}$'s are points on $G(k, n)$ and do not intersect trivially in any dimension, i.e., if $\theta_1, \theta_2, \ldots, \theta_k$ are principal angles between any pair of subspaces in $C^{(1)}$ or $C^{(2)}$, then θ_i is not equal to 0 identically for all i. We then say that $C^{(1)}$ and $C^{(2)}$ are *Grassmann d-separable in k* if there exists a ℓ-truncated Grassmannian semi-distance d and a real number $\epsilon \geq 0$ (tolerance value) such that for each subject i,

$$\max_{1 \leq m,n \leq k_i} d(S_m^{(i)}, S_n^{(i)}) \leq \epsilon \quad \text{and} \quad \min_{\substack{1 \leq m \leq k_1 \\ 1 \leq n \leq k_2}} d\left(S_m^{(1)}, S_n^{(2)}\right) > \epsilon.$$

In other words, between-class distances are always greater than within-class distances. Therefore, we define the *separation gap* $g_s = m - M$, where M is the maximum of the match distances and m is the minimum of the non-match distances, to quantify Grassmann separability, since $g_s > 0$ implies that the two classes are Grassmann separable.

In some of the cases of geometric object recognition, we are interested in the case where $k_i = 2$ for all i. Namely, split the available data in each class into two disjoint sets from which the subspaces $S_{k_i}^{(i)}$'s are formed. This is because gallery models are better learned from more images than fewer. Now, it follows naturally from the definition that if the two associated classes of Grassmannian distance is linearly separable, then the two classes $C^{(1)}$ and $C^{(2)}$ are Grassmann d-separable. To simplify the notion of Grassmann separability between classes, we then define the following.

Definition 4.1.4. (2-class Grassmann separability) We say that two classes are Grassmann separable if there exists a ℓ-truncated Grassmannian semi-distance d such that they are Grassmann d-separable in some k.

It is worth mentioning that in the extreme case when $d(S_m^{(i)}, S_n^{(i)}) = 0$, there exist two subspaces that perfectly estimate the class $C^{(i)}$. In other words, $S_m^{(i)}$ is a unitary transformation of $S_n^{(i)}$, since the metric being a nonnegative function implies that all principal angles between $S_m^{(i)}$ and $S_n^{(i)}$ are identically zero whenever their distance is 0. Ideally, in the absence of noise and given enough samples, any two subspaces estimated from vectors of the same class should be identical. Moreover, any two subspaces estimated from vectors of different classes should be non-intersecting. However, in a practical setting, vectors are noisy in their high-dimensional ambient space. As a result, two subspaces estimated from points of a single class are not only not identical; they rarely even intersect nontrivially in practice. The following example illustrates the fact that two subspaces estimated from the same class are closer in their Grassmannian distances than any other subspace estimated from a different class. We construct the sets in a way to further demonstrate the robustness of the notion of Grassmann separability. Notice that the four matrices in the following example give natural representations of four linear subspaces that are elements of $G(2, 4)$. For convenience, instead of writing formally that a class is a collection of subspaces, we will in general write a class's subspace configuration as a collection of matrices that give natural correspondences to equivalence classes in $G(2, 4)$.

Example 4.1.2. Let

$$\begin{aligned} C^{(1)} &= \left\{ S_1^{(1)} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}^T, S_2^{(1)} = \begin{bmatrix} 1.1 & 1 & 0.1 & 1 \\ 0 & 0 & 1 & 1.1 \end{bmatrix}^T \right\} \\ C^{(2)} &= \left\{ S_1^{(2)} = \begin{bmatrix} 10 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}^T, S_2^{(2)} = \begin{bmatrix} 10.01 & 1 & 0 & 0 \\ 0 & 1 & 0.01 & 1.01 \end{bmatrix}^T \right\}. \end{aligned}$$

With this notation, $C^{(1)}$ and $C^{(2)}$ are Grassmann d-separable, where d, for example, is the chordal distance. $S_2^{(1)}$ can be seen as a noisy version of $S_1^{(1)}$ and similarly for $S_2^{(2)}$ and $S_1^{(2)}$. See Table 4.1 for the pairwise chordal distances. Clearly the within-class distances are always smaller than the between-class distances. Moreover, this notion of distance is less sensitive to perturbation, i.e., noisy subspaces of an identity are not confused with subspaces of other identities.

within-class		between-class			
$d_{1,2}^{(1,1)}$	$d_{1,2}^{(2,2)}$	$d_{1,1}^{(1,2)}$	$d_{2,2}^{(1,2)}$	$d_{1,2}^{(1,2)}$	$d_{2,1}^{(1,2)}$
0.1013	0.0086	0.9613	0.9416	0.9598	0.9435

Table 4.1: $d_{i,j}^{(m,n)}$ = chordal distance between the ith and jth subspaces of m and n.

So far, we have built the notions of separability upon subspaces. We will now introduce the notion of separability on a collection of classes where elements in those classes form the basis vectors of the subspaces whose separability conditions give rise to the separability conditions of the entire collection.

Definition 4.1.5. (multi-class Grassmann separability)
A set $\mathcal{P} = \left\{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\right\}$ of N distinct classes with dimension-k subspace configurations $C^{(i)} = \left\{S_1^{(i)}, S_2^{(i)}, \ldots, S_{k_i}^{(i)}\right\}$ for each i is **Grassmann d-separable in** k if $C^{(i)}$ and $C^{(j)}$ are pairwise Grassmann d-separable in k for all $1 \leq i, j \leq N$.

Similarly, we can drop the metric and define Grassmann separability for a collection of classes and derive a useful theorem concerning the subspace configurations.

Definition 4.1.6. (multi-class Grassmann separability)
A set $\mathcal{P} = \left\{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\right\}$ of N distinct classes having subspace configurations $C^{(i)} = \left\{S_1^{(i)}, S_2^{(i)}, \ldots, S_{k_i}^{(i)}\right\}$ for each i is **Grassmann separable** if there exists a ℓ-truncated Grassmannian semi-metric d such that $\mathcal{P}$ is Grassmann d-separable in some k.

Theorem 4.1.1. *If $\mathcal{P} = \left\{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\right\}$ is Grassmann separable with a dimension-k complete subspace configuration, then it is Grassmann separable with any dimension-k subspace configuration.*

Proof. Assume that $\mathcal{P}$ having dimension-k complete subspace configurations for each distinct class is Grassmann separable, then the within-class distances are always less than the between-class distances. Now, if $\mathcal{P}$ has dimension-k subspace configuration for each distinct class, then for some class i there exists a pair of subspaces $S_{m_i}^{(i)}$ and $S_{n_i}^{(i)}$ such that $\theta_{\min}(S_{m_i}^{(i)}, S_{n_i}^{(i)}) = 0$. This makes the within-class distances smaller than or equal to the ones from before. Hence $\mathcal{P}$ with a dimension-k subspace configuration is Grassmann separable. □

Converse is not necessarily true. Thus, for simplicity, we will adapt the notion of complete subspace configuration without loss of generality in our examples and applications. In terms of the distance matrix $D = (d_{ij})$ given in Figure 4.3(a), $\mathcal{P}$ is Grassmann separable if $\max_{i=j} d_{ij} < \min_{i \neq j} d_{ij}$. Equivalently, $\mathcal{P}$ is Grassmann separable if D has zero percent FAR at zero FRR as defined earlier in Chapter 3.3. Now, Grassmann separability can be recast in this new language.

Definition 4.1.7. (**Separation Gap**) Let $\mathcal{P}$ be a set of N distinct classes with dimension-k (complete) subspace configurations for each i. Further let M be the maximum of the match distances and m be the minimum of the non-match distances for some ℓ-truncated Grassmannian semi-distance d. Then define *separation gap* to be $g_s = m - M$.

It follows immediately from Definition 4.1.7 that $g_s > 0$ implies $\mathcal{P}$ is Grassmann separable. Figure 4.3(b) is obtained using the arc length on a Grassmann separable data set with dimension-2 subspace configurations. It can be seen clearly that the diagonal entries (match distances) have much lower intensity than the off-diagonal entries (non-match distances). Figure 4.4 shows the box-whisker plot of Grassmann separable and a non-Grassmann separable data sets. For a non-zero separation gap, the amount of overlap between the maximum of the matching distances and the minimum of the non-matching distances is measured by the FAR score.

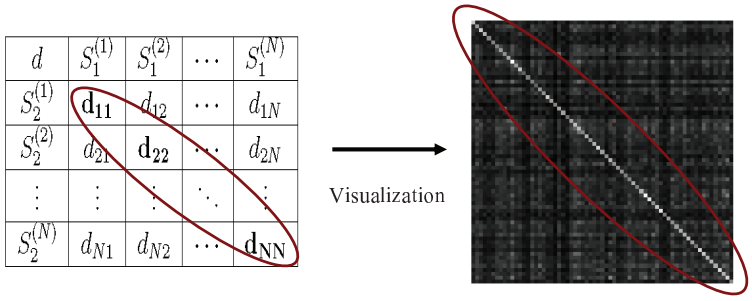
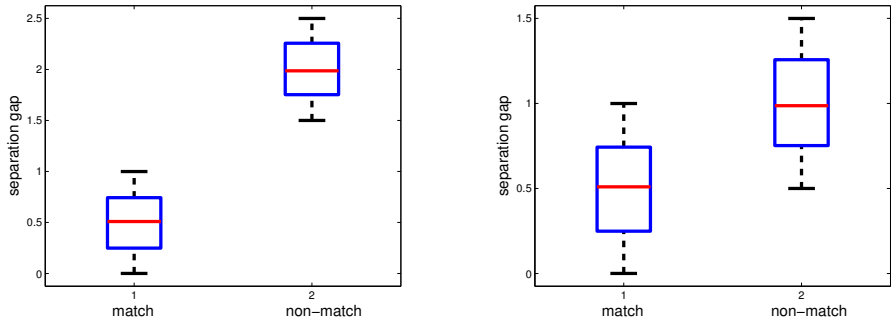


Figure 4.3: Visualization of the distance matrix for a Grassmann separable data set where distances of matching subspaces (diagonal entries) are smaller than the distances of non-matching subspaces (off-diagonal entries).

Finally, since the central idea of this thesis is about classifying *a set of images*, the common terms gallery and probe sets used in objection recognition are altered here to



(a) Grassmann separable (b) non-Grassmann separable

Figure 4.4: Illustrations of Grassmann and non-Grassmann separable data sets.

describe *sets of sets of images*. Thus, we consider the gallery and probe data to consist of a set of points on a Grassmann manifold where each point is generated by computing a basis from a set of images associated with a given class. For simplicity, let *cardinality* of a point on the Grassmann manifold be the number of images used to construct such a point on the manifold.

Definition 4.1.8. (Cardinality) Let *cardinality* of a point on the Grassmann manifold be the number of distinct images used to construct such a point on the manifold. In matrix notation, denote *cardinality* of a data matrix by the rank of that matrix.

4.2 Classification of Gender

We now illustrate the concept of Grassmann separability with two examples starting with the classification problem of gender. It has been established that gender classification of digital images of faces is a tractable problem [19]. Here we explore a slightly different question: Is a set of images drawn from the male or female population? To this end we use frontal images with neutral expressions and a single illumination setting taken from the CMU-PIE database [74]. The complete image as supplied by CMU is used, and no cropping or re-sampling is performed. There are 50 men and 17 women in the CMU-PIE database. Therefore, 50 and 17 images are available for training and testing the male class C^+ and female class C^-, respectively. The images of men are partitioned into 40 training and 10 testing images. The images of women are partitioned into 10 training and 7 testing images. For men, training points on the Grassmann manifold are generated by randomly sampling 8 out of 40 images. Thus, there are $76,904,685$

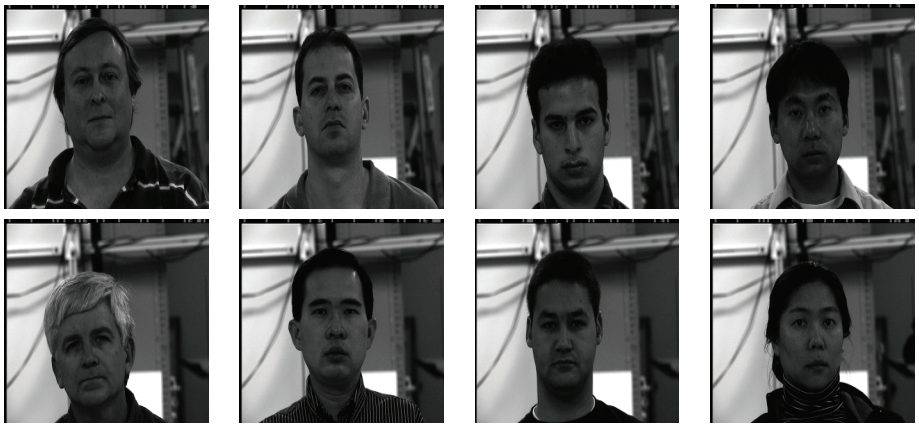


Figure 4.5: Example images used to construct the training and testing points for the male and female classes.

ways to construct labeled points associated with C^+. Testing points are generated by randomly sampling 5 out of 10 images: there are 252 possible points associated with C^-. For women, training points are generated by randomly sampling 8 out of 10 images and test points are generated by randomly sampling 5 out of 7 images. Clearly, the particulars of sampling could vary. Example images used to construct points on the Grassmann manifold are shown in Figure 4.5.

As described above, many-to-many set comparisons were carried out for the male-female classification problem where male and female classes are represented by sets of images. We present results for a training set consisting of 3 labeled points associated with each class. The resulting misclassification rates are presented in Table 4.2 where we vary the cardinality of the training points for the male and female classes. We found that the number of labeled points belonging to each class used on the Grassmann manifold impacts the classification outcome. Therefore ten trials were carried out to give a broader idea of how this set-to-set comparison behaves. A total of 40 testing points were used, where 20 are of male class and 20 are of female class. Admittedly the cardinality of the testing and training points are *ad hoc* and additional exploration is warranted. This example is illustrative of a non-Grassmann separable data set when using the 1-truncated semi-distance and not intended to be a detailed study of the gender recognition problem. The data set might become Grassmann separable if different parameters are used, such as the value of ℓ in the (semi-)distance measures.

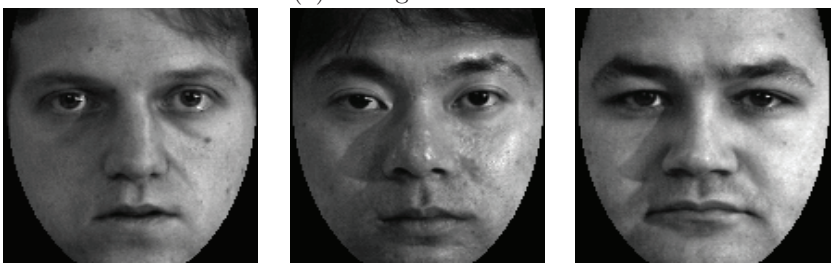
	Trial number										
ℓ	1	2	3	4	5	6	7	8	9	10	$\mu \pm \sigma$
1	0	0	0	0	0	0	22.5	0	10	0	3.25±7.46
2	35	0	0	5	12.5	27.5	22.5	5	10	7.5	12.5±11.96
3	30	0	0	0	0	27.5	27.5	5	22.5	0	11.25±13.66

Table 4.2: Error rates out of 40 testing points where 20 are of male class and 20 are of female class. ℓ-truncated semi-chordal metric is used. The experiment is repeated ten times where the mean and standard deviation is reported in the last column of the table.

4.3 Classification of Glasses

The data set used here is a subset of the "lights" portion of the CMU-PIE database, where images were captured in neutral expression under a single illumination condition with ambient lights on. Among the 67 subjects in this portion of the CMU-PIE database, 39 subjects were seen without glasses and 28 subjects were seen with glasses. Geometric normalization is performed with this data set. A total of 39 images are available for training and testing for the without-glasses class while 28 images were available for training and testing for the with-glasses class. For the with-glasses class, we divide the 28 images into mutually disjoint sets of 20 and 8 for training and testing, respectively. We further construct training points by randomly selecting 10 images in the list of 20 so that there are $184,756$ ways to construct a point for the training set. Similarly, we construct testing points by randomly choosing 5 images in the list of 8 so that there are 56 ways to construct this testing point. For the without-glasses class, we divide the 39 images into mutually disjoint sets of 30 and 9 for training and testing, respectively. We then have $30,045,015$ and 126 possibilities for the training points and testing points, respectively. See Figure 4.6 for the example images from the data set that are used to construct the training and testing points for the with-glasses and without-glasses classes. We observed that the average classification errors are less than two percent for training points of cardinality 3 in each class and it is zero when training points of cardinality 10 are used for each class. See Table 4.3 for more details of the classification outcome. Thus, when the 1-truncated semi-distance is used, this data set is Grassmann separable.

(a) with-glasses class.



(b) without-glasses class.

Figure 4.6: (a) Example images used to construct training and testing points for the with-glasses class. (b) Example images used to construct training and testing points for the without-glasses class.

	Trial number										
ℓ	1	2	3	4	5	6	7	8	9	10	$\mu \pm \sigma$
1	0	0	0	0	0	0	0	0	0	0	0
2	50	37.5	12.5	45	35	47.5	27.5	37.5	35	40	36.8±10.8
3	52.5	55	52.5	62.5	47.5	47.5	47.5	52.5	70	45	53.3±7.7

Table 4.3: Error rates out of 40 testing points where 20 are of with-glasses class and 20 are of without-glasses class. ℓ-truncated semi-chordal metric is used. The experiment is repeated ten times with the mean and standard deviation shown in the last column.

Chapter 5

FACE RECOGNITION UNDER VARYING ILLUMINATION

A popular multi-class problem that fits into the Grassmann framework is the face recognition problem under varying illumination conditions. This is due to the fact that the set of images of a general object illuminated by an arbitrary number of distinct point light sources forms a convex polyhedral cone that lies near a low-dimensional linear subspace [7, 4]. We will first review the theoretical and empirical evidence that lead the these facts and present a few state-of-the-art techniques in solving the illumination problem in Chapter 5.1. This Chapter is concluded with two successful classification results on CMU-PIE and YDB with the Grassmann method in Chapters 5.2 and 5.3, respectively. Readers who are interested in the experimental details and results are referred to [14] and [13].

5.1 Background

Belhumeur and Kriegman have shown that the set of n-pixel monochrome images of a convex, Lambertian object illuminated by an arbitrary number of distinct point light sources forms a convex polyhedral cone in $\mathbb{R}^n$, which they refer to as *the illumination cone* [7]. They also established that the statement remains true for non-convex objects under relaxed lighting conditions. Unfortunately, for most objects the exact illumination cone is difficult to obtain, especially for non-convex human faces. However, experimental work by Belhumeur et al. [34], and Kriegman et al. [54] show that this cone lies near a low-dimensional linear subspace in the space of all possible images.

Further, Basri and Jacobs have demonstrated both theoretically and empirically that the set of images of a convex, Lambertian object seen under arbitrary distant light

sources can be well approximated by a 9-dimensional linear subspace [4]. In this context, "well approximated" means that over 99% of the statistical variance is captured in the subspace. Basri and Jabobs prove their result by representing reflectance functions as linear combinations of spherical harmonics and further relating the representation of images to the reflectance functions. Their theory shows that the first 9 harmonics capture over 99% of the energy and that the high frequency components of the lighting function have little effect on the reflectance function. Therefore, one can confidently approximate the representation of the high-dimensional illumination face images with as few as 9 harmonics.

In addition, Ramamoorthi transforms the problem of linear approximation with spherical harmonics into linear approximation with principal components which can be shown to be identical to the spherical harmonic basis functions evaluated at the surface normal vectors under the same assumptions of Basri and Jacobs [70]. This prior work is ample evidence for the utility of modeling illumination variation with low-dimensional linear subspaces. Ramamoorthi provides an analytic construction of the principal components which he uses to address the reason behind the variation in combinations of the principal components when the distribution of the surface normals is given by a face. The total energies captured by the first 9 principal components for images of a face as well as images of a sphere under varying illumination condition are both over 99%.

In terms of building a classifier, Belhumeur et al. [34] developed a generative procedure that ultimately gives rise to a representation based on pose-specific illumination cones for each face class, where single-to-many comparison is performed at the classification stage. In short, a single face representation consists of multiple pose-specific illumination cones that are approximated linearly to capture over 99% of the variability, i.e., $\mathcal{R}_f = \bigcup_{p=1}^{\text{many}} I_p$, where I_p is a linear approximation of the sub-sampled illumination cone, C_p, for each pose p. This face representation is then projected down to $\mathcal{D}_f$ to further reduce the dimension. Recognition of a test image, x, is performed by first normalizing it to unit length and then computing its distance to $\mathcal{D}_f$, for each f. Namely, x is assigned to the identity of the closest face representation based on Euclidean distance within the image space, i.e., x is assigned to the face identity f^* for which f^* satisfies

$$(f^*, p^*) = \arg\min_{f,p} \sqrt{||x - \mathcal{D}_f||^2 + \min_{I_p \in \mathcal{R}_f} ||x - I_p||^2}.$$

This way, the pose and identity information of the test image x can be revealed.

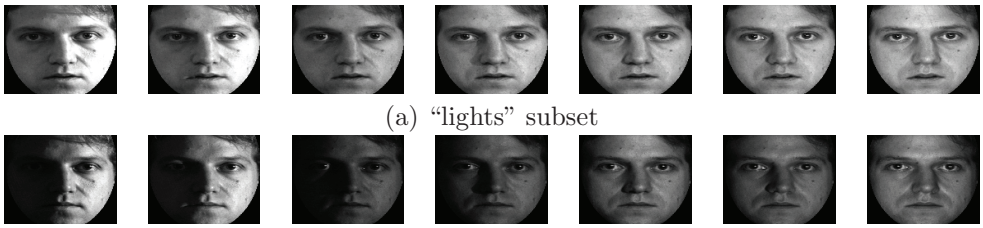
A similar but slightly different approach by Gross et al. is given in [39]. A set of "ideal" eigen light-fields can be learned from a collection of training images from which the identities are not necessarily found in the testing images. Images in the gallery and probe are then written as linear combinations of the known eigen light-fields with appropriate weights. A probe identity x is then assigned to the identity in the gallery whose eigen light-field representation is the closest to that of x in the Euclidean sense. The classification is based on many-to-many image comparisons where any number of gallery and probe images captured at arbitrary poses per subject can be used. Better classification outcomes are achieved as the number of images used in both gallery and probe increases.

As powerful as both of these methods appear to be in terms of dealing with both pose and illumination variations [7] or just pose variations [39], one major drawback lies in the notion of training. For both of these classification schemes to work, a significant amount of training has to be done *a priori*, thus making it harder to apply in real-time face recognition. We will present in this section successful examples using the Grassmann method without requiring any form of training. Namely, we will show that a subset of the CMU-PIE data set and the Yale Face Database B are both Grassmann separable as described in Chapter 4.1.

Two face data sets, Yale Face Database B (YDB) [34] and CMU-PIE [74], are frequently used by researchers to explore how changing illumination alters the appearance of human subjects; they are the largest publicly available data sets of face imagery to offer many (64 and 21, respectively) illumination of each subject (10 and 67, respectively). We will focus our attention on a subset of these two data sets where images are acquired with frontal pose and neutral facial expression, so named the illumination data sets. As mentioned in the previous section, such an illumination data set is lose to lying on a low-dimensional linear subspace. Therefore, classification on such data sets satisfies the framework for classification on the Grassmannians.

5.2 The Grassmann Separability of CMU-PIE

The PIE database includes imagery of 68 people under different pose, illumination conditions, and expressions. Our focus here concerns illumination variations rather than



(a) "lights" subset

(b) "illum" subset

Figure 5.1: (a) 7 example images of the "lights" subset of CMU-PIE database where ambient lights are on. (b) 7 examples images of the "illum" subset of CMU-PIE database where ambient lights are off.

pose, so only frontal (c27) images are used. For these frontal images, there are 21 distinct sources of lights used to illuminate the face. Preprocessing of the images include geometric normalization based on known eye-coordinates and clipping prior to the classification. In addition, these 21 sources are sampled both with the background room lights on (see Figure 5.1(a)) and the background room lights off (see Figure 5.1(b)).

For each of the two types of imagery, room lights on and room lights off, we have randomly selected two disjoint sets of images X_i and X_j for $1 \leq i, j \leq 67$ people in the PIE Database [1]. This sampling is "balanced" in so much as it is randomized relative to the specific illumination settings. The use of only 10 images to estimate the illumination space is probably approaching a lower bound on the necessary number of samples. To augment the sample, the mirror reflection of each image is also included in the image set when estimating illumination subspaces. See Figure 5.2 for an example mirror image created by reflecting through the midline. Augmentation of the data set via inclusion of the mirror images effectively increases the available data [51]. Furthermore, this *symmetrization* of the data set imposes even and odd symmetry on the basis functions analogous to sinusoidal expansions. For sets of facial images under varying illumination conditions, reflection augmentation drastically improves the estimated linear representation by both increasing the effective sample set size and introducing novel illumination conditions. As a consequence, the approximation of illumination spaces can be improved without acquiring more data.

[1] One subject was not used because 3 images of subject 39 were missing due to hardware problems during the process of image acquisition.

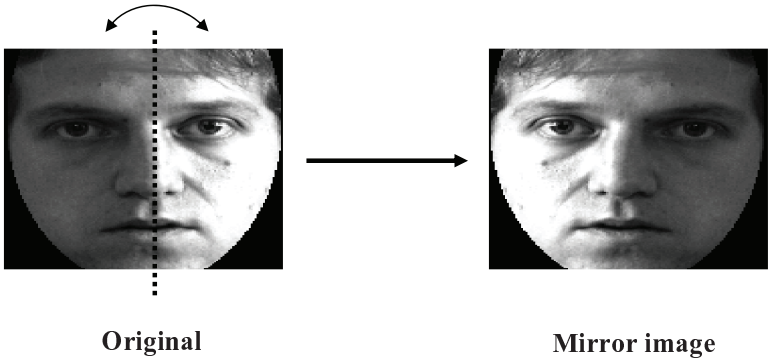


Figure 5.2: A mirror image is created by reflecting column pixels with respect to the vertical midline of the face.

Figure 5.3 summarizes the results for the room lights off and room lights on. Each plot shows the FAR at zero FRR using four ℓ-truncated Grassmannian semi-distance measures, for $1 \leq \ell \leq 20$. The values shown are averages taken over ten trials, therefore creating 670 match pairs and $10(67^2 - 67)$ non-match pairs. In each trial, random pairs of disjoint sets are created for each of the 67 people in the PIE database. Notice that, for instance, the chordal semi-distance perfectly separates all subjects in the PIE data set for all values of ℓ in both cases of lighting conditions. Therefore, this subset of the CMU-PIE data set is Grassmann separable. See [13] for different sampling and lighting conditions where Grassmann separability is observed.

5.3 The Grassmann Separability of YDB

The YDB [34] has far fewer subjects than PIE. Nonetheless, it is worth studying because it is the oldest and most studied illumination database, and because it has a large number of images per subject. See Figure 5.4 for example variations of illumination of a single subject. For each of the ten subjects, two disjoint sets have been created by randomly sampling from the 64 images per person. Then, these sets have been compared using four (ℓ-truncated) Grassmannian semi-distance measures applied to the estimated illumination subspaces. The results when image sets contain only 8 and 16 images are summarized in Figure 5.5. Notice that none of the Grassmannian semi-distances perfectly separate all subjects in YDB when only using 8 images plus their mirrors to estimate subspaces. However, the chordal semi-distance perfectly separates all subjects in YDB for all values of ℓ when using 32-dimensional (16 plus mirrors) subspace representations.

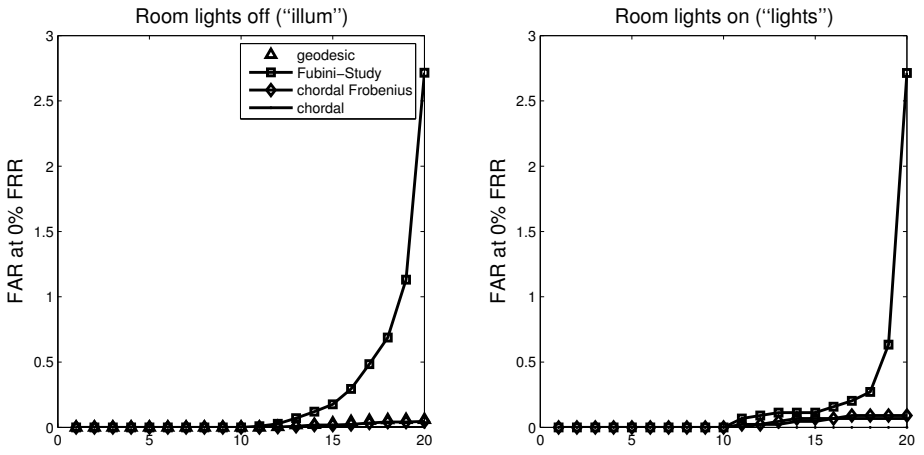


Figure 5.3: Plots of False Accept Rate (FAR) at a zero False Reject Rate (FRR) for the PIE database divided into frontal images with room lights on and room lights off. Between 1 and 20 principal angles are included in the Grassmannian semi-distance computation as shown along the horizontal axis.

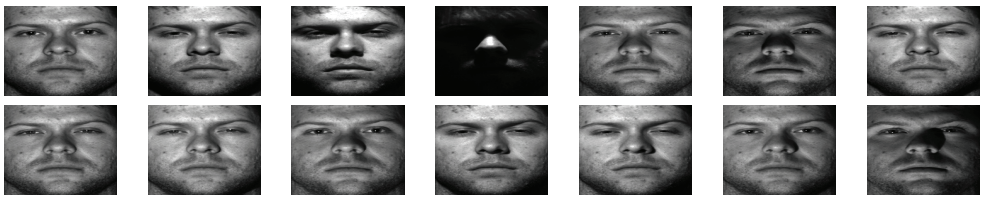


Figure 5.4: 14 example images of the Yale Face Database B.

The YDB experiment was also carried out for 21 and 32 samples and the results are summarized in Table 5.1.

It is important to notice that although YDB is not Grassmann separable when using 8 samples and their mirrors to create subspace representation, it is Grassmann separable when using 16 and more. Grassmann separability greatly depends on ways of partition on the available data, i.e., subspace configuration and dimension k. Also note that Fubini-Study tends to decrease its performance as we increase the number of principal angles in the construction of the metric. We suspect that this is due to the nature of the cosine function and the fact that two random vectors in a high-dimensional space are most likely to be orthogonal. While the later angles tend to be close to orthogonal, cosine of these angles tend to be close to zero. These small numbers erase any discriminatory ability offered by the smaller angles in the expression of Fubini-Study, hence causing

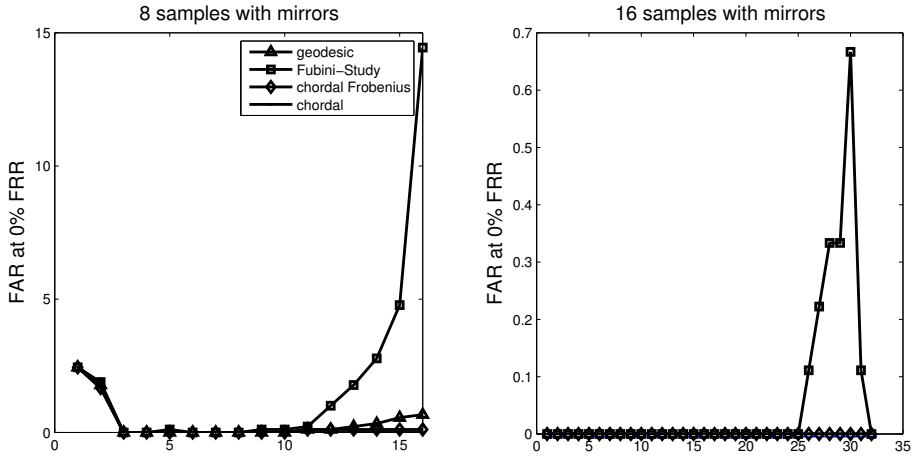


Figure 5.5: Plots of False Accept Rate (FAR) at a zero False Reject Rate (FRR) for YDB. Left: 8 samples and their mirrors are used in creating subspaces. Right: 16 samples and their mirrors are used in creating subspaces. The horizontal axis shows various various values of ℓ.

n	mirror	d_g	d_{cF}	d_c	d_{FS}
8	no	20.19	18.00	10.97	38.44
8	yes	1.07	1.04	1.02	1.06
16	no	0.07	0.06	0.02	0.13
16	yes	0.00	0.00	0.00	0.00
21	no	0.00	0.00	0.00	0.00
21	yes	0.00	0.00	0.00	0.00
32	no	0.00	0.00	0.00	0.00
32	yes	0.00	0.00	0.00	0.00

Table 5.1: Symmetric comparison. FAR at zero FRR for various ℓ-truncated Grassmannian semi-distance measures averaged over $5 \leq \ell \leq 10$, applied to illumination spaces estimated from n disjoint samples from YDB. d_g: geodesic; d_{cF}: chordal Frobenius; d_c: chordal (projection F); d_{FS}: Fubini-Study.

the metric to be ineffective. To understand the optimal number of angles needed in construction of each metric in order to ensure successful classification performance is itself an interesting topic and warrants careful examination in the future. In particular, methods which consider only the first principal angle or all the principal angles will fail to discover the Grassmann separability discovered by the ℓ-truncated measures, see the left plot of Figure 5.5. Furthermore, metric selection in achieving optimal classification outcome also plays an important role in geometric data analysis and is an open question in the community.

Chapter 6

FACE RECOGNITION UNDER VARYING ILLUMINATION AND POSE

Face recognition under variations in illumination and pose has long been recognized as a difficult problem with pose appearing somewhat more challenging to handle than variations in illumination [96]. A direct approach to deal with such images has been to develop algorithms that normalize for variations in illumination and then to focus on a solution for pose [71], [39]. In contrast, as is shown in [4] and [13], it is an appealing and plausible idea that sets of images acquired under varying or non-uniform illumination conditions possess valuable discriminatory information. Furthermore, both theoretical and empirical evidence have demonstrated that there exist low-dimensional representations for a set of images of a fixed object under variations in illumination conditions [7, 4]. This suggests that a wider range of discriminatory information can be captured in a low dimensional model as opposed to discarding a portion of the data as noise. This observation is not new and examples of algorithms that attempt to solve the face recognition problem under variations of illumination and pose without factoring out the illumination variations are [97, 40, 34, 10]. Readers who are interested in the details of the following discussions are referred to [16].

Algorithms that are successful in recognizing subjects in uncontrolled environments rely on good models for both illumination and pose variations. In the typical representation of image data, variation in illumination is inherently linear. More precisely, images collected under a convex set of illumination conditions themselves form a convex set [34] and a vast majority of the energy of such data can be captured with a relatively low-dimensional linear space [4]. In contrast, the collection of images captured under variations in pose is not inherently linear. As a consequence, linear methods such as

those based on the SVD perform poorly when pose variations are included. One natural nonlinear approach for addressing pose variations is with a 3D Morphable Model as described in [10]. Such non-linear approaches often come with the expense of a training phase and manual feature extraction at the recognition stage.

When a collection of images are available for a subject, we view the data as sampling (with noise) an underlying manifold. In this context, the underlying data manifold, M, captures pose and illumination variations of a fixed subject. If we fix an illumination condition, then the underlying data manifold, X, captures pose variation. There is a natural map $\phi : M \to X$ (under the fixed illumination condition). The fibers of the map (i.e. the inverse images of points on X) capture variations in illumination for a fixed pose. For each pose we can capture, with a low-dimensional linear space, the variations in illumination. Fixing the dimension of the linear space used to capture the illumination data to be k, we obtain a map of X into the parameter space of k-dimensional linear spaces inside the ambient space used to represent the images. We proceed with this model in the background. As we shall see, the Grassmann method works well in this face recognition problem without training. Literature review for this problem is presented in Chapter 6.1 and we present a suite of experiments that illustrate the effectiveness of the Grassmann method in Chapter 6.2. Note that we attempt to demonstrate the robustness of the Grassmann method by purposely omitting the preprocessing stage: we employ the Extended Yale Face Database B (E-YDB) [34] and CMU-PIE [74] Database with none of the images geometrically normalized and with registration essentially ignored.

6.1 Background

We review a few start-of-the-art models here to compare and contrast with the experimental results shown in Chapter 6.2. Works of [29, 50, 64] also used the set-to-set framework to solve a general object recognition problem under varying illumination and viewpoints. However, variations in illumination are normalized away before identification. We will focus on algorithms that model both the illumination and pose variations in the following discussions since we have established in the previous Chapter that illumination variations are potentially discriminatory.

A work which utilizes joint information given by variations in both illumination and pose and that is related to our work is given by Belhumeur et al. [34]. They developed a generative procedure that gives rise to a representation based on pose-specific

illumination cones for each face class. A single face representation consists of multiple pose-specific illumination cones that are approximated linearly to capture over 99% of the variability. Recognition of a test image is performed by first normalizing its vector representative to unit length and then computing its distance to each face representation. This way, the pose and identity information of the test image can be revealed. When tested on the Yale Face Database B [34], the average error rate reported in [34] is about 2.9% out of 4050 (45 illuminations $\times$ 9 poses $\times$ 10 subjects) images tested. The algorithm performs the worst on extreme illumination and pose conditions with 12.6% error rate out of 420 (14 illuminations $\times$ 3 poses $\times$ 10 subjects) images tested. Note that the most extreme illumination conditions were not even considered and it is reported in [72] that this is the best result obtained on YDB. Further note that the illumination cone method uses a single-to-many classification scheme, which is fundamentally different from the one afforded by the set-to-set method proposed in this thesis.

Another example by Gross et al. [40] uses the concept of light field [39] to handle pose variations and Fisher Discriminant Analysis (FDA) to handle the illumination variations, as described in Chapter 5.1.

The average error rate when using testing images from CMU-PIE Database (see Figure 6.1) that are comparable to those of the Extended YDB is about 10.5% (c07, c09, c11, c37), while the average error rate is about 14.5% on FaceIt, the commercial face recognition system from Visionics, using the same images from CMU-PIE Database [39]. When tested on the CMU-PIE Database, the average error rate for the Fisher light-field method is about 53%, while it is 59% for the eigen light-field method when the gallery set consists of simply the frontal pose and frontal illumination. The pose variations in CMU-PIE are significantly more difficult to recognize compared to those in the Extended YDB. The error rate is improved slightly when the variation of illumination is handled by a Lambertian reflectance model [97] with 47% in the most difficult case. In all the cases above, pose and illumination conditions for the probes are different than the ones from the gallery. Moreover, in all of the methods reported above, a significant amount of training is required.

It is reported in a recent survey paper [72] that the best recognition result on the "lights" subset of PIE is achieved by a 3D Morphable Model [10] when using only front, side, and profile views in both gallery and probe set. In particular, when the gallery and

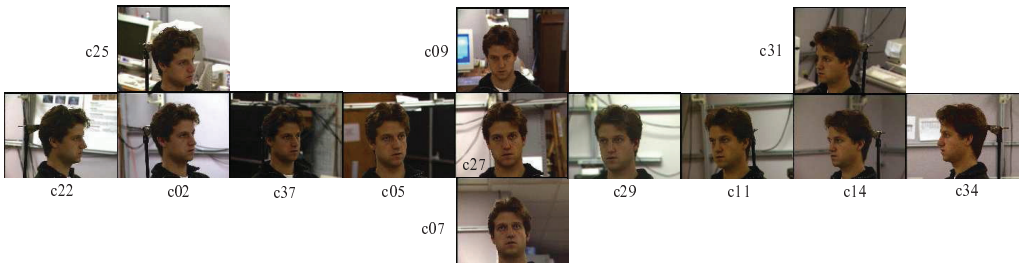


Figure 6.1: Illustration of pose variations in CMU-PIE Database.

probe sets both contain images of profile views, the recognition error rate is 10.6% across all illuminations. However, there are two limitations to the practical version of the 3D morphable model. Training of the faces are required in order to build a 3D model and it is necessary to manually select 7 landmark points on probe images to provide a good estimate of 3D pose. This hinders the algorithm from being automatic.

6.2 Empirical Results

The data sets we used to empirically test our algorithm are the Extended Yale Face Database B (E-YDB) [34] and the "illum" subset of the CMU-PIE Database [74]. For the E-YDB, there are 28 different subjects each recorded under 9 poses and 65 illumination conditions. For the CMU-PIE Database, there are 67 subjects each recorded under 13 poses and 21 illumination conditions. We denote the image corresponding to subject s, pose p and illumination condition i by $J_{s,p,i}$. See Figure 6.2 for an illustration of images with variations in pose and illumination from the E-YDB. As you can see, the images in these databases are coarsely centered and coarsely controlled. The rough nature of the data makes the experiments applicable to a wider variety of real-life applications. We consider three experiments to test the proposed algorithm.

In Experiment I, the poses are treated separately. In Experiments II and III, the poses are pooled for each subject. In Experiments I and II, we include the probe pose in the gallery and the probe and gallery images each use a distinct set of k illuminations taken from each pose. These two experiments are merely a sanity check. Any recognition algorithm that claims to be successful in dealing with variations of illumination and pose should perform very well in these two experiments. In Experiment III, we remove the

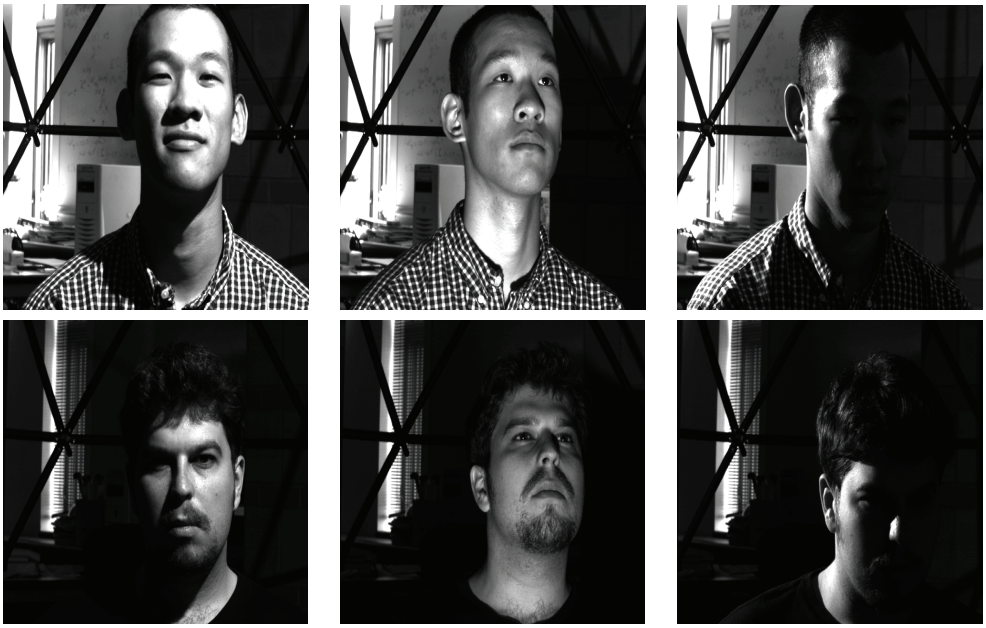


Figure 6.2: Example images in the E-YDB that are used in the experiments.

probe pose from the gallery in addition to using a distinct set of k illuminations. To this end, we test the algorithm's ability to recognize novel viewpoints.

We describe the experiments below for a single probe set. Let the number of distinct subjects in either the E-YDB or the CMU-PIE Database be s_0, the number of distinct poses be p_0, and the number of distinct illuminations be i_0. The distance measure used will be $d = d_c^\ell$ in the following experiment descriptions. We will describe how the set of probe and gallery images are selected, indicate how error statistics are compiled and analyze the results.

Experiment I

In Experiment I we view each pose as an additional subject in the database while retaining the information that each of the p_0 poses are associated with a given subject. The probe set (resp. gallery set) associated with subject α and pose β is written as $P_{\alpha,\beta}$ (resp. $G_{\alpha,\beta}$). We have

$$P_{\alpha,\beta} = \bigcup_{i \in I_P} J_{\alpha,\beta,i} \quad \text{and} \quad G_{\alpha,\beta} = \bigcup_{i \in I_G} J_{\alpha,\beta,i},$$

where I_P denotes the set of illuminations associated with the probe and I_G denotes the set of illuminations associated with the gallery. The set of indices defining I_P and I_G is

chosen randomly with the restriction that $I_P \cap I_G = \emptyset$. In Experiments I, II, and III, we let the cardinality of I_G and I_P be k and let them be denoted by $|I_G|$ and $|I_P|$, with $k = 16$ for E-YDB and $k = 10$ for CMU-PIE. For a fixed α and β,

$$P_{s^*,p^*} = \arg\min_{s,p} d(P_{\alpha,\beta}, G_{s,p}).$$

If $s^* = \alpha$ and $p^* = \beta$, then we have achieved the correct classification. Thus a single pose of a single subject is compared individually to each pose set of each subject. In essence this requires the algorithm to recognize both pose and identity. See Figure 6.3 for an illustration of the experiment.

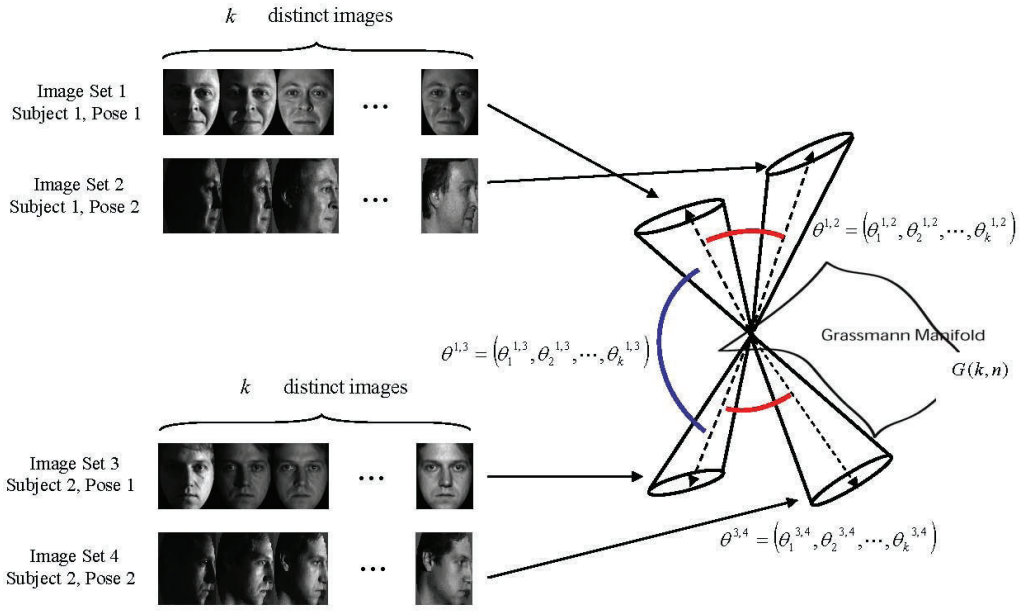


Figure 6.3: Illustration of Experiment I.

Experiment II

We now consider the case where, for the gallery, we pool the different poses and illuminations into a single set associated with each subject, i.e.,

$$G_s = \bigcup_{p=1}^{p_0} \bigcup_{i \in I_G} \{J_{s,p,i}\}.$$

Now the gallery set consists of s_0 sets of images where each set has p_0 different poses and $|I_G|$ illuminations. Again, a single probe set $P_{\alpha,\beta}$ is associated with one subject α and one pose β over a set of $|I_P|$ illuminations. For a fixed α and β, we solve

$$P_{s^*} = \arg\min_{s} d(G_s, P_{\alpha,\beta}).$$

If $s^* = \alpha$, then the classification is correct. See Figure 6.4 for an illustration of the experiment.

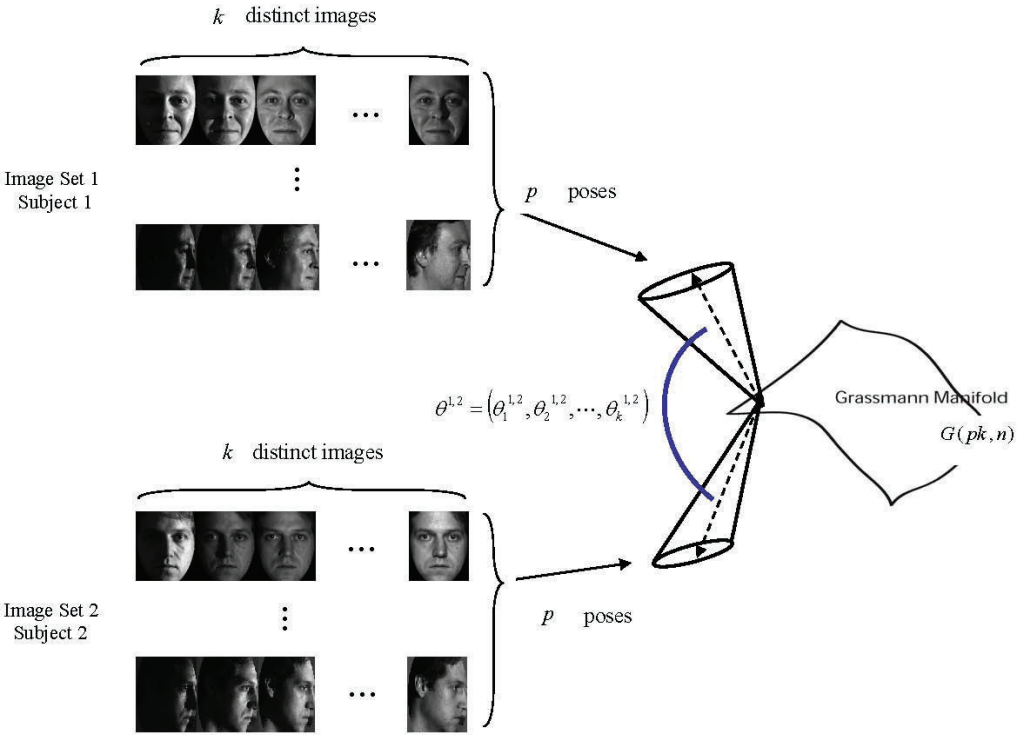


Figure 6.4: Illustration of Experiment II

Experiment III

In this experiment we remove the pose associated with the probe from all of the sets in the gallery. Hence, for each $\alpha = 1, \ldots, s_0$ and $\beta = 1, \ldots, p_0$ we seek to solve the equation

$$P_{s^*} = \arg\min_s d(G'_s, P_{\alpha,\beta}),$$

where

$$G'_s = \bigcup_{\substack{p=1 \\ p \neq \beta}}^{p_0} \bigcup_{i \in I_G} \{J_{s,p,i}\}.$$

If $s^* = \alpha$, then classification is correct. See Figure 6.5 for an illustration of the experiment.

For each experiment we compute the errors using each pose and each subject, therefore a total of $28 \times 9 = 252$ probe points for E-YDB and $67 \times 13 = 871$ for CMU-PIE. In addition, we randomly partition i_0 illumination conditions into two disjoint sets of k,

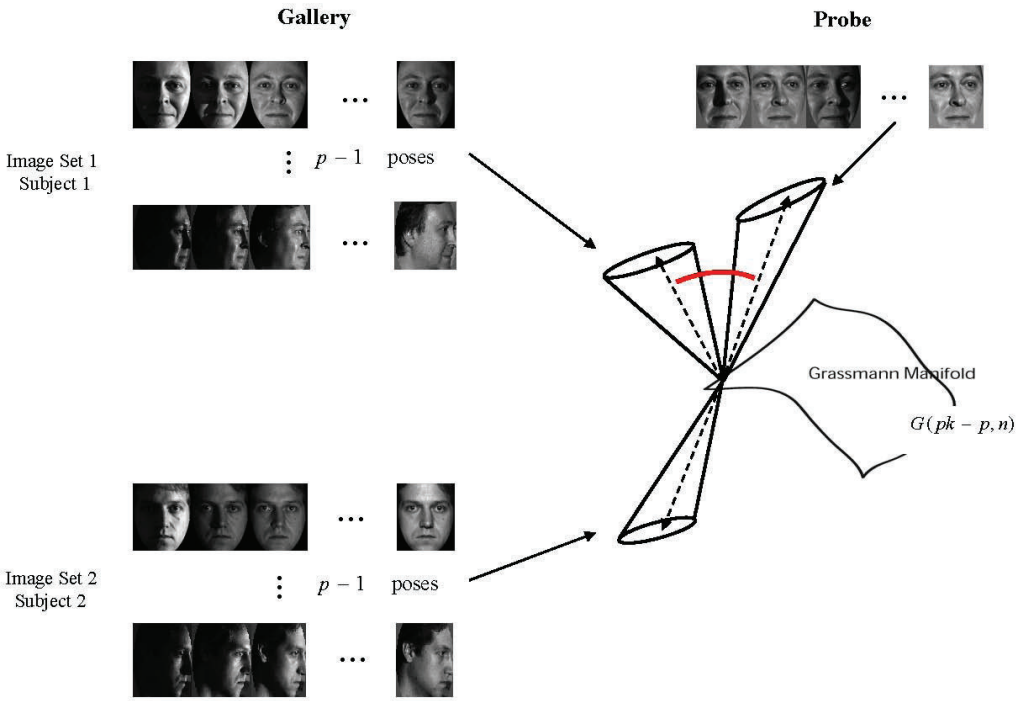


Figure 6.5: Illustration of Experiment III

Database	Experiment		
	I	II	III
Extended YDB	0	0	6.7
CMU-PIE	0	0	43.2

Table 6.1: Average recognition error rate for Experiments I – III with d_c^1 on both Extended YDB and CMU-PIE.

one for the gallery and the other for the probe. Table 6.1 shows the recognition rates in the nearest neighbor sense for experiments I – III on both databases when using the 1-truncated chordal distance d_c^1.

To visualize the contrast in performance of the algorithm for Experiments I and II versus Experiment III, we examine the first principal vector for a set of probe and gallery points using images in CMU-PIE in each Experiment in Figures 6.6 and 6.7. When the probe pose for each subject is present in the gallery, the algorithm is able to come up with a good representation in the gallery that models the pose in the probe. However, when the probe pose is not found in the gallery, the algorithm can only use the variations in the gallery to try and come up with a representation that matches the probe pose as

closely as possible. This observation is built upon the fact that a subject's pose manifold is nonlinear in its standard representation and we are using methods that are linear in nature.

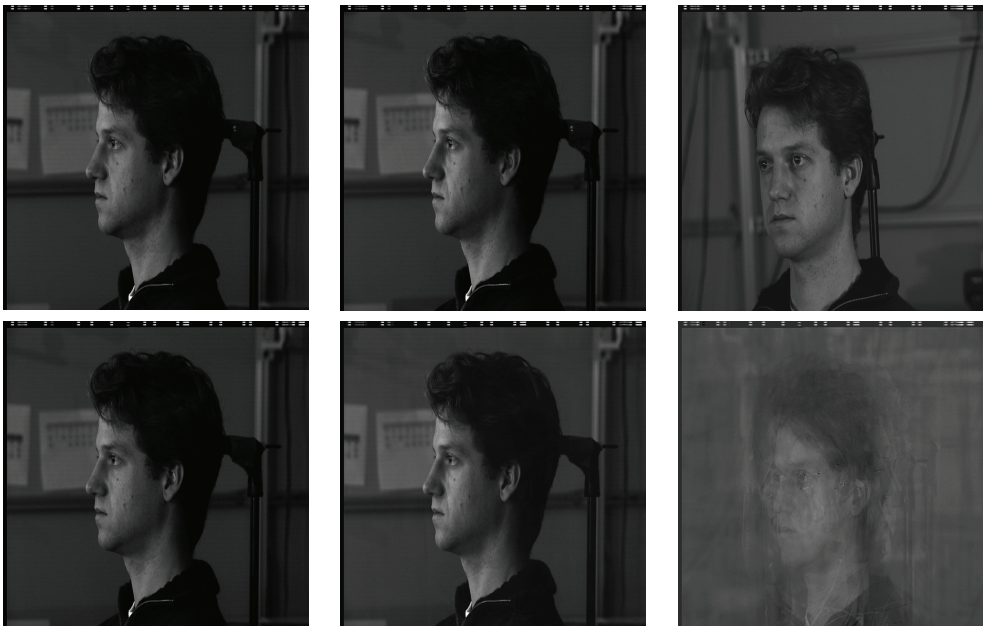


Figure 6.6: Top: first principal vector for a sample probe. Bottom: first principal vector for the correct gallery set that the algorithm identifies for experiments I, II , and III from left to right. The first principal angle between the top and bottom vector is 0.066, 0.069, and 0.269 radians, from left to right.

To further understand which viewpoints are difficult to handle, we look specifically at the individual error rates for each viewpoint in CMU-PIE in Table 6.2. We note that recognition results for probe poses c07, c09, c25, and c31 are not reported in studies [10, 97] and we observed the highest error rates on these poses in our experiments. Without consideration of those 4 poses, our error rate is approximately 26.9% for CMU-PIE in Experiment III. We suspect the reason for the degradation of the algorithm's performance for these 4 poses is due to the variation in depth of field on top of the actual pose variation.

We also conducted experiments where the only variations in the gallery and probe is the viewpoint. For each probe and gallery set in E-YDB and CMU-PIE, we randomly chose 4 and 6 non-overlapping poses and calculated their mirror images to create sets of 8 and 12 distinct pose images, respectively. Then for each $\alpha = 1, \ldots, s_0$, we solve

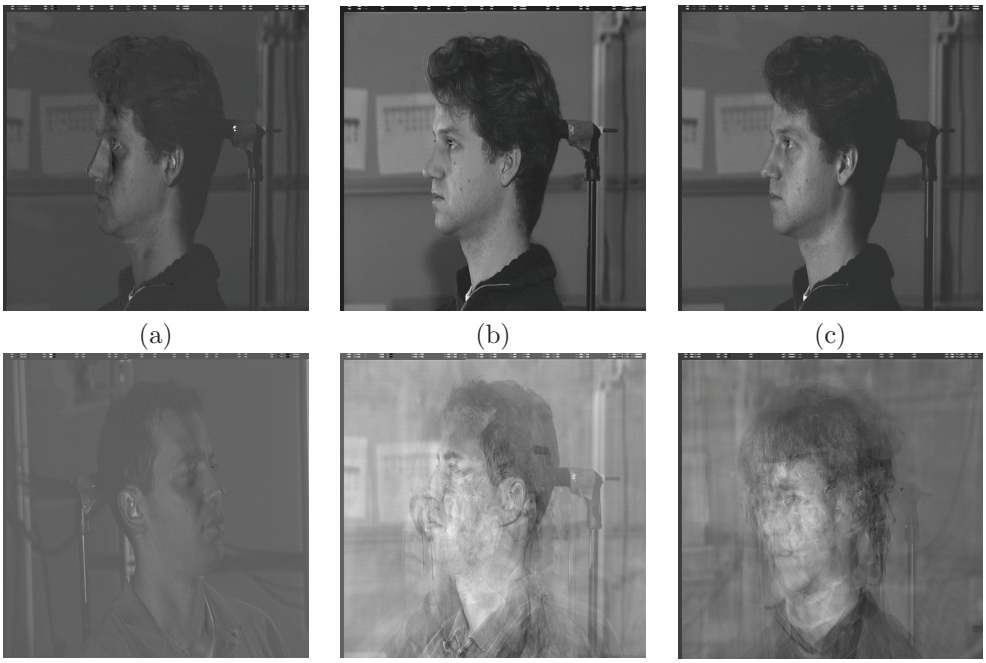


Figure 6.7: (c) The first principal vector for a sample probe. (f) The first principal vector for the incorrect gallery set that the algorithm identifies for experiments III. (a),(b) First principal vector for a sample probe point. (d),(e) First principal vector of a sample gallery point of a different subject from the ones in (a) and (b). The first principal angle between the top and bottom vector is 0.731, 0.275, and 0.369 radians, from left to right.

$P_{s^*} = \arg\min_s d(G_s, P_\alpha)$, where

$$G_s = \bigcup_{\substack{p \in P_G \\ i=\text{frontal}}} \{J_{s,p,i}\}$$

and P_G denotes the set of poses associated with this gallery set. We repeat this experiment 10 times to create a total of $10 \times s_0$ probe sets. The average error rate is 32.1% and 19.4% for E-YDB and CMU-PIE, respectively. We suspect the reason why the error rate for CMU-PIE is smaller than it is for E-YDB is because there are more pose variations in CMU-PIE, hence creating a better sampled characterization for the subjects. This observation supports the claim that the proposed algorithm captures the characteristics within a family of patterns and can be extended to handle other general object recognition problems.

As mentioned in Chapter 3.4, the Grassmann method comprises of two major steps: QR-decomposition of the representation matrices and SVD of the inner product of the

pose	c02	c05	c07	c09	c11	c14	c22
error (%)	13.4	31.3	83.6	73.1	0	1.5	23.9
pose	c25	c27	c29	c31	c34	c37	
error (%)	82.1	22.4	16.4	80.6	76.1	56.7	

Table 6.2: Average break-down recognition error rate for each pose in Experiment III using d_c^1 on CMU-PIE. We observe that some pose subsets perform much better than others.

	3DMM	Illum. Cone	Grassmann method
Data set used	"lights" of PIE	YDB	"illum" of PIE
Image resolution	200×200	42×36	367×401
Id time/probe	2.5 min. Pentium IV 2.0GHz	2.5 sec./gal. ind. Pentium II 300MHz	0.65 sec./gal. ind. AMD Opteron 2.8 GHz

Table 6.3: Computational speed of two state-of-the-art face recognition algorithms and the Grassmann method. Given the disparity in processors and in image resolution, care must be exercised in interpreting CPU time.

Q matrices. On a 2.8GHz AMD Opteron processor, it takes approximately 0.4 seconds to do a qr on a single probe set of size 147167×10 and 0.25 seconds to do an economy size SVD of a matrix of size 120×10. Therefore, it costs about 0.65 seconds to compare a single pair of probe and gallery point with the Grassmann method (in Experiment III on CMU-PIE). Table 6.3 shows the computational speed for a few state-of-the-art algorithms along with the Grassmann method. Note that our algorithm is significantly faster if the image resolution is smaller. Recall, we purposely omit the preprocessing stage in order to reflect the robustness of the algorithm, therefore the computational speed can be improved once low resolution images are used. On the other hand, we suspect that classification rates can be improved if images are geometrically registered. Moreover, novel probe pose can be better approximated from a larger pool of gallery poses. Thus, we envision a more successful classification result if a larger data set is used.

Chapter 7

FACE RECOGNITION WITH PATCH COLLAPSING

We concluded from Chapter 5 that digital images of a human face, collected under various illumination conditions, contain discriminatory information that can be used in classification. In this chapter, we will demonstrate that sufficient discriminatory information persists at ultra-low resolution to enable a computer to recognize specific human faces in settings beyond human capabilities. To obtain this result, we will introduce the notion of patch collapsing as a form of linear projection and review a well-known class of such projections in Chapter 7.2 after setting up the background work in Chapter 7.1. We will then utilize the Haar wavelet, which is a class of the patch collapsing, to modify a collection of images to emulate pictures from a 25-pixel camera and show that perfect classification results can be obtained on the CMU-PIE database with the Grassmann method in Chapter 7.3. Since facial imagery at ultra-low resolution is typically not recognizable or classifiable by human operators, we can envision saving large private databases of facial imagery at a resolution that is sufficiently low to prevent recognition by a human operator yet sufficiently high to enable machine recognition. Some discussions on why the method works well in this setting are given in Chapter 7.4. Readers who are interested in the details of the following discussions are referred to [15].

7.1 Background

A variety of studies consider the roles of data resolution and face recognition, including [53, 27, 60, 59, 86]. A common feature of these studies is the practice of using single-to-single image comparison in the recognition stage (with the exception of [86]). Among the techniques used to train the algorithms are PCA, LDA, ICA, Neural Network, and Radial Basis Functions. Some of the classifiers used are correlation, similarity

score, nearest neighbor, neural network, tangent distance, and multiresolution tangent distance. If variation in illumination is present in the data set, it is removed by either histogram equalization [25] or morphological nonlinear filtering [30]. Except in [86], the variation of illumination was treated as noise and eliminated in the preprocessing stage before the classification takes place.

In a more related study, Vasconcelos and Lippman proposed the use of transformation invariant tangent distance embedded in the multiresolution framework [86]. Their method, based on the (2-sided) tangent distance between manifolds, is referred to as the multiresolution tangent distance (MRTD) and is similar to our approach in that it requires a set-to-set image comparison. It is also postulated that the use of a multiresolution framework preserves the global minima that are needed in the minimization problems associated with computing tangent distances. The results of [86], however, are that when the only variation in the data is illumination, the performance of MRTD is inferior to that of the normal tangent distance and Euclidean distance. Hence, it appears that the framework of [86] does not sufficiently detect the idiosyncratic nature of illumination at low resolutions.

In summary, we use an algorithm for classification of image sets that requires no training and retains its high performance rates even at extremely low resolution. To our knowledge, no other algorithm has claimed to have achieved perfect separability of the CMU-PIE database at ultra low resolution.

7.2 Mathematics of Patch Collapsing

There are several families of linear transformations which are natural and useful to consider in the context of face recognition. Patch Collapsing is generally known as a technique to reduce resolution of images while maintaining global neighboring structures.

Definition 7.2.1. (Patch Collapsing) Consider a partition of the components of a vector, V, into disjoint sets $P_1 \cup P_2 \cup \cdots \cup P_d$. Patch collapsing is the operation of replacing, for each i between 1 and d, the components in P_i with a fixed weighted average of these components. This operation can be expressed as a linear map $L : \mathbb{R}^n \to \mathbb{R}^d \subset \mathbb{R}^n$.

Example 7.2.1. An example of the patch collapsing is the partitioning of a digital image into regions as provided by the scaling spaces in the Haar wavelet decomposition. See Figure 7.1 for illustrations of this type.

Figure 7.1: Illustration of patch collapsing using Haar wavelet.

Since we use a two-dimensional Discrete Wavelet Transform to emulate low resolution images, we will briefly review the general concept of *multiresolution analysis (MRA)* and its connection to the nested Grassmannians here.

MRA works by projecting data in a space V onto a sequence of nested subspaces

$$\cdots \subset V_{j+1} \subset V_j \subset V_{j-1} \subset \cdots \subset V_0 = V.$$

The subspaces V_j represent the data at decreasing resolutions and are called *scaling subspaces* or *approximation subspaces*. The orthogonal complements W_j to V_j in V_{j-1} are the *wavelet subspaces* and encapsulate the error of approximation at each level of decreased resolution. For any natural number j, we have an isomorphism

$$\phi^j : V_{j-1} \xrightarrow{\sim} V_j \oplus W_j.$$

Let $\pi^j : V_j \oplus W_j \to V_j$ denote projection onto the first factor and let $\psi^j = \pi^j \circ \phi^j$. This single level of subspace decomposition is represented by the commutative diagram in Figure 7.2(a).

Let $G(k, V)$ denote the Grassmannian of k-dimensional subspaces of a vector space V. Suppose that V, V' are vector spaces, and that $f : V \to V'$ is a linear map. Let $\ker(f)$ be its kernel and let

$$G(k, V)^\circ = \{A \subset V \mid \dim(A) = k \text{ and } A \cap \ker(f) = 0\}.$$

If $k + \dim \ker(f) \leq \dim V$, then $G(k, V)^\circ$ is a dense open subset of $G(k, V)$, so almost all points in $G(k, V)$ are in $G(k, V)^\circ$. Now if $A \cap \ker(f) = 0$, then $\dim f(A) = \dim A$, so f induces a map

$$f_k^\circ : G(k, V)^\circ \to G(k, V').$$

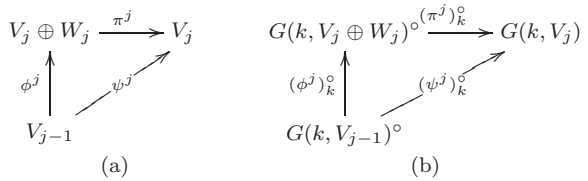


Figure 7.2: (a) Projection maps between scaling and wavelet subspaces for a single level of wavelet decomposition. (b) Projection maps between nested Grassmannians for a single level of decomposition.

Furthermore, if f is surjective then so is f°. The linear maps of the MRA shown in (a) of Figure 7.2 thus induce the maps between Grassmannians shown in (b) of the same figure.

Finally, we observe that if A, B are vector subspaces of V, then $\dim(A \cap B) = \dim(f(A) \cap f(B))$ if and only if $(A+B) \cap \ker(f) = 0$. In particular, when $(A+B) \cap \ker(f) = 0$ and $\ell \leq \min\{\dim A, \dim B\}$, then $d^\ell(A, B) = 0$ if and only if $d^\ell(f(A), f(B)) = 0$ for some Grassmannian distance d.

From this vantage point, we consider the space spanned by a linearly independent set of k images in their original space on the one hand, and the space spanned in their reduced resolution projections on the other hand, as points on corresponding Grassmann manifolds. Distances between pairs of sets of k distinct images or their low-resolution versions can then be computed using the truncated semi-distances d^ℓ on these Grassmann manifolds. The preceding observation insures that for sufficiently general resolution-reducing projections, spaces which were separable by d^ℓ remain separable after resolution reduction. Of course, taken to an extreme, this statement can no longer hold true. It is therefore of interest to understand the point at which the separability fails.

What we meant by *sufficiently general resolution-reducing projections* is really rank-preserving projections in the sense that after projection, the minimal principal angle does not become zero. In that case, the first principal angle between two elements on the Grassmannians decreases as the resolution decreases. Thus, as long as the wavelet transform preserves the rank of the subspaces at each level of decomposition, we are guaranteed to have decreasing minimal principal angles as the resolution decreases. The orthogonal Haar Wavelet is a sufficiently general resolution-reducing projection.

Now, we look at the result of applying MRA to digital face images. In a 2-dimensional Discrete Wavelet Transform (DWT), columns and rows of an image I each

(a) Original (b) LL (c) HL (d) LH (e) HH

Figure 7.3: An illustration of the sub-images from a single level of Haar wavelet analysis on an image in CMU-PIE. From left to right: original image, approximation, horizontal, vertical, and diagonal detail.

undergo a 1-dimensional wavelet transform. After a single level of a 2-dimensional DWT on an image I of size m-by-n, one obtains four sub-images of dimension $\lceil \frac{m}{2} \rceil$-by-$\lceil \frac{n}{2} \rceil$. If we consider each row and column of I as a 1-dimensional signal, then the *approximation* component of I is obtained by a low-pass filter on the columns then a low-pass filter on the rows and sampled on a dyadic grid. The other 3 sub-images are obtained in a similar fashion and collectively, they are called the *detail* component of I. The approximation component of an image after a single level of wavelet decomposition with the Haar wavelet is equivalent to averaging the columns, then the rows. See Figure 7.3 for an illustration of the sub-images obtained from a single level of Haar wavelet analysis.

To use wavelets to compress a signal, we sample the approximation and detail components on a dyadic grid. That is, keeping only one out of two wavelet coefficients at each step of the analysis. The approximation component of the signal, A_j, after j iterations of decomposition and down-sampling, will serve as the same image in level j with resolution $\lceil \frac{m}{2^j} \rceil$-by-$\lceil \frac{n}{2^j} \rceil$. In the subsequent discussions, we present results obtained by using the *approximation* subspaces. However, similar results obtained by using the *wavelet* subspaces are also observed.

7.3 Empirical Results

The experiment presented here follows the protocols set out in [13], where it was established that CMU-PIE is *Grassmann separable*. This means that using one of the distances d^ℓ on the Grassmannian, the distance between an estimated illumination space of a subject and another estimated illumination space of the same subject is always less than the distance to an estimated illumination space of any different subject. In this new experiment we address the question of whether this idiosyncratic nature of the illumination spaces persists at significantly reduced resolutions. As described below,

we empirically test this hypothesis by calculating distances between pairs of scaling subspaces. The results of the experiments performed on the "illum" subset of the CMU-PIE database is summarized in Figure 7.5. The results obtained by running the same experiment on the "lights" subset were not significantly different.

For each of the 67 subjects, we randomly select two disjoint sets of 10 images to produce two 10-dimensional estimates of the illumination space for the subject. Two estimated spaces for the same subject are called matching subspaces, while estimated subspaces for two distinct subjects are called non-matching subspaces. The process of random selection is repeated 10 times to generate a total of 670 matching subspaces and 44,220 non-matching subspaces. We mathematically reduce the resolution of the images using the Haar wavelet, effectively emulating a camera with a reduced number of pixels at each step. As seen in Figure 7.4, variations in illumination appear to be retained at each level of resolution, suggesting that the idiosyncratic nature of the illumination subspaces might be preserved. At the fifth level of the MRA the data corresponds to that which would have been captured by a camera with 5×5 pixels. We observe that at this resolution the human eye can no longer match an image with its subject.

The separability of CMU-PIE at ultra low resolution is verified by comparing the distances between the matching to the non-matching subspaces as points on a Grassmann manifold. When the largest distance between any two matching subspaces is less than the smallest distance between any two non-matching subspaces, the data is called Grassmann separable. This phenomenon can be observed in Figure 7.5. The three lines of the box in the box whisker plot shown in Figure 7.5 represent the lowest quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data and outliers are data with values beyond the ends of the whiskers.

Using d^1, i.e., a distance based on only one principal angle, we observe a significant separation gap between the largest and smallest distance of the matching and non-matching subspaces throughout all levels of MRA. Specifically, the separation gap between matching and non-matching subspaces is approximately 16°, 18°, 17°, 14°, 8°, and 0.17° when subspaces are realized as points in $G(10, 22080)$, $G(10, 5520)$, $G(10, 1400)$, $G(10, 360)$, $G(10, 90)$, and $G(10, 25)$, respectively. Note that the non-decreasing trend of the separation gap is due to the random selection of the illumination subspaces.

As expected, the separation gap given by the minimal principal angle between the matching and non-matching subspaces decreases as we reduce resolution. But never to the level where points on the Grassmann manifold are misclassified. In other words, individuals can be recognized at ultra-low resolutions provided they are represented by multiple image sets taken under a variety of illumination conditions.

It is curious to see if similar outcomes can be observed when using unstructured projections, e.g., random projections, to embed subject illumination subspaces into spaces of significantly reduced dimensions. To test this, we repeated the experiments described above in this new setting. Subject illumination subspaces in their original level of resolution were projected onto low dimensional spaces via randomly determined linear transformations. Error statistics were collected by repeating the experiment 100 times. Perfect separation between matching and non-matching subspaces occurred when subject illumination subspaces were projected onto random 35-dimensional subspaces. This validates the use of digital images at ultra low resolution and emphasizes the importance of illumination variations in the problem of face recognition. Furthermore, while unstructured projections perform surprisingly well in the retention of idiosyncratic information, structured projections that exploit similarities of neighboring pixels allow perfect recognition results at even lower resolutions.

We remark that the idiosyncratic nature of the illumination subspaces can be found not only in the scaling subspaces, but also in the wavelet subspaces. Indeed, we observed perfect separation using the minimal principal angle in almost all scales of the wavelet subspaces.

7.4 Discussions

We have shown that a mathematically emulated ultra low-resolution illumination space is sufficient to classify the CMU-PIE database when a data point is a set of images under varying illuminations, represented by a point on a Grassmann manifold. We assert that this is only possible because the idiosyncratic nature of the response of a face to varying illumination, as captured in digital images, persists at ultra low resolutions. This is perhaps not so surprising given that the configuration space of a 25-pixel camera consists of 256^{25} different images and we are comparing only 67 subjects using some 20 total instances of illumination. The representation space is very large compared to the

(a) (b)

Figure 7.4: Left to right: four distinct illumination images of subjects 04006 (a) and 04007 (b) in CMU-PIE. To to bottom: level one to level five approximation obtained from applying 2D discrete Haar wavelet transform to the top row.

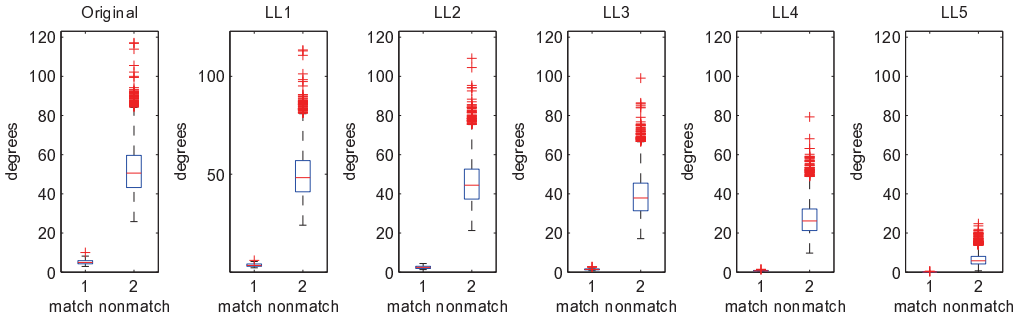


Figure 7.5: Box whisker plot of the minimal principal angles of the matching and non-matching subspaces. Left to right: original (resolution 160×138), level one Haar wavelet approximation (80×69), level two (40×35), level three (20×18), level four (10×9), level five (5×5). Perfect separation of the matching and non-matching subspaces is observed throughout all levels of MRA.

amount of data being stored. Furthermore, the reduction of resolution that was utilized takes advantage of similarities of neighboring pixels. The algorithm introduced here is computationally fast and can be implemented efficiently. In fact, on a 2.8GHz AMD Opteron processor, it takes approximately 0.000218 seconds to compute the distance between a pair of 25-pixel 10-dimensional illumination subspaces.

Chapter 8

FACE RECOGNITION WITH PATCH PROJECTION

Experience suggests that whenever identification of an individual is made by a human operator, a computer, or some combination of the two, both the holistic characteristics and the local characteristics of the individual can aid in a successful classification. The role, suggested by experience, of both local and global data has been confirmed by psychophysicists, neuroscientists and data analysts [18]. One can imagine a situation where only local features of an individual are available (perhaps due to an occlusion or some other modification). In this setting, it is useful to be able to extract as much information as possible from the data on hand.

We concluded from Chapter 5 that digital images of a human face, collected under various illumination conditions, contain discriminatory information that can be used in classification. Other work has focussed on the ability to classify using feature patches of the human face (i.e. sub-images of an image of a human face such as nose, eyes, lips, etc.). In this chapter, we will first set up the background works in Chapter 8.1 and define a mathematical notion of feature patches in Chapter 8.2. We combine the use of feature patches and subject illumination spaces to demonstrate that sufficient discriminatory information persists in feature patch illumination spaces to enable a computer to recognize specific human faces in settings far beyond human capabilities. Not unexpectedly, the amount of discriminatory information contained in a feature patch varies depending on the size and location of the patch. It is shown in Chapter 8.3 that for some feature patches, perfect classification rates were achieved for the 67 individuals in the CMU-PIE database using well under one percent of the image. In particular, it is shown that when the images are geometrically normalized based on known eye locations and coupled

with variations of illumination, characteristics given by the rough locations of mouth, nose, eyes, and even cheeks offer discriminatory information and can be used to classify identities without error when applied to the 67 individuals in the CMU-PIE database. Discussions on future use of these feature patches in classification problems are given in Chapter 8.4. Readers who are interested in the details of the following discussions are referred to [17].

8.1 Background

The most prominent and known techniques among an appearance-based holistic approach in face recognition are via eigenfaces [76, 84] and Fisherfaces [6]. In a more general object recognition scheme, Multiresolution Analysis (MRA) and Fourier Analysis (FA) are best known for extracting global features, where frequency information of the images can be seen as a global characteristic. A wide variety of face recognition algorithms are driven, at some level, by the usage of MRA and FA.

The conventional sense of geometric feature-based matching for face recognition is based on relative position and other parameters of distinctive features such as eyes, mouth, nose, and eyebrows. It started in the early 1970's with Kanade [45]. This way of performing face recognition is far from automatic. Feature points need to be labeled manually for every single image in the gallery as well as every probe image. An example of a feature-based face recognition method with a relatively high success rate is Elastic Bunch Graph Matching [89]. Coldstein [35] and Kaya [47] showed that a face recognition algorithm using features extracted manually enjoys a relatively successful performance. In general, feature-based methods are less sensitive to variations in illumination and viewpoint [96]. Nevertheless, automatic feature extraction techniques do not have sufficient reliability to justify the accuracy of feature-based face recognition approaches.

A related but philosophically different technique to feature-based matching is template matching, which uses specific local template(s) of the face to represent the whole face. Brunelli and Poggio showed in [11] that template matching is superior in recognition performance than feature-based matching. Moreover, templates of eyes are more

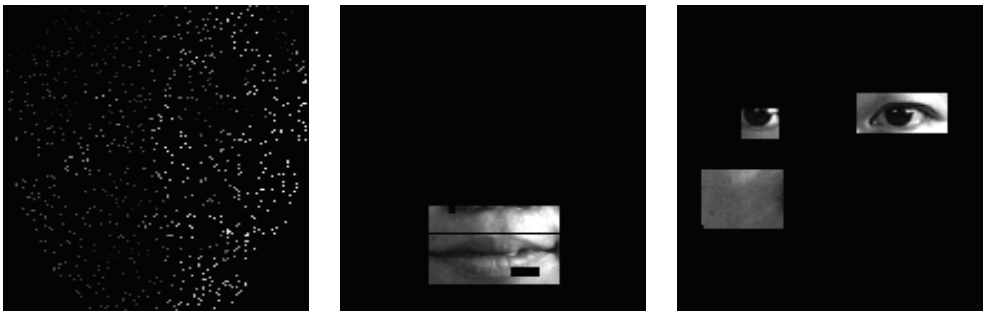


Figure 8.1: Illustration of patch projections. Patches do not have to be selected from a connected nor a rectangular region.

discriminating than templates of noses, which are then more discriminating than templates of the mouth in a correlation-based algorithm. They also suggested increasing the number of images used per template to boost recognition performance.

As it may seem that the use of global (MRA) and local (template matching) features in face recognition are completely different from each other, it turns out that they are fundamentally the same. Both manners of selecting features correspond to mapping the data, consisting of the original digital image of the whole face, into a representation space of smaller size that preserves discriminatory structure.

8.2 Mathematics of Patch Projection

Definition 8.2.1. (Patch Projection) Given a partition of the components of a vector, V, into disjoint sets $P_1 \cup P_2 \cup \cdots \cup P_d$. A family of patch projections is given by the natural projection maps $L_i : \mathbb{R}^n \to \mathbb{R}^{|P_i|}$.

Example 8.2.1. An example of the patch projection is the restriction of a digital image to a region of the image. For instance, the restriction of a digital image of a face to the region surrounding the lips. See Figure 8.1 for illustrations of this type.

When the notion of patch projection is applied to images, a feature patch of an image is simply a sub-image and is naturally amenable to the mathematics of moduli of linear spaces. A few comments on the relationship between projections and Grassmannians are in order:

Let K be the kernel of a linear map $L : \mathbb{R}^n \to \mathbb{R}^m$. Let $\bar{\Omega}(K) \subseteq G(k,n)$ denote the Schubert variety defined by

$$\bar{\Omega}(K) = \{E \in G(k,n) \,|\, \dim(E \cap K) \geq 1\}.$$

L induces a natural map $L_G : G(k,n) - \bar{\Omega}(K) \to G(k,m)$ since the image of any k-dimensional subspace $V \subset \mathbb{R}^n$ under L remains k-dimensional precisely if the point $p \in G(k,n)$ corresponding to V lies outside of $\bar{\Omega}(K)$. Suppose $\dim(K) + k \leq n$, then $\bar{\Omega}(K)$ is a proper subset of $G(k,n)$ and the dimension of $\bar{\Omega}(K)$ is strictly less than the dimension of $G(k,n)$ at a generic point. Thus, with probability one, a point chosen at random from $G(k,n)$ will lie in $G(k,n) - \bar{\Omega}(K)$. Due to the method we use to determine points on $G(k,n)$ and the special nature of patch projections, it is possible for the corresponding linear spaces to have a non-trivial intersection with the kernel of the projection map. However, as one would expect, never have we observed a point chosen within $\bar{\Omega}(K)$.

8.3 Empirical Results

For the empirical analysis that follows, we consider two types of classification schemes — FAR and the nearest neighbor (NN) classifiers. A pair of points on a Grassmannian is correctly classified in the FAR sense if the largest distance between a point in the gallery and a point in the probe for a matched pair is smaller than the smallest distance between any non-match pairs. On the other hand, for a data set that attains a zero in the nearest-neighbor sense, it means that all points in the probe are mapped correctly to the identities among the gallery.

To compare performance associated with direct image comparison using the same data, we employ a baseline similarity $S(X,Y)$ algorithm for comparing sets of images X and Y.

Definition 8.3.1. Let $X \in \mathbb{R}^{n \times k_x}$, $Y \in \mathbb{R}^{n \times k_y}$. Set

$$s(x^{(j)}, Y) = \max_{1 \leq i \leq k_y} \{\mathrm{Cor}(x^{(j)}, y^{(i)})\},$$

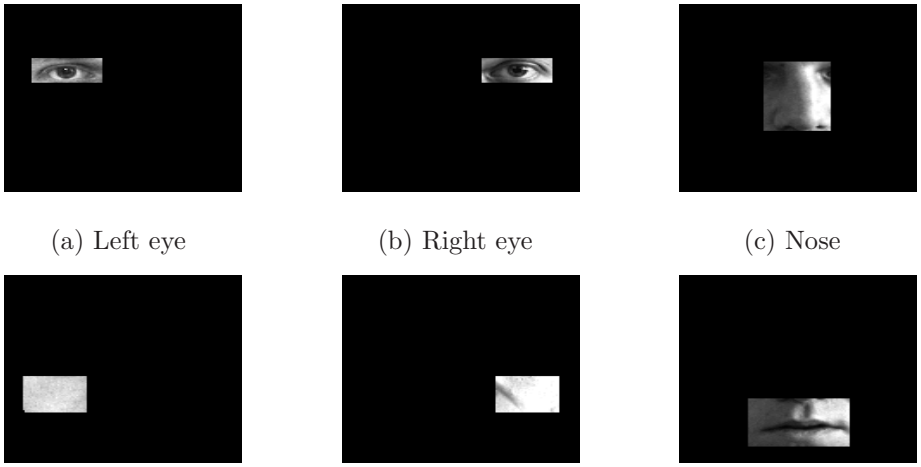
where $x^{(j)}$ and $y^{(j)}$ is a column vector in X and Y, respectively. Then a *baseline similarity* is defined as

$$S(X,Y) = \frac{1}{2} \sum_{i=1}^{k_x} \sum_{j=1}^{k_y} \left(s(x^{(i)}, Y) + s(y^{(j)}, X) \right),$$

where $\mathrm{Cor}(x, y)$ stands for the standard Pearson's correlation between signals x and y.

Notice that $S(X, Y)$ is nothing more than a straight across comparison between images, and therefore serves as an excellent baseline in evaluating the performance of our method.

The data sets we used here are the "illum" and "lights" subsets of the CMU-PIE Database[1]. The images are geometrically normalized according to known eye coordinates. The viewpoint is fixed to be frontal and subsets of 21 distinct illumination conditions are used to form the probe and gallery. We perform two experiments on the lip, nose, left eye, right eye, left cheek, and right cheek patches, see Figure 8.2, to show their effectiveness as illumination feature classifiers. Error rates for both classification schemes are reported in this section.



(a) Left eye (b) Right eye (c) Nose

(d) Left cheek (e) Right cheek (f) Lip

Figure 8.2: Example feature patches that are used for the algorithm.

Experiment I: Connected Patches

For each of the 67 people in CMU-PIE data set, we randomly select two sets of images of equal size with disjoint illumination conditions. Since illumination spaces can be well-approximated by 9- or 10-dimensional linear subspaces [4, 7], we randomly select two

[1]Note that results achieved on the "illum" subset are comparable to those done on the "lights" subset and are not reported here.

disjoint sets of size 10 for the points in the probe and gallery. This process is repeated 10 times, thus making a total of 670 probe points. Now, instead of the whole face image, selected feature patches are used. The result of this experiment is given in Table 8.1 along with the patch resolutions and the computational time it takes to calculate the distance between a single pair of probe and gallery points. Results for the baseline algorithm are also shown. Notice that while the Grassmann method performs without error on this task, the baseline algorithm performs poorly and is computationally more expensive than classification on Grassmannians.

It is apparent from the results of Experiment I that when the cardinality of points in the gallery and probe is ten, the algorithm is able to separate all people in the data set using each of the selected patches without error. To further speed up the classification time and to see how sensitive the proposed algorithm is to the location of the feature patches, we repeat the experiment while reducing the patch resolution until the perfect recognition rates cease to exist. Table 8.2 gives the conditions for perfect recognition results in the FAR sense while Table 8.3 gives the same thing but in the NN sense. Notice that the baseline algorithm is extremely sensitive to patch resolutions and less efficient. For example, while using 30-pixel nose patches with NN classifier, the baseline algorithm attains an error rate of 97.46% and it takes 74 times longer to compute the similarity between a single pair of probe and gallery points than it takes with the proposed algorithm. The time it takes to identify a single probe point in a database of 67 persons with 30-pixel nose patches using the proposed algorithm is only 0.014627 seconds. The results here suggest that locally correlated feature patches consisting of an extremely small number of pixels provide sufficient information for recognition.

Experiment II: Imbalanced Cardinality

In this experiment, we examine the effect of varying the cardinality of the probe and gallery. Often times, it is unrealistic to collect equal number of images at enrollment and during operation. Therefore, it is hard to avoid comparisons of sets of images of asymmetric sizes. In such cases, we would like to know the minimal number of images needed to represent a person while still achieving perfect separation. Figure 8.3 shows the classification error rates for each of the six selected patches. The cardinality of the probe points increase from 1 to 20 while the cardinality of the gallery points decreases from 20 to 1 simultaneously. The illumination conditions for the probe are always different from

		lip	nose	left eye	right eye	left cheek	right cheek
	Exp. img.						
GS method	Resolution	41×59	59×39	21×41	21×41	31×37	31×37
	CPU time	0.0037	0.0034	0.0011	0.0011	0.0014	0.0014
	FAR	0	0	0	0	0	0
	NN	0	0	0	0	0	0
Baseline	CPU time	0.0254	0.0249	0.0187	0.0187	0.0198	0.0198
	FAR	0.3008	1.2234	2.5690	4.8937	2.2388	4.8937
	NN	11.7910	0	6.4179	1.7910	0.5970	4.4776

Table 8.1: Error rates (in %) for individual feature patches where 10 images are used to compute each point in the probe and gallery. On a 2.8GHz AMD Opteron processor, the CPU time is how long it takes to calculate the distance/similarity between a probe and a gallery point in seconds.

		lip	nose	left eye	right eye	left cheek	right cheek
	Exp. img.						
	Resolution	3×29	35×13	21×41	21×41	31×37	31×37
	CPU time	2.7×10^{-4}	6.3×10^{-4}	0.0011	0.0011	0.0014	0.0014
Bsl.	FAR	6.8204	1.2121	3.1592	6.5762	4.1995	0.5812
	CPU time	0.0158	0.0171	0.0187	0.0186	0.0199	0.0196

Table 8.2: Conditions for perfect separation in the FAR sense for individual feature patches where 10 images are used to compute each point in the probe and the gallery.

		lip	nose	left eye	right eye	left cheek	right cheek
	Exp. img.						
	Resolution	1×33	30×1	4×22	4×23	19×6	23×27
	CPU time	2.2×10^{-4}	2.2×10^{-4}	2.8×10^{-4}	2.8×10^{-4}	3.0×10^{-4}	8.1×10^{-4}
Bsl	NN	28.0597	97.4627	14.1791	17.9104	44.4776	7.9104
	CPU time	0.0151	0.0151	0.0157	0.0157	0.0159	0.0178

Table 8.3: Conditions for perfect separation in the NN sense for individual feature patches where 10 images are used to compute each point in the probe and the gallery.

the ones in the gallery. The plot suggests the performance of the algorithm is optimal when the cardinality of the probe and gallery points approaches each other in the FAR sense. For example, when using only 1 image per person in the probe and 20 images per person in the gallery, the error rate is about 2.2% in the FAR sense, while the error rate is diminished to zero when using 3 images per person in the probe and 18 images per person in the gallery.

In the worse case scenario, if it is only possible to collect a single image for each probe, then we would like to know the minimal number of images required for each person in the gallery in order to obtain perfect separation. For this set of experiments, we use a single image for each probe and let the cardinality of the gallery points vary from 1 to 20. The classification error rates for each of the selected patches are given in Figure 8.4. For example, when the lip feature is used, the algorithm performs perfectly using only 9 images and 16 images per person in the gallery in the NN sense and FAR sense, respectively. However, when the cheek features are used, even the use of 20 images per person in the gallery could not force a perfect recognition in both classifiers. Suggestively, certain features (e.g., nose, lip) provide more discriminatory information than others (e.g., cheeks) when classification is done on the Grassmannians.

Experiment III: Random Patches

Here we explore whether or not patches that consist of small number of disconnected pixels will contain enough discriminatory information to accomplish the task of face recognition on a Grassmannian. Specifically, we employ feature patches consisting of a random (but the same for each image) selection of 36 pixels. See Figure 8.5(a) for an example image of such patch. A set of 10 different illuminations is used for both the gallery and probe. Hence, the data is represented as points on $G(10, 36)$. It turns out that the idiosyncratic nature of the patches persists in this case. We perform Experiment I again, but now using randomly projected low-dimensional patches and still observe error-free identification for all people in the PIE Database. Perhaps surprisingly, a similar result is observed even when we use a thin horizontal strip of 33 pixels across the left eye. See Figure 8.5(b) for an example image of such a patch.

Experiment IV: Robustness to Registration

To further understand the effect of inconsistent registration, we repeat Experiment I with varying registration. All images are randomly shifted either horizontally or vertically

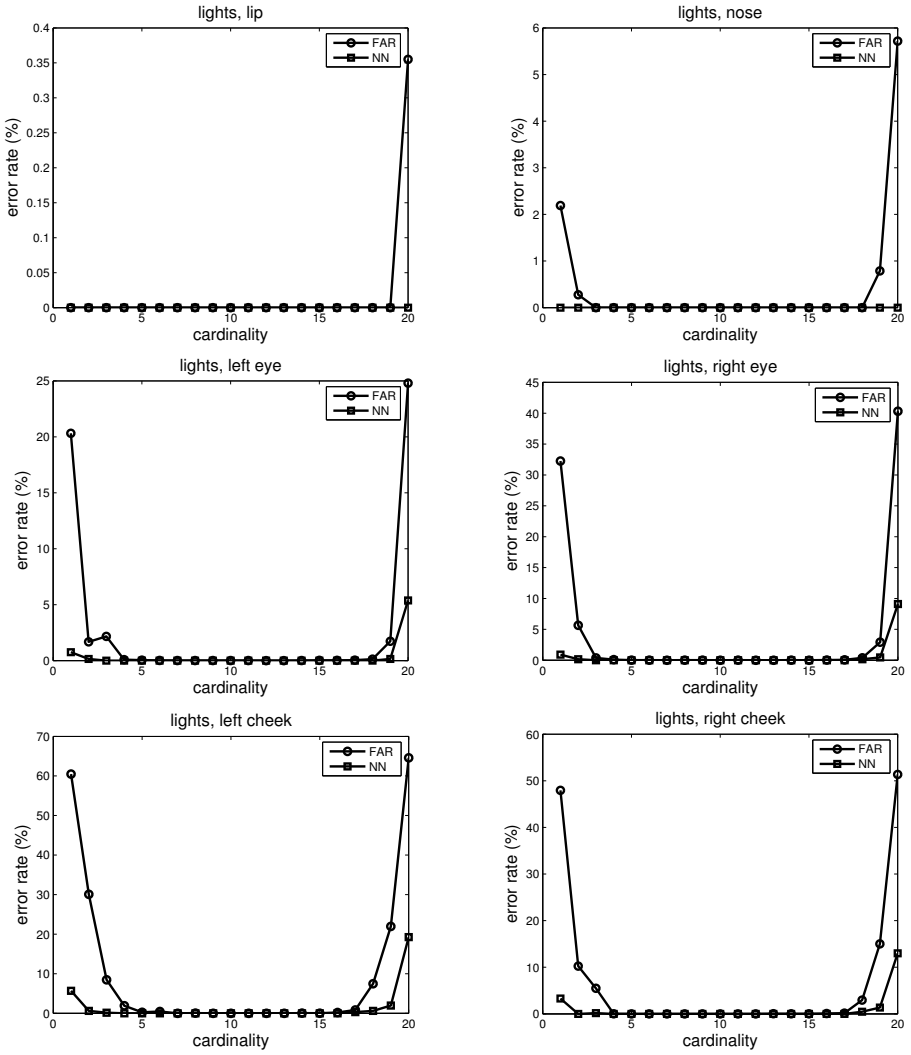


Figure 8.3: Error rates (in %) on the nose patch. The cardinality of points in the probe increases from 1 to 20 while the cardinality of points in the gallery simultaneously decreases from 20 to 1.

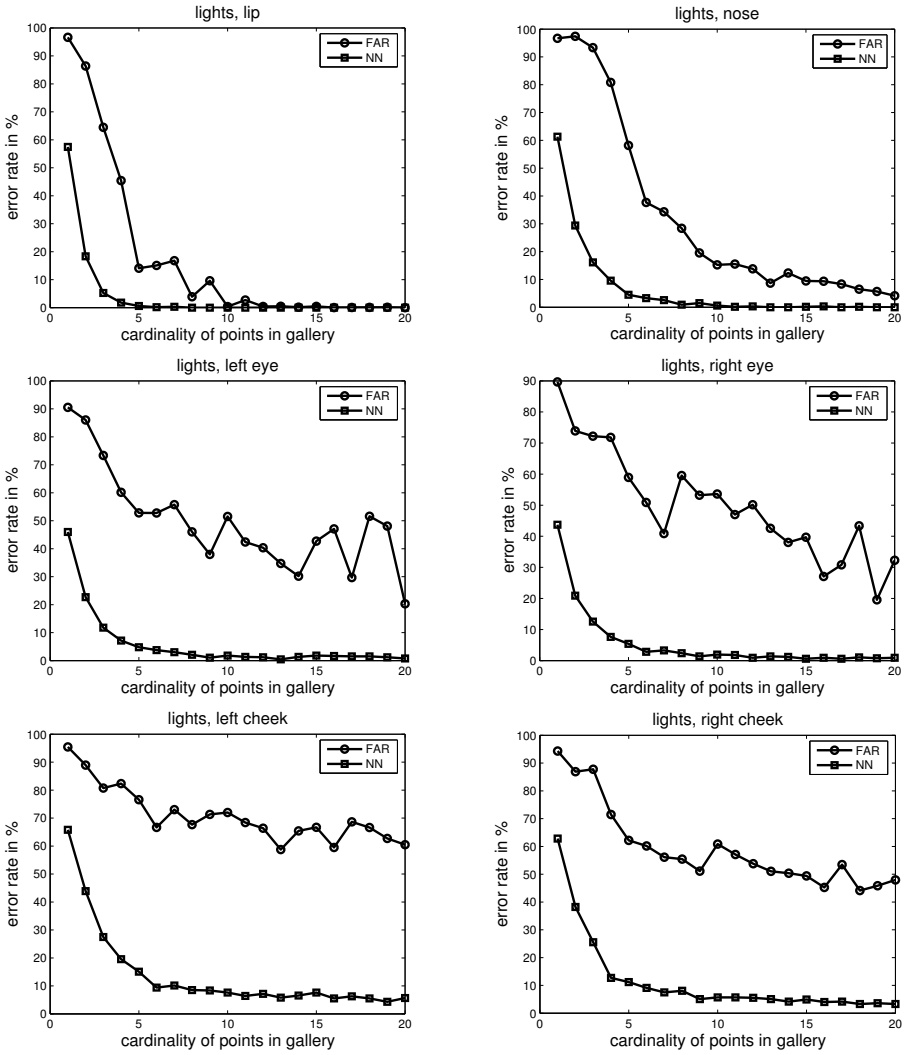
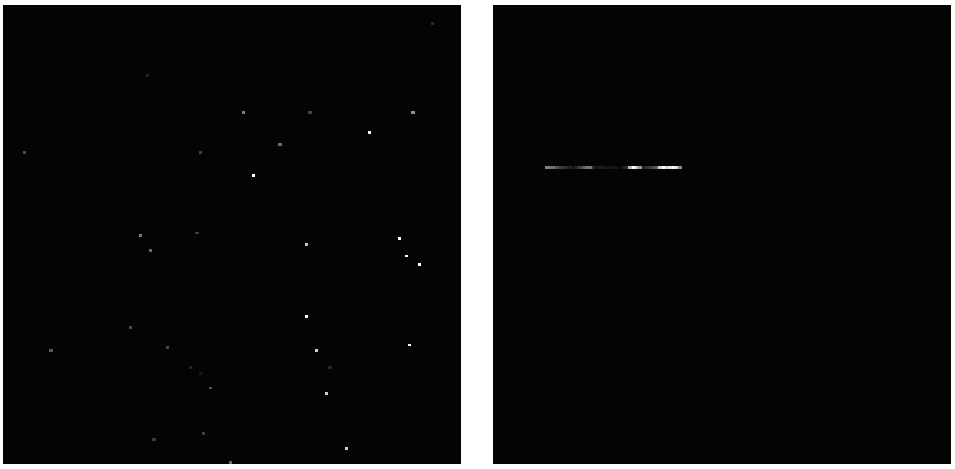


Figure 8.4: Classification error rates (in %) for each selected feature patch. The cardinality of the probe points is one while the cardinality of the gallery points ranges from 1 to 20.



(a) (b)

Figure 8.5: (a) A patch of 36 pixels are randomly selected from a face. (b) A patch of 33 pixels across the left eye.

one pixel at a time. Classification is repeated for every pixel shift up to 10 pixels using the new registered images to obtain error statistics in both classifiers. The lip and nose patches are least sensitive to perturbation of registration, see, e.g., the error rates given in Figure 8.6. The result here implies that if a human operator registers the gallery patches in a certain way, then another human operator can have about 2 pixels of freedom in registering the probe patches. Of course, expanding the data sets to include data that is poorly registered might improve this tolerance further.

8.4 Discussions

The work in this chapter and Chapter 7 build on the notion that variations in the state of an object can provide discriminatory information. Further, that the nature of this information may arise from global features of the pattern or, alternatively, from local features that possess their own special characteristics under a variation of state.

We see that the notion of a feature, or image patch generally defined, provides enhanced opportunities for pattern classification. The fact that certain local feature patches are more discriminatory than others suggests that weakly discriminating features can combine to form stronger ones. While this idea of course is itself not new, it is presented in the geometric context of Grassmann manifolds where large quantities of

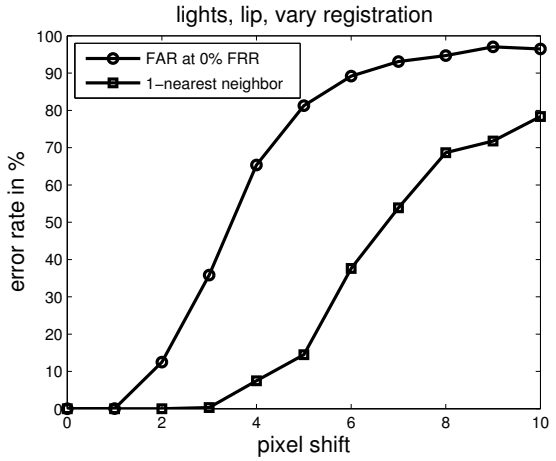


Figure 8.6: Error rates (in %) for the lip patch where probe and gallery points have cardinality 10. The raw images are randomly shifted either horizontally or vertically by one pixel at a time.

illumination data associated with a person may be effectively encoded and compared as a generalized point on a Grassmannian.

Both Chapter 7 and the present chapter provide a way to compress data and accelerate the computations without sacrificing performance. This type of compression yields new collections of points on the Grassmann manifolds, since compression of this sort reduces n that corresponds to reducing the number of pixels representing a digital image. An alternative way of compressing data will be presented in Chapter 11.5 where the dimension of the subspace representing a set of digital images is reduced.

Chapter 9

MATRIX PERTURBATION THEORY

We want to further understand the robustness to the Grassmann framework described in the previous Chapters. For example, if data is corrupted during acquisition or has noise (e.g., missing pixels), will this impact the distances we compute and the classification results? If possible, it would be beneficial to diminish this problem by transforming the data into a space where the intrinsic characteristics of the data are enhanced. To this end, we appeal to idea from the perturbation theory of matrices, see e.g., [79]. It was shown in Chapter 3.2 that all unitarily invariant metrics for subspaces are functions of principal angles between subspaces. Hence, we are able to focus on the theoretical perturbation results of principal angles between linear subspaces. In this chapter, we review perturbation theorems that will help to explore the robustness of the Grassmann method with an emphasis on results from Sun [81]. In particular, Theorem 9.2.3 provides a simple framework to discuss the robustness of canonical correlations between a pair of subspaces using only their matrix representations. Other perturbation theorems are presented as a reference for future exploration of the topic.

General results derived for the perturbation of principal angles can be separated into two main categories as described in Sections 9.1 and 9.2, respectively. First, the quantitative change in principal angles between a matrix A and its perturbed version $\tilde{A} = A + \Delta A$ is given in terms of several unitarily invariant norms for normal, Hermitian, diagonal, and general matrices. Secondly, changes in the set of principal angles between a pair of subspaces are also given in terms of several unitarily invariant norms. Namely, if $\theta(\mathcal{R}(A), \mathcal{R}(B))$ and $\theta(\mathcal{R}(\tilde{A}), \mathcal{R}(\tilde{B}))$ are the principal angle vectors between $\mathcal{R}(A)$ and $\mathcal{R}(B)$ and $\mathcal{R}(\tilde{A})$ and $\mathcal{R}(\tilde{B})$, respectively, then lower and upper bounds on $\triangle\theta = \theta(\mathcal{R}(A), \mathcal{R}(B)) - \theta(\mathcal{R}(\tilde{A}), \mathcal{R}(\tilde{B}))$ are available. Often we will let $\theta(A, B) = \theta(\mathcal{R}(A), \mathcal{R}(B))$ for convenience.

9.1 Perturbation Analysis of a Linear Subspace

Wedin described in [87] the connection between singular value decomposition and perturbation bounds when a matrix A is perturbed. In particular, he gave an upper bound for sine of the invariant subspaces of AA^H and A^HA (left and right singular subspaces) in terms of the magnitude of the perturbation and the singular values of the appropriate matrices. We will first develop notation and then state Wedin's main result — *The Generalized* sin θ *Theorem.*

Definition 9.1.1. The subspace $\mathcal{X}$ is an invariant subspace of A if $A\mathcal{X} \subset \mathcal{X}$.

Let $A \in \mathbb{C}^{m\times n}$ and its singular value decomposition be given by

$$A = U\Sigma V^H = U_1\Sigma_1 V_1^H + U_0\Sigma_0 V_0^H, \tag{9.1}$$

where

$$V_1 = [v_1, \ldots, v_r], V_0 = [v_{r+1}, \ldots, v_p], V = [V_1, V_0],$$

$$U_1 = [u_1, \ldots, u_r], U_0 = [u_{r+1}, \ldots, u_p], U = [U_1, U_0],$$

and

$$\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \Sigma_0 = \mathrm{diag}(\sigma_{r+1}, \ldots, \sigma_p), \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p).$$

V_1, V_0, V and U_1, U_0, U are assumed to be partial isometries satisfying

$$V^HV = U^HU = I_p, V_1^HV_1 = U_1^HU_1 = I_r, V_0^HV_0 = U_0^HU_0 = I_{p-r}.$$

The rank of A is p and $r \leq p$ is arbitrary.

Now, for the perturbation of A, $B = A + T$, a corresponding singular value decomposition can be made similarly. Take

$$A_j = U_j(A)\Sigma_j(A)V_j^H(A), \quad B_j = U_j(B)\Sigma_j(B)V_j^H(B), \quad j = 0, 1. \tag{9.2}$$

It is easy to see that $\mathcal{R}(A_1)$ and $\mathcal{R}(A_0)$ are invariant subspaces of the Hermitian matrix AA^H as are $\mathcal{R}(A_1^H)$ and $\mathcal{R}(A_0^H)$ of A^HA. Phrased differently, $\mathcal{R}(A_1)$ and $\mathcal{R}(A_0)$ are the left singular subspaces of A while $\mathcal{R}(A_1^H)$ and $\mathcal{R}(A_0^H)$ are the right singular subspaces of A. The goal of the following discussion will be to estimate the angles between the subspaces $\mathcal{R}(A_1)$ and $\mathcal{R}(B_1)$ as well as the subspaces $\mathcal{R}(A_1^H)$ and $\mathcal{R}(B_1^H)$.

If we let the distance between the subspaces $\mathcal{R}(A_1)$ and $\mathcal{R}(B_1)$ be $\rho_{p,\nu}(\mathcal{R}(A_1), \mathcal{R}(B_1))$, then by Theorem 3.2.5,

$$\rho_{p,\nu}(\mathcal{R}(A_1), \mathcal{R}(B_1)) = ||(I - P_{\mathcal{R}(B_1)})P_{\mathcal{R}(A_1)}||_\nu = ||\sin\Theta(\mathcal{R}(A_1), \mathcal{R}(B_1))||_{\nu'}$$

for some arbitrary unitarily invariant norms ν and ν'. Thus, it is then equivalent to get good upper bounds for the expressions

$$||\sin\Theta(\mathcal{R}(A_1), \mathcal{R}(B_1))|| \quad \text{and} \quad ||\sin\Theta(\mathcal{R}(A_1^H), \mathcal{R}(B_1^H))||$$

when we have estimates of $||T||$ and the gap between the least singular value of B_1 and the largest singular value of A_0.

We will now define residuals which can be used instead of T. The reason for the definition of these residuals become apparent in the proof of the generalized $\sin\theta$ theorem.

Let $y_1, \ldots, y_r$ and $x_1, \ldots, x_r$ be orthonormal vectors spanning the subspaces $\mathcal{R}(B_1)$ and $\mathcal{R}(B_1^H)$, respectively, then with $Y_1 = [y_1, \ldots, y_r]$ and $X_1 = [x_1, \ldots, x_r]$,

$$Y_1^H Y_1 = X_1^H X_1 = I_r \quad \text{and} \quad Y_1 Y_1^H = P_{\mathcal{R}(B_1)}, X_1 X_1^H = P_{\mathcal{R}(B_1^H)}.$$

Now, take $D_1 = Y_1^H B X_1$. A convenient choice from the SVD of B_1 is to let $X_1 = V_1$ and $Y_1 = U_1$. With this choice, $D_1 = \Sigma_1(B)$. Now define the residuals

$$\begin{cases} R_{11} = AX_1 - Y_1 D_1 \\ R_{21} = A^H Y_1 - X_1 D_1^H. \end{cases} \tag{9.3}$$

Then $R_{11} = -TX_1$ and $R_{21} = -T^H Y_1$. Up to this point, we have been developing the necessary notations for stating the generalized $\sin\theta$ theorem. We will now recall the $\sin\theta$ *theorem* about the perturbation of Hermitian operators by Davis and Kahan [21]. The generalized $\sin\theta$ theorem will follow rather nicely as a generalization of Davis and Kahan's theorem in terms of singular values instead of the spectrum.

Theorem 9.1.1. *[21] (The* $\sin\theta$ *theorem) Assume there is an interval* $[\beta, \alpha]$ *and a* $\delta > 0$ *such that the spectrum of* A_0 *lies entirely in* $[\beta, \alpha]$ *while that of* B_1 *lies entirely outside of* $(\beta - \delta, \alpha + \delta)$ *(or vice versa). Then for every unitarily invariant norm,* $\delta||\sin\theta(A_1, B_1)|| \leq ||R_1||$ *where* $R_1 = R_{11} = R_{21}$ *is a direct consequence of* A *and* B *being Hermitian.*

A formulation of the $\sin\theta$ theorem in terms of singular values instead of the spectrum is provided in [87].

Theorem 9.1.2. *[87],[21] Let A and B be Hermitian matrices. Assume there exists an $\alpha \geq 0$ and a $\delta > 0$ such that*

$$\sigma_{\min}(B_1) \geq \alpha + \delta \quad \textit{and} \quad \sigma_{\max}(A_0) \leq \alpha,$$

then for every unitary invariant norm,

$$\delta ||\sin\theta(A_1, B_1)|| \leq ||R_1||.$$

We are now ready to formulate a generalization of the $\sin\theta$ theorem to general $m \times n$ matrices.

Theorem 9.1.3. *[87] (The generalized $\sin\theta$ theorem) Assume there exists an $\alpha \geq 0$ and a $\delta > 0$ such that*

$$\sigma_{\min}(B_1) \geq \alpha + \delta \quad \textit{and} \quad \sigma_{\max}(A_0) < \alpha.$$

Take $\epsilon = \max\{||R_{11}||, ||R_{21}||\}$ where R_{11} and R_{21} are defined by (9.3). Then for every unitarily invariant norm,

$$\begin{cases} ||\sin\theta(A_1, B_1)|| \leq \epsilon/\delta \\ ||\sin\theta(A_1^H, B_1^H)|| \leq \epsilon/\delta. \end{cases} \tag{9.4}$$

The estimates can be sharpened to be

$$||\sin\theta(A_1, B_1)|| \leq \frac{||R_{11}||}{\delta}$$
$$||\sin\theta(A_1^H, B_1^H)|| \leq \frac{||R_{21}||}{\delta},$$

if $\dfrac{\alpha}{\alpha + \delta}$ is small.

Namely, the generalized $\sin\theta$ theorem gives an upper bound on how much change can be induced on the sine of the principal angles between the left singular subspace of A and the left singular subspace of a perturbation of A. The change in the magnitude of the principal angles is not large as long as the perturbation is kept relatively small.

9.2 Perturbation Analysis of a Pair of Linear Subspaces

The generalized $\sin\theta$ theorem only provides perturbation analysis for a single subspace with its perturbation. In this section, we will present a few general perturbation theorems that discuss perturbation analysis for a pair of subspaces with their perturbations. Perturbation theory of this type will be more related to the perturbation analysis of canonical correlations and principal angles of a pair of subspaces. First, we present two theorems that are needed for the main result:

Theorem 9.2.1. *[58] Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q$ and $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_q$ be the singular values of A and $\tilde{A} \in \mathbb{C}^{p\times q}$, respectively, $p \geq q$. Then for every unitarily invariant norm,*

$$||\operatorname{diag}(\sigma_1 - \tilde{\sigma}_1, \ldots, \sigma_q - \tilde{\sigma}_q)||_* \leq ||A - \tilde{A}||_*. \tag{9.5}$$

Theorem 9.2.2. *[80] Let Z, $W \in \mathbb{C}^{n\times m}$. If rank(W) = rank(Z), then*

$$||P_W - P_Z||_* \leq \mu \min\{||Z^\dagger||_2, ||W^\dagger||_2\}||W - Z||_*, \tag{9.6}$$

where μ is given in the following table:

$\|\quad\|_*$	arbitrary u.i.n.	Frobenius	spectral
μ	2	$\sqrt{2}$	1

Now, the main result by Sun [81] that provides perturbation analysis for the canonical correlations of a pair of matrices:

Theorem 9.2.3. *[81] Let A, $\tilde{A} \in \mathbb{C}^{n\times p}$, B, $\tilde{B} \in \mathbb{C}^{n\times q}$, $p \geq q$. Suppose that $\sigma(A, B) = \{\cos\theta_k\}_{k=1}^q = \{c_k\}_{k=1}^q$, $0 \leq \theta_1 \leq \ldots \leq \theta_q \leq \frac{\pi}{2}$, $\sigma(\tilde{A}, \tilde{B}) = \{\cos\tilde{\theta}_k\}_{k=1}^q = \{\tilde{c}_k\}_{k=1}^q$, $0 \leq \tilde{\theta}_1 \leq \ldots \leq \tilde{\theta}_q \leq \frac{\pi}{2}$. Let*

$$C = \operatorname{diag}(c_1, \ldots, c_q), \quad \tilde{C} = \operatorname{diag}(\tilde{c}_1, \ldots, \tilde{c}_q),$$

and

$$S = \operatorname{diag}(s_1, \ldots, s_q), \quad \tilde{S} = \operatorname{diag}(\tilde{s}_1, \ldots, \tilde{s}_q),$$

where $s_k = \sin\theta_k$ and $\tilde{s}_k = \sin\tilde{\theta}_k$, $k = 1, \ldots, q$. Then for every unitarily invariant norm $||\cdot||_*$,

$$||C - \tilde{C}||_*, ||S - \tilde{S}||_* \leq \delta_*(A, \tilde{A}) + \delta_*(B, \tilde{B}), \tag{9.7}$$

where

$$\delta_*(X, \tilde{X}) = \mu||X||_*||X^\dagger||_2 \cdot \frac{||X - \tilde{X}||_*}{||X||_*}$$

and μ is the same as what was given in Theorem 9.2.2.

Proof. First assume that the columns of U_A, $U_{\tilde{A}}$, U_B and $U_{\tilde{B}}$ form unitary bases for $\mathcal{R}(A)$, $\mathcal{R}(\tilde{A})$, $\mathcal{R}(B)$ and $\mathcal{R}(\tilde{B})$, respectively. By the hypotheses, we have

$$\sigma(U_A^H U_B) = \{c_k\}_{k=1}^q, \quad \sigma(U_{\tilde{A}}^H U_{\tilde{B}}) = \{\tilde{c}_k\}_{k=1}^q.$$

Let W_A and $W_{\tilde{A}}$ be such that (U_A, W_A) and $(U_{\tilde{A}}, W_{\tilde{A}})$ are $n \times n$ unitary matrices. Then from Lemma 3.2.1

$$\sigma(W_A^H U_B) = \{s_k\}_{k=1}^q, \quad \sigma(W_{\tilde{A}}^H U_{\tilde{B}}) = \{\tilde{s}_k\}_{k=1}^q.$$

By Lemma 3.2.2 and Theorem 9.2.1 we get

$$||C - \tilde{C}|| \leq ||P_A P_B - P_{\tilde{A}} P_{\tilde{B}}|| \leq ||P_A - P_{\tilde{A}}|| + ||P_B - P_{\tilde{B}}|| \tag{9.8}$$

and

$$||S - \tilde{S}|| \leq ||(I - P_A)P_B - (I - P_{\tilde{A}})P_{\tilde{B}}|| \leq ||P_A - P_{\tilde{A}}|| + ||P_B - P_{\tilde{B}}||. \tag{9.9}$$

We then apply Theorem 9.2.2 to Equations (9.8) and (9.9) to get the desired result. □

Observe that if

$$|\cos\theta - \cos\tilde{\theta}| \leq h, \quad |\sin\theta - \sin\tilde{\theta}| \leq h, \quad \theta, \tilde{\theta} \in [0, \frac{\pi}{2}],$$

then

$$|\theta - \tilde{\theta}| \leq \frac{\pi}{2} h.$$

Hence by Theorem 9.2.3, we can deduce the following result.

Corollary 9.2.4. *[81] Assuming the hypotheses of Theorem 9.2.3, then we have*

$$\sqrt{\sum_{k=1}^{q} (\theta_k - \tilde{\theta}_k)^2} \leq \frac{\pi}{2} \left(\delta_F(A, \tilde{A}) + \delta_F(B, \tilde{B}) \right) \tag{9.10}$$

and

$$|\theta_k - \tilde{\theta}_k| \leq \frac{\pi}{2} \left(\delta_2(A, \tilde{A}) + \delta_2(B, \tilde{B}) \right), \quad \forall k = 1, \ldots, q. \tag{9.11}$$

In the meanwhile, [37] offers improved bounds given in [81] without restricting the dimension of the subspaces to be equal. However, the bounds are given on the canonical correlations which are the cosine of the principal angles instead of the principal angles themselves.

Theorem 9.2.5. *[37] Let rank(A) = rank($\tilde{A}$) = p, and rank(B) = rank($\tilde{B}$) = q. For any unitarily invariant norm, define the condition numbers of A and B to be*

$$\kappa(A, ||\cdot||) = ||A|| \cdot ||A^\dagger||_2, \quad \kappa(B, ||\cdot||) = ||B|| \cdot ||B^\dagger||_2.$$

If $\triangle \cos\Theta = \cos\Theta(A,B) - \cos\Theta(\tilde{A},\tilde{B})$ *and* $\triangle \sin\Theta = \sin\Theta(A,B) - \sin\Theta(\tilde{A},\tilde{B})$*, then*

$$|| \triangle \cos\Theta|| \leq \mu \left\{ \kappa(A, ||\cdot||)\cos\theta_1 \frac{||A-\tilde{A}||}{||A||} + \kappa(B, ||\cdot||)\cos\phi_1 \frac{||B-\tilde{B}||}{||B||} \right\}$$

and

$$|| \triangle \sin\Theta|| \leq \mu \left\{ \kappa(A, ||\cdot||)\cos\theta_2 \frac{||A-\tilde{A}||}{||A||} + \kappa(B, ||\cdot||)\cos\phi_2 \frac{||B-\tilde{B}||}{||B||} \right\}$$

with

$$\begin{aligned} \theta_1 &= \theta_{\min}(\mathcal{C}(A,\tilde{A}), \mathcal{R}(B)), \quad \theta_2 = \theta_{\min}(\mathcal{C}(A,\tilde{A}), \mathcal{R}(B)^{\perp}), \\ \phi_1 &= \theta_{\min}(\mathcal{C}(B,\tilde{B}), \mathcal{R}(\tilde{A})), \quad \phi_2 = \theta_{\min}(\mathcal{C}(B,\tilde{B}), \mathcal{R}(\tilde{A})^{\perp}), \end{aligned}$$

where $\mathcal{C}(A,\tilde{A})$ *is the orthogonal complement of* $\mathcal{R}(A) \cap \mathcal{R}(\tilde{A})$ *in* $\mathcal{R}(A) + \mathcal{R}(\tilde{A})$*, and* $\mathcal{C}(B,\tilde{B})$ *is the orthogonal complement of* $\mathcal{R}(B) \cap \mathcal{R}(\tilde{B})$ *in* $\mathcal{R}(B) + \mathcal{R}(\tilde{B})$*. Moreover, for the spectral norm* $\mu = 1$*, while for any arbitrary invariant norm* $\mu = \sqrt{2}$*.*

An important thing to realize from Theorem 9.2.5 that is not found in Theorem 9.2.3 is that perturbations of the canonical correlations (cosine of the principal angles) depend on the matrix pair (A,B) instead of A and B as two individual matrices. For example, if we perturb A and B to the effect that $\mathcal{R}(\tilde{A})$ and $\mathcal{R}(\tilde{B})$ remain the same as $\mathcal{R}(A)$ and $\mathcal{R}(B)$, respectively, then the canonical correlations are unchanged [37].

However, the perturbation bounds given in Theorem 9.2.5 are not invariant under the column scaling of A and B, i.e., perturbation bounds for A and B are not necessarily the same as the ones for AD_1 and BD_2 where D_1 and D_2 are some positive definite diagonal matrices. A way to get around it is to consider the following theorem for the spectral norm:

Theorem 9.2.6. *[37] Let* $A \in \mathcal{M}_{m,p}$ *and* $B \in \mathcal{M}_{m,q}$ *be of full column rank, and let* $\tilde{A} = A + \triangle A$ *and* $\tilde{B} = B + \triangle B$ *with* $|\triangle A| \leq \epsilon G_A |A|$ *and* $|\triangle B| \leq \epsilon G_B |B|$ *be such that* $\tilde{A}$ *and* $\tilde{B}$ *are also of full column rank, where* ϵ *is small and* G_A *and* G_B *are matrices with nonnegative elements. Then*

$$\begin{aligned} || \triangle \cos\Theta||_2 \leq\ & \epsilon \Big[\sqrt{p(m-p)}||G_A||_2 \cos\theta_1 \kappa_s(A) \\ & + \sqrt{q(m-q)}||G_B||_2 \cos\phi_1 \kappa_s(B) \Big]. \end{aligned}$$

and

$$\begin{aligned} || \triangle \sin\Theta||_2 &\leq \epsilon \left[\sqrt{p(m-p)}||G_A||_2 \cos\theta_2 \kappa_s(A) \right. \\ &\quad \left. + \sqrt{q(m-q)}||G_B||_2 \cos\phi_2 \kappa_s(B) \right], \end{aligned}$$

where θ_i, ϕ_i, $i = 1, 2$, *are defined in Theorem 9.2.5. Moreover,* $\kappa_s(A) = |||R||R^{-1}|||_2$ *if the QR decomposition of A is* $A = QR$. *Obviously,* $\kappa_s(A)$ *is independent of column scaling of A since* $\kappa_s(AD) = \kappa_s(A)$ *for any positive definite diagonal matrix D.*

When the matrices are ill-conditioned, the perturbation bounds are terrible. Since column scaling does not change the canonical correlations, so in general, one would want to adjust the condition number of a ill-conditioned matrix. All of these perturbation theorems give "absolute" perturbation bounds. Relative perturbation bounds exist only if no "cancelation" occurs (see [37]).

Certainly, there are more perturbation theorems available. For the applications we are interested in, Theorem 9.2.3 will be used as an initial attempt to solve the optimization problem formulated in Chapter 10.

Chapter 10

PERTURBATION THEORY FOR GRASSMANN SEPARABILITY

In this chapter, we adapt our normal notation to let an orthogonal matrix $S \in \mathbb{R}^{n\times k}$ represent a point on the Grassmannian, $G(k,n)$. Furthermore, let $\tilde{S} = S + \Delta S$ be a perturbation of S. Given points on the Grassmann manifold that are Grassmann separable, it is curious to see how much we can perturb those points before the separation ceases to exist. We will illustrate with a simple example how perturbation theory discussed in Chapter 9.2 can be used to derive bounds for $||\Delta S||$ so that whenever a point on the Grassmannian is perturbed by no more than such bounds, perfect Grassmann separation is guaranteed. Once this is established in Chapter 10.1, it will be clear that a data set is more likely to be Grassmann separable if its perturbation bound is large. To improve the robustness of the separability condition, we formulate an optimization problem in Chapter 10.2 with the objective function motivated by matrix perturbation theory. In doing so, we are able to arrive at the objective function given in [50] that is motivated by *Linear Discriminant Analysis* as a special case of our final objective function using perturbation theory. And we formulated the problem from totally different framework than the one given in [50]. In Chapter 10.3, numerical algorithms and solutions using the *Steepest Descent* and *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* optimization methods are derived for the low resolution data set mentioned in Chapter 7.3. The separation gap for this particular data set was improved after data points are transformed using numerical solutions of the optimization problem. These results show initial promise although the improvements might be modest. It will require further investigation to establish just how effective the objective function might be.

	Target			
	d	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$
Query	$\tilde{Q}^{(1)}$	$\tilde{d}_{11}$	$\tilde{d}_{12}$	$\tilde{d}_{13}$
	$Q^{(2)}$	d_{21}	d_{22}	d_{23}
	$Q^{(3)}$	d_{31}	d_{32}	d_{33}

Table 10.1: Distance matrix between the target points and the query points under a one-sided perturbation.

10.1 Framework for Grassmann Separability

Here we replace the conventional terms *probe* and *gallery* by *target* and *query* since the purpose of this section is not to classify unknown identities, but to examine a data set's Grassmann separability. We assume that we know the labels of the points in both the target and query permitting us to examine the robustness of the Grassmann method.

Consider three subjects with dimension-2 complete subspace configurations: $C^{(1)} = \left\{T^{(1)}, Q^{(1)}\right\}$, $C^{(2)} = \left\{T^{(2)}, Q^{(2)}\right\}$, and $C^{(3)} = \left\{T^{(3)}, Q^{(3)}\right\}$, where $T^{(i)}$ and $Q^{(i)}$ denote target and query points, respectively.

Suppose that the data set consisting of these three subjects is Grassmann separable in dimension k with the ℓ-truncated Grassmannian semi-distance d. Let d_{ij} be the distance matrix of the pairwise distances between each pair of target and query points. The fact that this data set is Grassmann separable implies that the separation gap $g_s = \min\limits_{i \neq j} \{d_{ij}\} - \max\limits_{i=j} \{d_{ij}\}$ is greater than zero.

Now, we want to investigate the maximum amount of perturbation that can be applied to points on the Grassmannian in some unitarily invariant norm $||\cdot||_*$ before perfect Grassmann separability breaks. First, consider one-sided perturbation applied only to the query points. To simplify the matter even more, imagine that we accidentally acquired a noisy version of subject one, i.e., replace $Q^{(1)}$ by $\tilde{Q}^{(1)} = Q^{(1)} + \Delta Q^{(1)}$. The noise could come from the corruption of the image or operator error in collecting the image. Then notice that the new distance matrix, if we were to compute one, will be of the form in Table 10.1, where only the distances in the first row are effected. This motivates us to examine the relationships between the perturbed query point and the target points only.

Proposition 10.1.1. (One-sided Perturbation Bound)

Suppose $\mathcal{P} = \left\{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\right\}$ *is a set of subjects with dimension-2 complete subspace configuration for each* $C^{(i)}$*, i.e.,* $C^{(i)} = \left\{T^{(i)}, Q^{(i)}\right\}$*. Let* $D_{ij} = d\left(Q^{(i)}, T^{(j)}\right)$ *be the pairwise distance between each target and query point for some* ℓ*-truncated Grassmannian semi-distance,* d*.*

Set

$$M = \max_{1\leq i\leq N} d\left(T^{(i)}, Q^{(i)}\right),\ m = \min_{1\leq i\neq j\leq N} d\left(Q^{(i)}, T^{(j)}\right),\ g_s = m - M,$$

and let $\tilde{D} = \left(\tilde{d}_{ij}\right)_{i,j=1}^{N}$ *be the distance matrix computed from replacing* $Q^{(k)}$ *with* $\tilde{Q}^{(k)} = Q^{(k)} + \Delta Q^{(k)}$*, for some* k*. Then the perturbed data set is Grassmann separable if*

$$g_s > 0, \text{ and } \left|\tilde{d}_{kj} - d_{kj}\right| < \frac{g_s}{2}, \text{ for all } j = 1, 2, \ldots, N.$$

Proof. Notice that

$$\tilde{D} = \begin{pmatrix} & D(1:k-1,:) & \\ \tilde{d}_{k1} & \ldots & \tilde{d}_{kN} \\ & D(k+1:N,:) & \end{pmatrix}$$

The only interesting case is when $\tilde{d}_{kk} > d_{kk}$ and $\tilde{d}_{kj} < d_{kj}$ for all $j \neq k$. By the assumptions, we have for all k and $j \neq k$

$$\begin{aligned} \tilde{d}_{kk} - d_{kk} < \frac{g_s}{2} \quad &\Rightarrow \quad \tilde{d}_{kk} < d_{kk} + \frac{g_s}{2} < M + \frac{g_s}{2}, \\ d_{kj} - \tilde{d}_{kj} < \frac{g_s}{2} \quad &\Rightarrow \quad \tilde{d}_{kj} > d_{kj} - \frac{g_s}{2} > m - \frac{g_s}{2}. \end{aligned}$$

Therefore,

$$\tilde{M} := \max_{i=j} \tilde{d}_{ij} < M + \frac{g_s}{2}$$

and

$$\tilde{m} := \min_{i\neq j} \tilde{d}_{ij} > m - \frac{g_s}{2}.$$

Now the new separation gap,

$$\tilde{g}_s = \tilde{m} - \tilde{M} > \left(m - \frac{g_s}{2}\right) - \left(M + \frac{g_s}{2}\right) = m - M - g_s = 0.$$

Hence the perturbed data set is Grassmann separable. □

Notice that less tight bounds than the one given in Proposition 10.1.1 can be found to improve the approximation. Before utilizing Proposition 10.1.1, we consider Theorem 9.2.3 by Sun [81] in a way illustrated in the following. For example, for the pair of target and query points of subject one, we have

$$\begin{aligned}\left\|\sin\Theta\left(\tilde{Q}^{(1)},T^{(1)}\right)-\sin\Theta\left(Q^{(1)},T^{(1)}\right)\right\|_F &\leq \sqrt{2}\left\|Q^{(1)}\right\|_F\left\|\left(Q^{(1)}\right)^{\dagger}\right\|_2\frac{\left\|Q^{(1)}-\tilde{Q}^{(1)}\right\|_F}{\left\|Q^{(1)}\right\|_F}\\ &+ \sqrt{2}\left\|T^{(1)}\right\|_F\left\|\left(T^{(1)}\right)^{\dagger}\right\|_2\frac{\left\|T^{(1)}-\tilde{T}^{(1)}\right\|_F}{\left\|T^{(1)}\right\|_F}.\end{aligned}$$

But in our problem $T^{(1)}=\tilde{T}^{(1)}$, so the second expression on the right vanishes. Observe from this inequality, the expression on the right is similar for the target points $T^{(1)}$, $T^{(2)}$, and $T^{(3)}$. Thus, if we let T denote one of those three target points, then

$$\left\|\sin\Theta\left(\tilde{Q}^{(1)},T\right)-\sin\Theta\left(Q^{(1)},T\right)\right\|_F\leq\sqrt{2}\left\|\left(Q^{(1)}\right)^{\dagger}\right\|_2\left\|\Delta Q^{(1)}\right\|_F.$$

Now, let d be the Projection F-norm. In order to maintain Grassmann separability we need the condition $\left|d\left(\tilde{Q}^{(1)},T\right)-d\left(Q^{(1)},T\right)\right|<g_s/2$ to hold by Proposition 10.1.1. Thus

$$\begin{aligned}\left|d\left(\tilde{Q}^{(1)},T\right)-d\left(Q^{(1)},T\right)\right| &= \left|\left\|\sin\Theta\left(\tilde{Q}^{(1)},T\right)\right\|_F-\left\|\sin\Theta\left(Q^{(1)},T\right)\right\|_F\right|\\ &\leq \left\|\sin\Theta\left(\tilde{Q}^{(1)},T\right)-\sin\Theta\left(Q^{(1)},T\right)\right\|_F\\ &\leq \sqrt{2}\left\|\left(Q^{(1)}\right)^{\dagger}\right\|_2\left\|\Delta Q^{(1)}\right\|_F<\frac{g_s}{2},\end{aligned}$$

which leads to

$$\left\|\Delta Q^{(1)}\right\|_F<\frac{g_s}{2\sqrt{2}\left\|\left(Q^{(1)}\right)^{\dagger}\right\|_2}. \tag{10.1}$$

Similar bounds can be derived for subjects two and three if their query points were subject to perturbation:

$$\left\|\Delta Q^{(2)}\right\|_F<\frac{g_s}{2\sqrt{2}\left\|\left(Q^{(2)}\right)^{\dagger}\right\|_2},\quad \left\|\Delta Q^{(3)}\right\|_F<\frac{g_s}{2\sqrt{2}\left\|\left(Q^{(3)}\right)^{\dagger}\right\|_2}.$$

It is clear that the perturbation bounds derived in such a way depends only on the spectral norm of the pseudo-inverse of the query points $Q^{(1)}$, $Q^{(2)}$, or $Q^{(3)}$. To simplify notations, we let Q denote some general query point and ΔQ its perturbation in the following discussions.

Proposition 10.1.2. *Let d be the Projection F-norm (chordal), Q any query point, T any target point, and ΔQ the amount of perturbation applied to Q, then*

$$\|\Delta Q\|_F < \frac{g_s}{2\sqrt{2}\,\|Q^\dagger\|_2}.$$

Proof. See discussions above. □

Similarly,

Proposition 10.1.3. *Let d be the arc length (geodesic), Q any query point, T any target point, and ΔQ the amount of perturbation applied to Q, then*

$$\|\Delta Q\|_F < \frac{g_s}{\pi\,\|Q^\dagger\|_2}.$$

Proof.

$$\begin{aligned}\left|d\left(\tilde{Q},T\right)-d\left(Q,T\right)\right| &= \left|\left\|\Theta\left(\tilde{Q},T\right)\right\|_F-\left\|\Theta\left(Q,T\right)\right\|_F\right| \\ &\leq \left\|\Theta\left(\tilde{Q},T\right)-\Theta\left(Q,T\right)\right\|_F \\ &\leq \frac{\pi}{2}\left\|Q^\dagger\right\|_2\left\|\Delta Q\right\|_F<\frac{g_s}{2}.\end{aligned}$$

Thus,

$$\|\Delta Q\|_F < \frac{g_s}{\pi\,\|Q^\dagger\|_2}. \tag{10.2}$$

□

Notice that this is a tighter bound than the one given in Equation (10.1). Moreover,

Proposition 10.1.4. *Let d be the Projection 2-norm, Q any query point, T any target point, and ΔQ the amount of perturbation applied to Q, then*

$$\|\Delta Q\|_2 < \frac{g_s}{2\,\|Q^\dagger\|_2}.$$

Proof.

$$\begin{aligned}\left|d\left(\tilde{Q},T\right)-d\left(Q,T\right)\right| &= \left|\left\|\sin\Theta\left(\tilde{Q},T\right)\right\|_2-\left\|\sin\Theta(Q,T)\right\|_2\right| \\ &\leq \left\|\sin\Theta\left(\tilde{Q},T\right)-\sin\Theta\left(Q,T\right)\right\|_2 \\ &\leq \left\|Q^\dagger\right\|_2\left\|\Delta Q\right\|_2<\frac{g_s}{2}.\end{aligned}$$

Thus,

$$\|\Delta Q\|_2 < \frac{g_s}{2\,\|Q^\dagger\|_2}. \tag{10.3}$$

□

A summary of the perturbation bounds obtained from considering various Grassmannian metrics in the perturbation Theorem 9.2.3 is given in Table 10.2. In particular, since the spectral norm of any orthogonal matrix is identically one, the perturbation bounds simplifies for points on the Grassmannians.

	P. Norm	**Matrix P. B.**	**P. B. for Points on** $G(k,n)$
Geodesic	$\|\Delta Q\|_F$	$\frac{g_s}{\pi \|Q^\dagger\|_2}$	$\frac{g_s}{\pi}$
Chordal	$\|\Delta Q\|_F$	$\frac{g_s}{2\sqrt{2} \|Q^\dagger\|_2}$	$\frac{g_s}{2\sqrt{2}}$
Projection 2-norm	$\|\Delta Q\|_2$	$\frac{g_s}{2 \|Q^\dagger\|_2}$	$\frac{g_s}{2}$

Table 10.2: A summary of various perturbation bounds obtained from considering various Grassmannian metrics in the perturbation Theorem 9.2.3 where g_s denote the separation gap.

In general, when given a data set of N subjects with the usual notations for target and query points, we have the following corollary:

Corollary 10.1.1. *For any query point Q with $\tilde{Q} = Q + \Delta Q$,*

$$\|\Delta Q\|_* \leq \frac{c(d) \cdot g_s(d)}{q}, \tag{10.4}$$

where

$$q = \min_{1 \leq i \leq N} \left\| \left(Q^{(i)} \right)^\dagger \right\|_2 .$$

The values of $c(d)$ are given in the following table:

d	$\|\| \cdot \|\|$	$c(d)$
Geodesic	$\|\| \cdot \|\|_F$	$\frac{1}{\pi}$
Chordal	$\|\| \cdot \|\|_F$	$\frac{1}{2\sqrt{2}}$
Projection 2-norm	$\|\| \cdot \|\|_2$	$\frac{1}{2}$

For points on the Grassmannians, the perturbation bounds simplifies to:

$$\|\Delta Q\|_* \leq c \cdot g_s. \tag{10.5}$$

On the other hand, if we consider two-sided perturbation where both target and query points of a single subject are perturbed, then a similar result to that of Proposition 10.1.1 can be obtained.

Proposition 10.1.5. (Two-sided Perturbation Bound)

Suppose $\mathcal{P} = \left\{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\right\}$ *is a set of subjects with dimension-2 complete subspace configuration for each* $C^{(i)}$*, i.e.,* $C^{(i)} = \left\{T^{(i)}, Q^{(i)}\right\}$*. Let* $D_{ij} = d\left(Q^{(i)}, T^{(j)}\right)$ *be the pairwise distance between each target and query point for some* ℓ*-truncated Grassmannian semi-distance,* d*.*

Set

$$M = \max_{1\leq i\leq N} d\left(Q^{(i)}, T^{(i)}\right),\ m = \min_{1\leq i\neq j\leq N} d\left(Q^{(i)}, T^{(j)}\right),\ g_s = m - M,$$

and let $\tilde{D} = \left(\tilde{d}_{ij}\right)_{i,j=1}^{N}$ *be the distance matrix computed from replacing* $T^{(k)}$ *with* $\tilde{T}^{(k)} = T^{(k)} + \Delta T^{(k)}$*, for some* k *and* $Q^{(r)}$ *with* $\tilde{Q}^{(r)} = Q^{(r)} + \Delta Q^{(r)}$*, for some* r*. Then the perturbed data set is Grassmann separable if*

$$g_s > 0,\ \textit{and}\ \left|\tilde{d}_{kj} - d_{kj}\right| < \frac{g_s}{2},\ \textit{and}\ \left|\tilde{d}_{ir} - d_{ir}\right| < \frac{g_s}{2}\ \textit{for all}\ i, j = 1, 2, \ldots, N.$$

Proof. The proof is similar to the proof in Proposition 10.1.1, it is therefore omitted here. □

Now, suppose that the entire data set is subject to perturbation.

Proposition 10.1.6. *Given a data set of* N *subjects with the usual notations for target and query points and let* d *be the one of the* ℓ*-truncated Grassmannian semi-distances. The perturbed data set having target and query points in the form* $\tilde{T} = T + \Delta T$ *and* $\tilde{Q} = Q + \Delta Q$ *satisfy the following:*

$$\|\Delta S\|_* < \frac{c(d) \cdot g_s(d)}{\alpha},$$

where

$$\|\Delta S\|_* = \max_{\substack{1\leq i\leq N\\ 1\leq j\leq N}} \left\{\left\|\Delta Q^{(i)}\right\|_*, \left\|\Delta T^{(j)}\right\|_*\right\}$$

and

$$\alpha = \min_{\substack{1\leq i\leq N\\ 1\leq j\leq N}} \left\{\left\|\left(Q^{(i)}\right)^{\dagger}\right\|_2 + \left\|\left(T^{(j)}\right)^{\dagger}\right\|_2\right\}.$$

The values of $c(d)$ *are given in the following table:*

d	$\|\|\cdot\|\|$	$c(d)$
Geodesic	$\|\|\cdot\|\|_F$	$\frac{2}{\pi}$
Chordal	$\|\|\cdot\|\|_F$	$\frac{1}{\sqrt{2}}$
Projection 2-norm	$\|\|\cdot\|\|_2$	1

Proof. For any pair of target and query points T and Q,

$$\begin{aligned}\left|d\left(\tilde{Q},\tilde{T}\right)-d(Q,T)\right| &\leq c_1\left\|Q\right\|_*\left\|Q^\dagger\right\|_2\frac{\left\|Q-\tilde{Q}\right\|_*}{\left\|Q\right\|_*}+c_1\left\|T\right\|_*\left\|T^\dagger\right\|_2\frac{\left\|T-\tilde{T}\right\|_*}{\left\|T\right\|_*}\\ &= c_1\left(\left\|Q^\dagger\right\|_2\left\|\Delta Q\right\|_*+\left\|T^\dagger\right\|_2\left\|\Delta T\right\|_*\right)\\ &\leq c_1\left(\left\|Q^\dagger\right\|_2\left\|\Delta S\right\|_*+\left\|T^\dagger\right\|_2\left\|\Delta S\right\|_*\right)\end{aligned}$$

By Proposition 10.1.5, we have

$$\left\|\Delta S\right\|_* < \frac{g_s}{2c_1\left(\left\|Q^\dagger\right\|_2+\left\|T^\dagger\right\|_2\right)}.$$

But

$$\left\|Q^\dagger\right\|_2+\left\|T^\dagger\right\|_2\geq\alpha,$$

hence the result with the appropriate value of $c(d)$. □

In particular, if $T^{(i)}$ and $Q^{(i)}$ are points on the Grassmannian, then

$$\left\|\Delta S\right\|_* < \frac{c(d)\cdot g_s(d)}{2}.$$

That is, if we do not perturb any point by more than $\frac{c\cdot g_s}{2}$, then Grassmann separability of the new data set is guaranteed.

In general, when subjects do not have consistent subspace configurations, we assume that a total of m data points are given as $\left\{S^{(1)},S^{(2)},\ldots,S^{(m)}\right\}$. Each point belongs to one of the subject class denoted by $C^{(i)}$. Define $W_i=\left\{j\mid S^{(j)}\in C^{(i)}\right\}$, the within-class set of subject i, and $B_i=\left\{j\mid S^{(j)}\notin C^{(i)}\right\}$, the between-class set of subject i. Then Proposition 10.1.6 can be extended to the general case.

Corollary 10.1.2. *Given a data set of m points given as $\left\{S^{(1)},S^{(2)},\ldots,S^{(m)}\right\}$ with each point belonging to one of the subject class $C^{(i)}$. Let $\tilde{S}^{(i)}=S^{(i)}+\Delta S^{(i)}$ be a perturbation of $S^{(i)}$ and d be one of the ℓ-truncated Grassmannian semi-distances. Further denote $D_{ij}=d(S^{(i)},S^{(j)})$. Then*

$$\left\|\Delta S\right\|_* < \frac{c\cdot g_s}{\alpha},$$

where

$$g_s = \min_{1\le i\le m} \min_{j\in B_i} D_{ij} - \max_{1\le k\le m} \max_{l\in W_k} D_{kl},$$

$$\|\Delta S\|_* = \max_{1\le i\le m} \left\{ \left\| \Delta S^{(i)} \right\|_* \right\},$$

and

$$\alpha = \min_{\substack{1\le i\le m \\ 1\le j\le m}} \left\{ \left\| \left(S^{(i)} \right)^\dagger \right\|_2 + \left\| \left(S^{(j)} \right)^\dagger \right\|_2 \right\}.$$

The values of c are given in the table above.

d	$\|\|\cdot\|\|$	$c(d)$
Geodesic	$\|\|\cdot\|\|_F$	$\frac{2}{\pi}$
Chordal	$\|\|\cdot\|\|_F$	$\frac{1}{\sqrt{2}}$
Projection 2-norm	$\|\|\cdot\|\|_2$	1

In particular, if $S^{(i)}$'s are points on the Grassmann manifold, then

$$\|\Delta S\|_* < \frac{c}{2} \left(\min_{1\le i\le m} \min_{j\in W_i} D_{ij} - \max_{1\le k\le m} \max_{l\in B_k} D_{kl} \right). \tag{10.6}$$

In summary, we have found perturbation bounds based on a perturbation theorem given by Sun [81] that characterize Grassmann separability of data sets. In the next section, we will use these quantities (perturbation bounds) to motivate an optimization problem so that the solutions of the optimization problem will improve Grassmann separability of data sets.

10.2 Derivation of Grassmann Potential

The perturbation bounds derived in the previous section serve as indicators on how likely data sets are to be Grassmann separable. In particular, the bigger the perturbation bound is for a particular data set, the more likely the data set is Grassmann separable and more tolerant to noise. Therefore, in an attempt to simultaneously increase the robustness of the Grassmann method and improve data sets' ability to be Grassmann separable, we will search for maps that transform the data into a space where the Grassmann separability is optimized. Notice that it is sufficient to discuss distances between points on $G(k,n)$ by discussing distances between the n-by-k matrices used to represent these points on $G(k,n)$. This is because if P_i, P_j are orthonormal basis matrices

for the image sets $X_i, X_j \in \mathbb{R}^{n\times k}$, respectively, then for any Grassmannian distance (or Grassmann semi-distance) D, $D(\mathcal{R}(P_i), \mathcal{R}(P_j)) = D(\mathcal{R}(X_i), \mathcal{R}(X_j))$. Since the measures that we use here are the Grassmannian distances (or Grassmannian semi-distances), we will refer to the perturbation bound derived in the previous chapter as the *Grassmann potential* and use general image sets for the following discussions.

Assume that we are given a set of m matrices $\{X_1, X_2, \ldots, X_m\}$ each belonging to $\mathbb{R}^{n\times k}$. Moreover, each set belongs to one of the subject classes denoted by C_i. Let $D_{ij} = D(X_i, X_j)$ be the pairwise distance between the i^{th} and j^{th} subjects and define

$$g_s = \min_{1\le i\le m} \min_{j\in B_i} D_{ij} - \max_{1\le k\le m} \max_{l\in W_k} D_{kl} \quad (\textbf{separation gap}).$$

Furthermore, define a transformation matrix L such that $L \in \mathbb{R}^{n\times d}$ and $L : X_i \to Y_i = L^T X_i$, where $d \le n$. The matrix L will behave like a feature extractor that transforms general representations into a collection of optimal ones to ensure the distance between any two matching points is smaller than the distance between any two non-matching points using Grassmannian distances. Under this framework, the perturbation bound obtained from Corollary 10.1.2 gives us an immediate candidate objective function for optimizing the Grassmann potential. i.e., we search for a $L^* \in \mathbb{R}^{n\times d}$ such that

$$L^* = \arg\max_{L\in\mathbb{R}^{n\times d}} \frac{\min\limits_{1\le i\le m}\min\limits_{j\in B_i} D(L^T X_i, L^T X_j) - \max\limits_{1\le k\le m}\max\limits_{l\in W_k} D(L^T X_i, L^T X_j)}{\min\limits_{\substack{1\le i\le m\\ 1\le j\le m}} \left\{ \left\| L^T X_i^\dagger \right\|_2 + \left\| L^T X_j^\dagger \right\|_2 \right\}}. \tag{10.7}$$

This expression is sensitive to outliers and noise. For example, a pair of twins in a data set will cause the minimum of the between-class distances to be very small thus making the expression to be ineffective. Thus, we consider solving the optimization problem using the ensemble distances instead:

$$L^* = \arg\max_{L\in\mathbb{R}^{n\times d}} \frac{\sum\limits_{i=1}^{m}\sum\limits_{j\in B_i} D(L^T X_i, L^T X_j) - \sum\limits_{i=1}^{m}\sum\limits_{k\in W_i} D(L^T X_i, L^T X_k)}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} \left\| L^T X_i^\dagger \right\|_2 + \left\| L^T X_j^\dagger \right\|_2}. \tag{10.8}$$

It turns out that we can also obtain this objective function from Theorem 9.2.3 and Proposition 10.1.5. Let $Y_i = L^T X_i$. Specifically, by Theorem 9.2.3,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \left| D(\tilde{Y}_i, \tilde{Y}_j) - D(Y_i, Y_j) \right| \le c \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\left\| Y_i^\dagger \right\|_2 \|\Delta Y_i\|_* + \left\| Y_j^\dagger \right\|_2 \|\Delta Y_j\|_* \right).$$

Let $\|\Delta Y\|_* = \max_{1\leq i\leq m} \|\Delta Y_i\|_*$, then

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\left|D(\tilde{Y}_i,\tilde{Y}_j) - D(Y_i,Y_j)\right| \leq c\,\|\Delta Y\|_* \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|Y_i^\dagger\right\|_2 + \left\|Y_j^\dagger\right\|_2\right). \tag{10.9}$$

By Proposition 10.1.5, Expression (10.9) is bounded above by $\frac{m^2 g_s}{2}$. Thus,

$$\begin{aligned} 2c\cdot\|\Delta Y\|_* &< \frac{m^2 g_s}{\beta} \\ &= m^2\cdot\frac{\min\limits_{1\leq i\leq m}\min\limits_{j\in B_i} D_{ij} - \max\limits_{1\leq i\leq m}\max\limits_{j\in W_i} D_{ij}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\left\|Y_i^\dagger\right\|_2 + \left\|Y_j^\dagger\right\|_2}. \end{aligned} \tag{10.10}$$

Expression (10.10) is then less than

$$\frac{\sum\limits_{i=1}^{m}\sum\limits_{j\in B_i} D_{ij} - \sum\limits_{i=1}^{m}\sum\limits_{j\in W_i} D_{ij}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\left\|Y_i^\dagger\right\|_2 + \left\|Y_j^\dagger\right\|_2} := P. \tag{10.11}$$

Now, in order to optimize Grassmann potential (or separability), we would like the numerator of P to be as bigger than zero as possible. But that is the same as to have

$$\frac{\sum\limits_{i=1}^{m}\sum\limits_{j\in B_i} D_{ij}}{\sum\limits_{i=1}^{m}\sum\limits_{j\in W_i} D_{ij}}$$

as bigger than one as possible, since

$$a-b>0 \quad\Rightarrow\quad \frac{a}{b}-\frac{b}{b}>0 \quad\Rightarrow\quad \frac{a}{b}>1$$

for any $a, b\in\mathbb{R}$, $b\neq 0$. Further notice that in order to optimize Expression (10.11), we want the denominator to be as small as possible. Thus, it is as effective to consider the following objective function

$$E(L) = \frac{\sum\limits_{i=1}^{m}\sum\limits_{j\in B_i} D(Y_i,\, Y_j)}{\sum\limits_{k=1}^{m}\sum\limits_{l\in W_k} D(Y_k,\, Y_l)\cdot\left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\left\|(Y_i)^\dagger\right\|_2 + \left\|(Y_j)^\dagger\right\|_2\right)} \tag{10.12}$$

in the optimization problem

$$L^* = \arg\max_{L\in\mathbb{R}^{n\times d}} E(L). \tag{10.13}$$

We remark that solving this optimization problem depends on the choice of the metric. A similar optimization problem was proposed in [50] where a linear transformation $L^* \in GL_n$ is sought such that

$$L^* = \arg\max_{L} \frac{\sum_{i=1}^{m} \sum_{k \in W_i} F_{ik}}{\sum_{i=1}^{m} \sum_{l \in B_i} F_{il}}, \tag{10.14}$$

where F_{ij} is the similarity between two transformed data sets Y_i and Y_j. It is given by the sum of the canonical correlations between Y_i and Y_j. To distinguish the difference between this objective function and our proposed objective function, notice that we let D be a Grassmannian distance and have an extra term β in the objective function in this context. Besides, the cost function in our optimization problem is motivated from the study of matrix perturbation theory while the cost function used in [50] is inspired by the classical *Linear Discriminant Analysis*. They refer to their method as Discriminant Canonical Correlation (DCC).

10.3 Numerical Solutions to Solving Grassmann Potential

As an example and initial step towards solving the optimization problem given in Expression (10.13), we consider the 1-truncated geodesic distance and two numerical optimization techniques, *Steepest Descent* and *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* methods. It is worth noting that in the special case where L is a rotation matrix and image sets X_i's are represented by their orthonormal basis vectors, $\left\| \left(L^T X_i\right)^\dagger \right\|_2 = 1$ for all i. If we also note that the *similarity* between two image sets is the inverse of the *distance* between the sets, then Equation (10.12) is equivalent to Equation (10.14) up to constant multiples and the choice of metric. However, depending on the actual metric implemented, the result of the optimization problems will generally output different expressions of L.

The optimization problem in Equation (10.14) involves three variables. As the two other variables are not explicitly described by L, Kim et al. proposed an iterative optimization algorithm that computes an optimal solution for one of the three variables at a time by fixing the other two. The algorithm repeats for a certain number of iterations and converges within the first few iterations. Their algorithm is given in Algorithm 10.3.1 and a detail description of the procedure can be found in [50].

algorithm 10.3.1 [50] Discriminative Canonical Correlation (DCC)

This algorithm computes a linear transformation $L \in GL_n$ iteratively that optimizes Equation (10.14).
Input: All $P_i \in \mathbb{R}^{n\times k}$ (orthonormal basis for X_i).
Output: $L \in \mathbb{R}^{n\times n}$.

1. Initialize L to I_n.
2. Iterate the following:
   (a) For all i, find QR-decomposition: $L^T P_i = \Phi_i R_i$ and let $U_i = P_i R_i^{-1}$ (orthonormal basis for transformed image sets).
   (b) For every pair i, j, find SVD: $U_i^T L L^T U_j^T = Q_{ij} M_{ij} Q_{ji}^T$ and save the rotation matrices Q_{ij} and Q_{ji}.
   (c) Compute

$$\begin{aligned} S_b &= \sum_{i=1}^{m} \sum_{l\in B_i} (U_l Q_{li} - U_i Q_{il}) (U_l Q_{li} - U_i Q_{il})^T , \\ S_w &= \sum_{i=1}^{m} \sum_{k\in W_i} (U_k Q_{ki} - U_i Q_{ik}) (U_k Q_{ki} - U_i Q_{ik})^T . \end{aligned}$$

   (d) Compute eigenvectors $\{\mathbf{l}_i\}_{i=1}^n$ of $(S_w)^{-1} S_b$ and let $L = [\mathbf{l}_1, \ldots, \mathbf{l}_n]$.

We propose to maximize the objective function in Equation (10.12) numerically by the Steepest Descent and BFGS methods as an initial attempt towards solving this optimization problem. We quickly set up the notations for the Steepest Descent and BFGS methods and give the procedure in Algorithm 10.3.2 and 10.3.3, respectively. Let $F(L) = -E(L)$. We wish to minimize F over all $L \in \mathbb{R}^{n\times d}$. If $L = (l_{ij})$, we stack columns of L so that

$$\begin{aligned} \mathbf{l} &= \left[l_{11} \cdots l_{n1} \, l_{12} \cdots l_{n2} \cdots l_{1d} \cdots l_{nd} \right]^T \\ &= \left[l_1 \, l_2 \cdots l_N \right]^T , \end{aligned}$$

where $N = nd$. Then

$$\nabla F(\mathbf{l}) = \left[\tfrac{\partial F}{\partial l_1}, \tfrac{\partial F}{\partial l_2}, \cdots, \tfrac{\partial F}{\partial l_N} \right]^T ,$$

with

$$\frac{\partial F}{\partial l_i} = \lim_{\Delta_i \to 0} \frac{F(l_1, \ldots, l_{i-1}, l_i + \Delta_i, l_{i+1}, \ldots, l_N) - F(l_1, \ldots, l_i, \ldots, l_N)}{\Delta_i}.$$

The Steepest Descent method is based on the fact that $F(L)$ decreases the fastest if one goes in the direction of the negative gradient of F at a given point $\mathbf{l}$. If follows that, if

$$\mathbf{t} = \mathbf{l} - \gamma \nabla F(\mathbf{l})$$

for an optimal choice $\gamma > 0$, then $F(\mathbf{l}) \geq F(\mathbf{t})$. A *line (linear) search* is generally performed to find the optimal step size γ. Thus, an iterative algorithm for finding a local minimum of F using the direction of negative gradient gives a sequence of solutions $\mathbf{l}^{(0)}, \mathbf{l}^{(1)}, \mathbf{l}^{(2)}, \ldots, \mathbf{l}^{(n)}, \ldots$ such that

$$\mathbf{l}^{(n+1)} = \mathbf{l}^{(n)} - \gamma_n \nabla F(\mathbf{l}^{(n)}), \quad n \geq 0.$$

One major weakness of the Steepest Descent method is that the algorithm can take many iterations to converge. On the other hand, Newton's method provides a better alternative for the search directions and does not get trapped into a local extreme as easily. However, Hessian of the function and its inverse information need to be computable, which is an expensive calculation in general cases. Thus, we consider a class of the Quasi-Newton methods, the Broyden-Fletcher-Goldfard-Shanno (BFGS) method for which an approximate Hessian is computed at each step.

---

algorithm 10.3.2 Steepest Descent Method

Input: Function to be minimized, $F : \mathbb{R}^N \to \mathbb{R}$.
Output: $\mathbf{l} \in \mathbb{R}^N$.

1. Initialization: Let $\mathbf{l}^{(0)}$ be a random vector, $\mathbf{g}_0 = \nabla F(\mathbf{l}^{(0)})$, $\mathbf{d}_0 = \mathbf{g}_0$.
2. Iterate the following:
   (a) Determine the step length γ_k by a *line search* method by solving the optimization
   $$\min_{\gamma_k > 0} F(\mathbf{l}^{(k)} - \gamma_k \mathbf{d}_k).$$
   (b) Update: $\mathbf{l}^{(k+1)} = \mathbf{l}^{(k)} - \gamma_k \mathbf{d}_k$.
   (c) Calculate the new search direction:
   $$\mathbf{d}_{k+1} = \mathbf{g}_{k+1} = \nabla F(\mathbf{l}^{(k+1)}).$$

10.4 Experimental Results

The goal that we set out to accomplish in the beginning of this chapter is to derive a quantity that characterizes a given data set's Grassmann separability from which we

algorithm 10.3.3 [55] (Memoryless) BFGS Method

Input: Function to be minimized, $F : \mathbb{R}^N \to \mathbb{R}$, start at any point $\mathbf{l}^{(0)}$.
Output: $\mathbf{l} \in \mathbb{R}^N$.
Start at $k = 0$ and denote $\mathbf{g}_k = \nabla F(\mathbf{l}^{(k)})$.

1. Set Hessian $\mathbf{H}_k = \mathbf{I}$.
2. Obtain search direction $\mathbf{d}_k = \mathbf{H}_k \mathbf{g}_k$.
3. Determine the optimal step length γ_k by a *line search* method by solving the optimization
$$\min_{\gamma_k > 0} F(\mathbf{l}^{(k)} - \gamma_k \mathbf{d}_k)$$
and obtain $\mathbf{l}^{(k+1)} = \mathbf{l}^{(k)} - \gamma_k \mathbf{d}_k$, $\mathbf{p}_k = \gamma_k \mathbf{d}_k$, and $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. (Select γ_k accurately enough to ensure $\mathbf{p}_k^T \mathbf{q}_k > 0$.)
4. If k is not an integer multiple of n, set
$$\mathbf{H}_{k+1} = \mathbf{I} - \frac{\mathbf{q}_k \mathbf{p}_k^T + \mathbf{p}_k \mathbf{q}_k^T}{\mathbf{p_k^T q_k}} + \left(1 + \frac{\mathbf{q}_k^T \mathbf{q}_k}{\mathbf{p}_k^T \mathbf{q}_k}\right) \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{q}_k}.$$
5. Update: Add 1 to k and return to step 2. If k is an integer multiple of n, return to step 1.

can search for linear subspace representations that increase the separation gap between matching and non-matching subspaces. In this section, we will show that numerical solutions to Equation (10.13) obtained via the Steepest Descent and BFGS methods are able to accomplish this task. In particular, we will start with a data set that is not originally Grassmann separable but become Grassmann separable after applying data points with the linear transformation obtained via numerical solutions of Equation (10.13). Furthermore, recognition results can be improved on this data set when images are first transformed using solutions of the optimization problem.

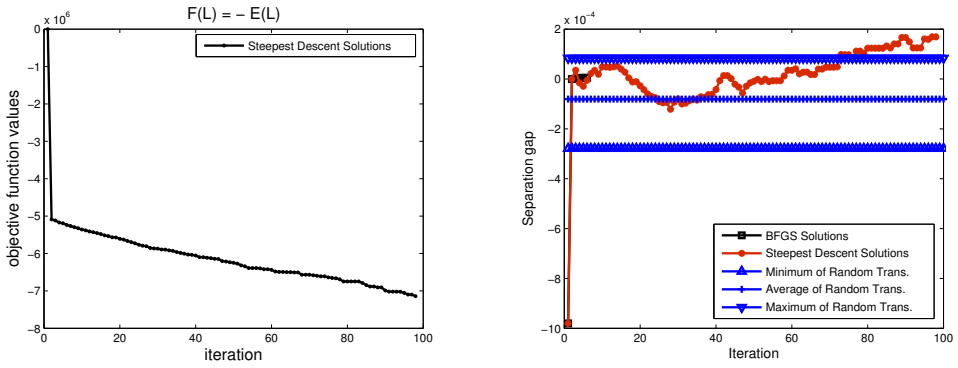
The data set used here is the "illum" subset of the CMU-PIE Database. The images are first projected down to 25 pixels via the scaling component of Haar wavelet analysis to speed up computational time. Two disjoint sets of ten illumination images are selected for each of the 67 subjects. The MATLAB codes for calculating the Grassmann potential objective function and the separation gap for this data set are given in Appendix B.7 and Appendix B.9, respectively. We use the code in Appendix B.6 to get numerical gradient of the objective function. The code in Appendix B.10, which is based on Algorithm 10.3.2, and the code in Appendix B.11, which is based on Algorithm 10.3.3, are used to obtain

numerical solutions to the Grassmann potential objective function for this particular data set with Steepest Descent and BFGS methods, respectively.

The results of applying these numerical solutions are illustrated in Figure 10.1. The separation gap is defined as the difference between the minimum between-class distances and the maximum of the within-class distances. Figure 10.1 (a) depicts the behavior of $F(L) = -E(L)$ as it searches for a local minimum as a function of iterations through the Steepest Descent method. A similar plot for BFGS is observed as well and therefore not included here. Figure 10.1 (b) compares the effects of the linear transformations applied to the data set via Steepest Descent and BFGS methods. Effects of random projections are also plotted for comparison. BFGS converges extremely fast; however, for a particular starting point, it might not converge to an ideal solution. On the other hand, Steepest Descent takes a lot longer to converge; however, it gradually improves to an optimal solution over time. While solutions of both methods turn a non-Grassmann separable data set into one that is Grassmann separable, random projections is shown to be inferior to the solutions of both methods on average. We have seen through this simple example, as a proof of concept, that the optimization problem formulated from studies of perturbation theory provides a useful search criterion for improving Grassmann separability.

As for how meaningful the improvement is, we propose to look at Figure10.2 where the distances between matching and non-matching pairs of subspaces on the 25-pixel data set are tallied. The plot shows that, on average, the distance between a pair of matching and non-matching subspaces is about 0.04 radian apart without transforming the data. This means that the linear transformations obtained via these two methods might not have a very strong ability to alter results of false positives. We suspect improvements of these results can come from better usage of the perturbation theorems.

We further examine the effect of transforming data as a preprocessing step in a face recognition problem. In order to speed up the classification process, PCA is applied to the "illum" subset of CMU-PIE Database to obtain a 15-dimensional feature space, i.e., PCA dimension is 15. We name this data set the 15-pixel data set. Now, randomly select two disjoint sets of 7 images for each person from which linear subspaces are formed. We then performed two sets of experiments by varying the cardinality of the test sets. In one experiment, we let the cardinality of the test sets be one and repeat



(a) (b)

Figure 10.1: (a) Descent of the objective function $F(L) = -E(L)$ using Steepest Descent method. (b) Separation gaps for the 25-pixel data set when images are transformed using random projections and solutions of $F(L) = -E(L)$ obtained via Steepest Descent method and BFGS method. Solutions for random projections are averaged over 100 times with the average, maximum, and minimum values plotted.

	Choice of L				
Cardinality of test sets	Identity	Random	Steepest Descent	BFGS	DCC [50]
1	38.8%	43.3%	33.4%	48.2%	41.5%
7	9.3%	38.1%	6.4%	17.7%	99.1%

Table 10.3: FAR with d^1 using different linear transformations on the 15-pixel data set. Result using the random projections is averaged over 100 times.

for a total of seven times. In another experiment, we let the cardinality of the test sets be seven and repeat for a total of ten times. The average classification errors in FAR for both experiments are shown in Table 10.3. As a comparison, FAR on data under no transformation and random transformation are also given in the same table. In addition, we compare the results obtained with the method described in [50]. Readers should make a note that we make no attempt to optimize the parameters used in [50] and do not claim that our results here are absolutely superior than the ones given in [50]. The bottom line is that by transforming the data using solutions of the optimization problem described in Expression (10.13), we are able to improve classification results in the FAR sense. We want to emphasize that the results described in the current section offers an initial understanding of the optimization problem described in the previous section and future research on ways to improve classification results and methods for solving the optimization problem is anticipated.

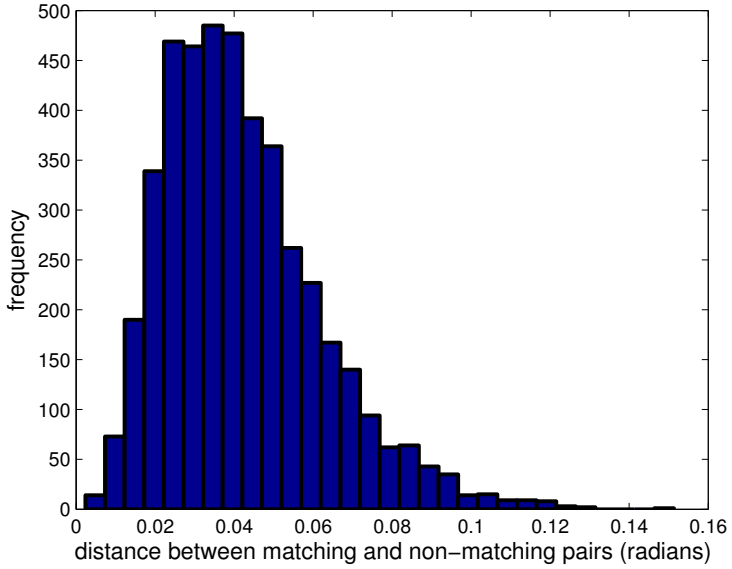


Figure 10.2: Frequency plot for the distance between matching and non-matching pairs for the 25-pixel data set without any linear transformation.

Chapter 11

KARCHER MEAN ON THE RIEMANNIAN MANIFOLDS

In searching for a robust prototype subspace representation that is less prone to perturbation, we conjecture that such a subspace coincides with the mean subspace on the manifold, hence the concept of Karcher mean. This use of mean subspaces in set-to-set object recognition problems will provide us a blueprint to contain discriminatory information with spaces of reduced dimensions. Although the definition of Karcher mean is well-established and mathematically easy to write down, calculating a Karcher mean even on a small collection of sets can be tedious. We will further investigate ways to speed up the convergence rate in search for a mean subspace on a collection of subspaces.

In the Euclidean space $\mathbb{R}$, the definition of the mean is simply the arithmetic average that is commonly known, i.e., for a set of P distinct objects $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(P)}\right\}$, its Euclidean mean is defined as

$$m = \frac{1}{P}\sum_{i=1}^{P} x^{(i)}.$$

Similarly, for a set of objects $\{x^{(1)}, x^{(2)}, \ldots, x^{(P)}\}$, $x^{(i)} \in \mathbb{R}^n$, its Euclidean mean is $m = (m_i)_{i=1}^n$ such that each m_i is defined as

$$m_i = \frac{1}{P}\sum_{j=1}^{P} x_i^{(j)}.$$

Furthermore, the distance between any two points $x^{(i)}$ and $x^{(j)} \in \mathbb{R}^n$ satisfies

$$d^2(x^{(i)}, x^{(j)}) = \sum_{l=1}^{n} \left(x_l^{(i)} - x_l^{(j)}\right)^2.$$

This straight-line distance and arithmetic mean make sense in a space with no curvature, which is the case of Euclidean spaces. We would like to extend the concepts of mean and distance to a more general and curved space, such as a Riemannian manifold, and

in particular, Grassmann manifold. We will briefly review the general definition of a geometric mean on Riemannian manifolds in Chapter 11.1 and a gradient descent method for calculating such means on compact Lie groups in Chapter 11.2. We present a toy example in 11.3 using the algorithm obtained in Chapter 11.2. The main results of the chapter are presented in Chapters 11.4 and 11.5 with an established algorithm for calculating Karcher mean on the Grassmann manifold and a novel algorithm for computing robust prototype representations for points on the Grassmann manifold, respectively.

11.1 Karcher Mean on the Riemannian Manifolds

A Riemannian manifold is a smooth differentiable manifold equipped with a symmetric positive definite metric known as the Riemannian metric. This metric arises from inheriting a canonical metric on the tangent space. Well-known examples of Riemannian manifolds include the Euclidean spaces with the standard inner product and the surface of a sphere such that the shortest distance between any two points on it lies along a great circle passing through the two points. In addition, the Grassmann manifold $G(k, n)$ is also a Riemannian manifold such that when endowed with different differential topology, different Riemannian metric on the Grassmannian is obtained. See Chapter 3.2 for the different geometries on the Grassmannian obtained in such ways.

Certainly the mean on a manifold minimizes the summed squared distance measured along the geodesics. Fréchet [31] in 1948 generalizes the notion of mean to manifolds that is defined globally as follows. If $\mathcal{M}$ is a Riemannian manifold, $d(x, y)$ is the geodesic distance between $x, y \in \mathcal{M}$, and μ is a probability measure on $\mathcal{M}$, then Fréchet mean minimizes

$$F(x) = \frac{1}{2} \int_{\mathcal{M}} d^2(x, y)\, d\mu(y).$$

In the discrete sense, since the estimates of the Fréchet mean derived from random samples of a distribution tend toward the Fréchet mean of the distribution from which the samples were drawn, one can define the Fréchet mean for a set of P samples $\{x^{(1)}, \ldots, x^{(P)}\}$ with respect to the distribution as the x that minimizes

$$F(x) = \frac{1}{2P} \sum_{j=0}^{P} d^2 \left(x, x^{(j)} \right).$$

The constant 2 is used for convenience of notation later on. It is then straightforward to see in the case of the sphere that the Fréchet mean between the north pole and south

pole is not unique since any point on the equator qualifies to be considered as a Fréchet mean. Therefore, Fréchet mean is not necessarily unique. However, a minor adjustment can be made in the definition of Fréchet mean to obtain an unique local minimum. Such means are known as the Karcher or Cartan means [46, 48]. Namely, for a set of P points $x^{(1)}, \ldots, x^{(P)} \in \mathcal{M}$, the Karcher mean q^* is defined as

$$q^* = \arg\min_{q \in \mathcal{M}} \frac{1}{2P} \sum_{i=1}^{P} d^2(x^{(i)}, q).$$

When distributions are limited to a sufficiently small region of a Riemannian manifold, it can be shown that an unique Karcher mean must exist in the restricted region [46].

11.2 Karcher's Local Test For Compact Lie Groups

An algorithm for calculating the Karcher mean on compact Lie groups that is motivated by Karcher's local test for a Karcher mean is given in [93, 57] and Appendix A.1. Karcher's local test for the Karcher mean gives rise to a Riemannian gradient descent algorithm for calculating the Karcher mean of a set of points on a Riemannian manifold that are sufficiently close to the identity. We will briefly review the algorithm in this section.

The heart of the algorithm lies in the ability to work on the tangent space of points on the manifold. In order to access points back and forth between the manifold and the tangent space, we will need the notions of exponential and logarithm maps on the Riemannian manifold.

Definition 11.2.1. Let M be a Riemannian manifold and for any point $p \in M$, denote the tangent space of M at p by T_pM. The Riemannian *Exponential map* $\mathrm{Exp}_{p_0} : T_pM \to M$ is defined as the solution of the geodesic equation, i.e., if $\gamma(t) : [0, 1] \to M$ is a geodesic on M with $\gamma(0) = p_0$ and $\gamma(1) = p_1$, and

$$\frac{d}{dt}\gamma^{'}(t) = 0 \quad \text{(acceleration free)}, \quad \gamma(0) = p_0, \quad \gamma'(0) = v_0 \in T_{p_0}M,$$

then $\mathrm{Exp}_{p_0}(\alpha v_0) = \gamma(\alpha)$.

In particular, $\mathrm{Exp}_{p_0}(v_0) = p_1$. This procedure of fixing a vector $v_0 \in T_{p_0}M$ as an initial velocity for a geodesic gives rise to a natural correspondence between $T_{p_0}M$ and a small neighborhood of p in M. This association is unique in a small ball of

radius ρ (injectivity radius) in $T_{p_0}M$. Within an appropriate neighborhood N_{p_0} of $p_0 \in M$, the inverse process is also unique and defined to be the Riemannian *Logarithm map*. With this notation, $\mathrm{Log}_{p_0} : N_{p_0} \subset M \to T_{p_0}M$ with $\mathrm{Log}_{p_0}(p_1) = v_0$. See a graphical illustration for the correspondence between points in the tangent space and the Riemannian manifold via *Exp* and *Log* maps in Figure 11.1 (analogous to the one given in [69]). In the figure, $\theta(t)$ is a line in the tangent space through 0 and $\gamma(t)$ is a geodesic on M.

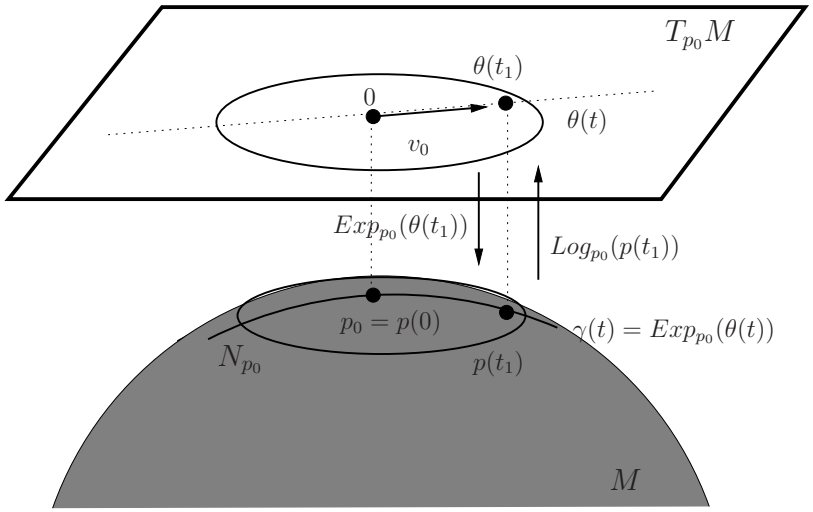


Figure 11.1: [69] Correspondence between points on a geodesic in a manifold M and points on a line in the tangent space $T_{p_0}M$.

For the clarity of the notations, let log and exp be the logarithm and exponential maps for the *Lie groups* and let Log_q and Exp_q be the logarithm and exponential maps for the Riemannian manifolds at the point q. If we adapt the notations from before and let $d(\cdot,\cdot)$ be a distance function on $\mathcal{M}$ and $f(q) = \frac{1}{2P}\sum_{i=1}^{P} d^2(x^{(i)}, q)$, then the following lemma gives the gradient of the cost function $f(q)$.

Lemma 11.2.1. *[57]*

$$\begin{aligned}
\mathrm{grad}\, f(q) &= -\frac{1}{P} q \sum_{i=1}^{P} \exp^{-1}\left(q^{-1}x^{(i)}\right) = -\frac{1}{P} q \sum_{i=1}^{P} \log(q^{-1}x^{(i)}), \quad \textit{or} \\
\mathrm{grad}\, f(q) &= -\frac{1}{P} \sum_{i=1}^{P} \mathrm{Exp}_q^{-1}\left(x^{(i)}\right) = -\frac{1}{P} \sum_{i=1}^{P} \mathrm{Log}_q(x^{(i)}).
\end{aligned}$$

Proof. Let $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ be the Riemannian exponential map, then its relation to the Lie group exponential map is

$$\mathrm{Exp}_x(xA) = x\exp(A). \tag{11.1}$$

Theorem 1.2 of [46] (theorem and proof are given in Appendix A.1) implies that

$$\begin{aligned}\operatorname{grad} f(q) &= -\frac{1}{P}\sum_{i=1}^{P}\mathrm{Exp}_q^{-1}(x^{(i)}) = -\frac{1}{P}\sum_{i=1}^{P}\mathrm{Exp}_q\left(q(q^{-1}x^{(i)})\right)\\ &= -\frac{1}{P}\sum_{i=1}^{P}q\exp^{-1}(q^{-1}x^{(i)}) = -\frac{1}{P}q\sum_{i=1}^{P}\log(q^{-1}x^{(i)}).\end{aligned}$$

□

A necessary condition for q to be the Karcher mean is for $\operatorname{grad} f(q)$ to be zero. But $\operatorname{grad} f(q) = 0$ if and only if $\frac{1}{P}\sum_{i=1}^{P}\log(q^{-1}x^{(i)}) = 0$. This gives us a search direction in the gradient descent algorithm. The following lemma gives us an update criterion.

Lemma 11.2.2. *[57] Let $\mathcal{M}$ be a Riemannian manifold and $f : \mathcal{M} \to \mathbb{R}$ a function whose Riemannian Hessian has all its eigenvalues in the interval $[\delta, 1]$ for $\delta > 0$. Let Exp_x denote the Riemannian exponential map about $x \in \mathcal{M}$. Then, for any initial point $q_0 \in \mathcal{M}$, the sequence*

$$q_{k+1} = \mathrm{Exp}_{q_k}\left(-\operatorname{grad} f(q_k)\right)$$

converges to the unique global minimum of f. Moreover, the distance from q to the minimum is bounded above by $\delta^{-1}||\operatorname{grad} f(q)||$.

Hence, we update the Karcher mean according to the following rule,

$$\exp q_k\left(-\operatorname{grad} f(q_k)\right) = \mathrm{Exp}_{q_k}\left(\frac{1}{P}q_k\sum_{i=1}^{P}\exp^{-1}\left(q_k^{-1}x^{(i)}\right)\right) = \mathrm{Exp}_{q_k}\left(q_kA\right),$$

where $A = \frac{1}{P}\sum_{i=1}^{P}\exp^{-1}\left(q_k^{-1}x^{(i)}\right)$. That is, $q_{k+1} = \mathrm{Exp}_{q_k}\left(q_kA\right) = q_k\exp(A)$.

Algorithm 11.2.1 gives a Riemannian gradient descent method for searching a local Karcher mean on a compact Lie group. Its Riemannian counterpart version is given in Algorithm 11.2.2. The algorithm is modeled after Newton's method (hence, quadratic convergence) and global convergence is guaranteed for points contained in an open ball of a small size.

algorithm 11.2.1 [57] Descent Method For Karcher Mean on Compact Lie Groups

This algorithm calculates the Karcher mean of a set of points on a compact Lie group, $\mathcal{G}$.
INPUT: Points $x^{(1)}, x^{(2)}, \ldots, x^{(P)} \in \mathcal{G}$, ϵ (machine zero).
OUTPUT: Karcher mean, q.

1. Set $q = x^{(1)}$.
2. Compute $A = \frac{1}{P} \sum_{i=1}^{P} \exp^{-1} \left(q^{-1} x^{(i)} \right)$.
3. If $||A|| < \epsilon$, stop and return q, else go to step 4.
4. Update $q := q \exp(A)$, and go to step 2.

algorithm 11.2.2 [57] Descent Method for Karcher Mean on Riemannian Manifolds

This algorithm calculates the Karcher mean of a set of points on a Riemannian manifold, $\mathcal{M}$.
INPUT: Points $x^{(1)}, x^{(2)}, \ldots, x^{(P)} \in \mathcal{M}$, ϵ (machine zero).
OUTPUT: Karcher mean, q.

1. Set $q = x^{(1)}$.
2. Compute $A = \frac{1}{P} \sum_{i=1}^{P} \mathrm{Exp}_q^{-1} \left(x^{(i)} \right)$.
3. If $||A|| < \epsilon$, stop and return q, else go to step 4.
4. Update $q := \mathrm{Exp}_q(A)$, and go to step 2.

11.3 Visualization of Karcher Mean on S^2

This is an example section where we try to visualize the convergence of the Karcher mean Algorithm 11.2.2 on the unit sphere in $\mathbb{R}^3$. $M = S^2$ is the collection of vectors $p \in \mathbb{R}^3$ with $||p|| = 1$. This space may be viewed as the quotient space $SO(3)/SO(2) \simeq M$ and an element can be represented by taking an orthogonal matrix U and retaining only the first column. The tangent space at $p \in M$ is the collection of vectors orthogonal to p, hence isomorphic to $\mathbb{R}^2$. The Riemannian metric is induced by the Euclidean metric on $\mathbb{R}^2$. Fixing a tangent vector $\theta \in T_{p_0}S^2$, the Riemannian exponential on S^2 at p_0 is

$$\mathrm{Exp}_{p_0}(\theta) = \cos(||\theta||)p_0 + \sin\theta\frac{\theta}{||\theta||}.$$

For any p_1 that is not antipodal to p_0, a unique expression for the Riemannian logarithm is

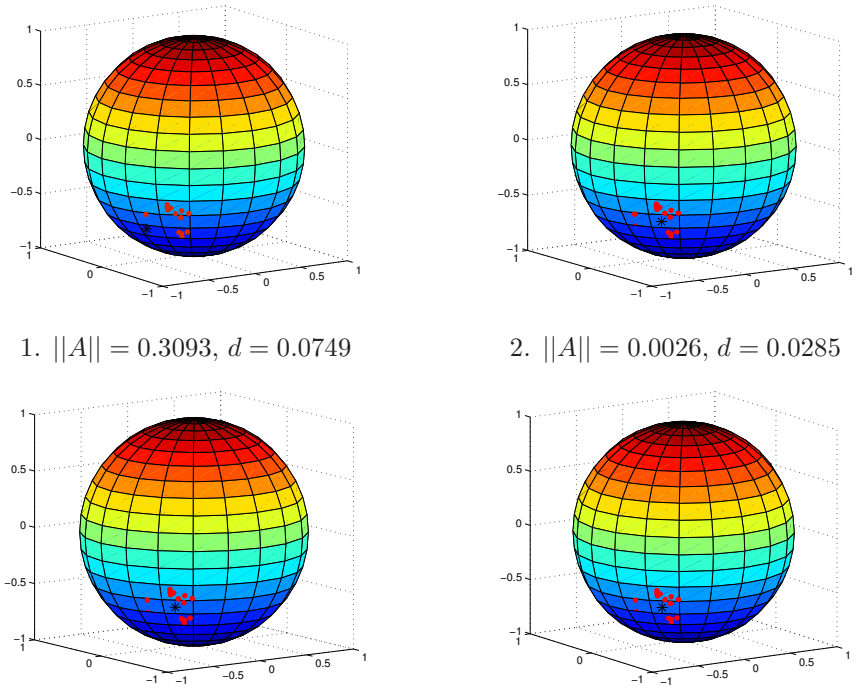
$$\mathrm{Log}_{p_0}(p_1) = cos^{-1}\left(\langle p_1, p_0\rangle\right)\left(\frac{v}{||v||_2}\right),$$

where $\langle\cdot,\cdot\rangle$ is the usual inner product on S^2 and $v = p_1 - \langle p_1, p_0\rangle\, p_0$.

An example of finding the Karcher mean on a random set of points in S^2 using Algorithm 11.2.2 and the *Exp* and *Log* maps given above is shown in Figure 11.2. The convergence is achieved in 4 iterations with $\epsilon = 10^{-6}$.

11.4 Karcher Mean on the Grassmann Manifold

Karcher mean on the Grassmann manifold can be obtained via the algorithm described above. This is because the Grassmann manifold can be realized as the quotient group of the orthogonal group, which is a compact Lie group. A complete and detailed algorithm for computing the Karcher mean on the Grassmann manifold based on the singular value decomposition is given in [5] and [22], which we will briefly review here. An alternative approach for computing Riemannian centroid in naturally reductive homogeneous spaces based on intrinsic Newton method is given in [68]. As the name suggests, the method requires knowledge of the first and second derivatives of the cost function that gives rise to the optimization problem for the Karcher mean. Details on how to find the gradient and Hessian vectors on the Grassmann manifold can be found in [24]. In addition, [24] also offers an algorithm for optimizing functions with Newton's method on the Grassmann manifold. A point $p \in G(k, n)$ gives an equivalence class $[p]$ such



1. $||A|| = 0.3093$, $d = 0.0749$

2. $||A|| = 0.0026$, $d = 0.0285$

3. $||A|| = 2.2215 \times 10^{-5}$, $d = 0.0273$

4. $||A|| = 1.9034 \times 10^{-7}$, $d = 0.0273$

Figure 11.2: Progression of Karcher mean on S^2. For a random set of 20 points (red dots), the Karcher mean (black star) is updated using Algorithm 11.2.2. $||A||$ is the value in Algorithm 11.2.2 and d is the arc length (geodesic) on S^2.

that $p \sim q$ if and only if $q = Q^T p$ for some $Q \in O_k$. The tangent space $T_p G(k,n)$ to $p \in G(k,n)$ is given by

$$T_p G(k,n) = \left\{ w \mid w = p_\perp g, \text{ where } g \in \mathbb{R}^{(n-k)\times n} \text{ and } p_\perp = N(p^T) \right\}.$$

Notice that $p_\perp$ is the orthogonal compliment of p. The Exp_p map that takes a point in the tangent space $T_p G(k,n)$ to a point in $G(k,n)$ is given by

$$\begin{aligned} \text{Exp}_p : T_p G(k,n) &\longrightarrow G(k,n) \\ w &\mapsto pV \cos\Theta + U \sin\Theta \end{aligned} \tag{11.2}$$

where $w \in T_p G(k,n)$ has the SVD $w = U\Theta V^T$. The Log_p map that takes a point in a neighborhood of $p \in G(k,n)$ to a point in $T_p G(k,n)$ is given by

$$\begin{aligned} \text{Log}_p : q \in U_p \subset G(k,n) &\longrightarrow T_p G(k,n) \\ q &\mapsto U\Theta V^T \end{aligned} \tag{11.3}$$

where $p_\perp p_\perp^T q(p^T q)^{-1} = U\Sigma V^T$ and $\Theta = \arctan(\Sigma)$, when it is well-defined. A numerical stable algorithm for finding the Log map is given in Algorithm 11.4.1 that can be found in [22] and MATLAB codes for Algorithm 11.4.1 and all other relevant subroutines are given in Appendix B.3 and B.5, respectively.

---

algorithm 11.4.1 [5, 22] Log Map on Grassmann Manifolds

This algorithm calculates the $\text{Log}_q(p)$ map on the Grassmann manifold.
INPUT: points $p, q \in G(k, n)$.
OUTPUT: $Log_q(p)$.

1. Find the CS decomposition $q^T p = VCZ^T$ and $q_\perp^T p = WSZ^T$, where V, W and Z are orthogonal matrices and C and S are diagonal matrices such that $C^T C + S^T S = I$ [36]. Note that C will always be a square, invertible matrix.
2. Delete(add) zero rows from(to) S so that it is square. Delete the corresponding columns of W (or add zero columns to W), so that it has a compatible size with S.
3. Let $U = q_\perp W$ and $\Theta = \arctan(SC^{-1})$.

Then U, Θ, and V are as in (11.3).

---

algorithm 11.4.2 [5, 22] Descent Method for Karcher Mean on Grassmann Manifolds

---

This algorithm calculates the Karcher mean for a set of points on the Grassmann manifold.
Input: Points $p_1, p_2, \ldots, p_m \in G(k, n)$, ϵ (machine zero).
Output: Karcher mean, q.

1. Set $q = p_1$.
2. Find (using Algorithm 11.4.1)
$$A = \frac{1}{m} \sum_{i=1}^{m} \text{Log}_q(p_i).$$
3. If $||A|| < \epsilon$, return q, else, go to step 4.
4. Find the SVD
$$U\Sigma V^T = A$$
and update
$$q \rightarrow qV \cos(\Sigma) + U \sin(\Sigma).$$
Go to step 2.

If we equip the tangent space with a Frobenius norm, then the induced Riemannian metric is simply the *arc length* or *geodesic distance* on $G(k.n)$, i.e., the distance between

$p, q \in G(k,n)$ can be written in terms of the principal angles $\theta(p,q) = (\theta_1, \theta_2, \ldots, \theta_k)$ (diagonal elements of Θ)as

$$d_g(p,q) = \left(\sum_{i=1}^{k} \theta_i^2 \right)^{1/2}.$$

Given the points $p_1, \ldots, p_m \in G(k,n)$, the Karcher mean is the point q^* that minimizes the sum of the squares of all the principal angles between q^* and p_i's, i.e.,

$$q^* = \arg\min_{q \in G(k,n)} \sum_{j=1}^{m} \sum_{i=1}^{k} \left(\theta_{j,i}(q, p_j)\right)^2. \tag{11.4}$$

The algorithm for finding the Karcher mean on the Grassmann manifold with this definition of the Karcher mean can be found in Algorithm 11.4.2 [5, 22] while a MATLAB code for Algorithm 11.4.2 is given in Appendix B.4. On the other hand, if we change the distance function in the optimization problem (11.4) to the chordal distance, then we obtain a new definition for the Karcher mean on the Grassmann manifold:

$$q^* = \arg\min_{q \in G(k,n)} \sum_{j=1}^{m} \sum_{i=1}^{k} \left(\sin^2 \theta_{j,i}(q, p_j)\right)^2. \tag{11.5}$$

Both definitions are asymptotically equivalent for small principal angles since all the Grassmannian distances generate the same topology. An algorithm for calculating the Karcher mean on the Grassmann manifold using the definition (11.5) is given in [1]. We remark that it will be very interesting to examine the topology generated by the ℓ-truncated Grassmannian semi-metrics and the Karcher mean obtained by replacing the distance function used in Equation (11.5) with these ℓ-truncated Grassmannian semi-metrics.

It is shown in [8] that points on a manifold that lie in a convex ball converge to a point that minimizes its sum squared distance to the points. The convexity of a ball depends on its radius (convexity radius, ρ_c), which satisfies the inequality

$$\rho_c \geq \min\left\{ \frac{1}{2}\rho, \frac{1}{2}\kappa \right\},$$

where ρ is the injectivity radius and κ is an upper bound on the sectional curvature. According to [91], any geodesic in $G(k,n)$ with $\min\{k, n-k\} \geq 2$ that intersects itself is closed, and the minimal length of a closed geodesic is π. Furthermore, the curvature of $G(k,n)$ is bounded by 4 [92]. Thus, as long as there is a $q \in G(k,n)$ that satisfies Karcher's local test for Karcher mean, then q is the unique Karcher mean for $p_1, p_2, \ldots, p_m \in B_{\frac{\pi}{4}}(q)$ [5]. On the other hand, the maximum distance between any two points in $G(k,n)$ is equal to $\min\left\{\sqrt{k}, \sqrt{n-k}\right\} \frac{\pi}{2}$ for $k, n-k \neq 1$ [91].

11.5 Karcher Compression for Face Recognition

For points on $G(k, n)$, reduction in the size of k corresponds to reducing the dimension of the subspace representing a set of digital images. We will accomplish this through a Karcher mean computation. Images obtained via patch collapsing and patch projection typically have a small enough dimension, e.g., 25–100, that the machinery of the Karcher mean is now computationally tractable. We will present in the current section how the calculation of Karcher mean may be used to perform robust classification at reduced computational cost. The results here provide supporting evidence for the potential of exploiting statistics on Grassmann manifolds.

We will follow the experimental protocols in Chapter 8.3 in this set of experiments. We illustrate the use of k-dimensional Karcher representation of illumination feature patches to compress data by comparing the recognition result when using k raw images. A k-dimensional Karcher representation is computed via Algorithm 11.5.1. In short, we randomly split the available data in the gallery into two disjoint sets of equal size. The first k left principal vectors of the pair is computed and saved. This process is repeated t times and the set of k-dimensional left principal vectors is used to compute the Karcher mean.

As an example, using 16 images to generate gallery points and 3 images to generate probe points yields an error-free classification result using the NN classifier on the lip patch of the "lights" data set. For the same gallery we replaced the 16 images by a k-dimensional Karcher representation, where k goes from 1 to 8. When tested on probes of cardinality 3, this resulted in an error-free classification for all k in the NN sense and $k \geq 4$ in the FAR sense as shown in Figure 11.3. The compression of a raw point on $G(16, 41 \cdot 59)$ to a Karcher representation on $G(4, 41 \cdot 59)$ and $G(1, 41 \cdot 59)$ without diminished performance in the FAR and NN sense, respectively, indicates the promise of what we are referring to as Karcher compression in the context of classification of points on Grassmannians. On the contrary, when using k raw images for each gallery point, the error rate is zero for $k \geq 7$ in the NN sense and never reaches zero in the FAR sense. The fact that using a 1-dimensional Karcher representation achieves a perfect recognition result in the NN sense while using 1 raw image in the gallery does not indicates that Karcher representations are able to pack useful information more efficiently. This

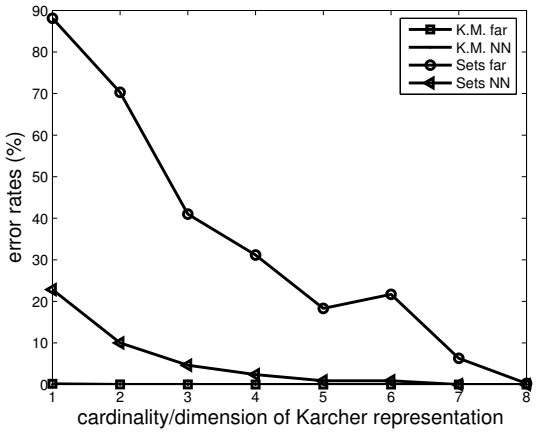


Figure 11.3: Error rate comparisons with k-dimensional Karcher representation and k raw images for points in the gallery corresponding to lip patches. Three images are used to compute points in the probe.

technique can potentially be used to store compact representations computed from video sequences or large data sets where a large number of images is available for the gallery.

Also, it is of potentially substantial interest to exploit the additional information provided by the distribution of distances from the Karcher representations in the gallery generated by either match or non-match probes. As we see in Figure 11.4, 3 images in the probe and a 2-dimensional Karcher representation in the gallery provide good separation between matches and non-matches. The plot is generated by training the Karcher representation 20 times and testing on 100 pairs of matching and 6600 pairs of non-matching incidences. The fact that these distributions are separated suggests a test for detecting false positives when a non-match is identified as a match. A footprint may be left if the identification (incorrectly) of a match is for the wrong reason, i.e., a large distance between the gallery principal vector(s) and the Karcher mean of true positive gallery principal vectors.

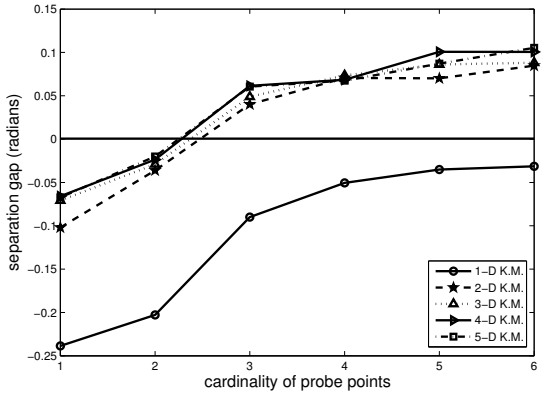


Figure 11.4: A plot of separation vs. cardinality of probe points under varying dimensions of Karcher representation.

---

algorithm 11.5.1 K.M. Representation

This algorithm computes the Karcher mean for a set of principal vectors trained from a set of images of a single person.
Input: k (Karcher dimension), t (training iteration), N (number of images given).
Output: Karcher mean, $< l >_K$.

1. For each training iteration $m = 1 : t$, do the following:

   (a) Let T_m and Q_m be two matrices such that $T_m, Q_m \in \mathbb{R}^{n \times \frac{N}{2}}$ and $\mathcal{R}(T_m)$ and $\mathcal{R}(Q_m)$ do not intersect trivially. Columns of T_m and Q_m are selected from the N input images.

   (b) Find the first k left principal vectors of the pair of subspaces $\mathcal{R}(T_m)$ and $\mathcal{R}(Q_m)$:

   $$T_m = Q_t R_t, \quad Q_t^T Q_t = I_{\frac{N}{2}}, \quad R_t \in \mathbb{R}^{\frac{N}{2} \times \frac{N}{2}}.$$
   $$Q_m = Q_q R_q, \quad Q_q^T Q_q = I_{\frac{N}{2}}, \quad R_q \in \mathbb{R}^{\frac{N}{2} \times \frac{N}{2}}.$$
   $$M = Q_t^T Q_q, \text{ compute the SVD: } M = Y S Z^T.$$

   The left p.v.'s are given by columns of $U = Q_t Y$.
   Let the first k left principal vectors be $l_m = U(:, 1 : k)$.

2. Since each $l_j \in G(k, n)$, find the Karcher mean of $\{l_j\}_{j=1}^m$, $< l >_K$, using Algorithm 11.4.2.

Chapter 12

CONCLUSIONS

In this dissertation, a novel geometric framework for the general classification problem of image sets is proposed. The power of the method is due, in part, to the fact that the geometry and statistics of the Grassmann manifold are well-understood and provide useful tools for quantifying the relationships between patterns. Moreover, improved classification outcomes are often observed when multiple sets of data per subject are available at both training and testing stages. This is perhaps not surprising since families of patterns with a common characterization often possess discriminatory variations that are useful for classification. As it was shown throughout the dissertation that the nature of this information may arise from global features of the pattern, or alternatively, from local features that possess their own special characteristics under a variation of state. Under the right conditions, these families of patterns can be viewed as points on a geometric parameter space called the Grassmannian where well-established distances are available for identifying neighborhood relationships. We made precise this connection, reviewed various ways these metrics on the Grassmannian arise, and how to efficiently compute distances between points on this manifold.

Under this framework, we achieved excellent classification results for a variety of applications in face recognition and offer new insights to the problem in general. As a proof of concept, we first presented two simple two-class classification problems along with state-of-the-art accuracies. We then tackled the well-known illumination problem and obtained perfect recognition results on two largest publicly available databases created for this purpose, CMU-PIE Database and Yale Face Database B. In an attempt to break the method, we were motivated to consider both nonlinear data sets and images of extremely low resolutions. Here the Grassmann framework is robust against resolution

reductions in the sense that the separation gap found in the original ambient space is still observed in the compressed spaces. The benefits of performing classification on such compressed data sets and potential applications were discussed and suggested.

In order to understand how robust the Grassmann framework is against perturbation, we employed tools from matrix perturbation theory where we exploited the natural correspondence between linear subspaces and points on the Grassmannians. Once we defined a notion of data sensitivity, we were then led to formulate an optimization problem using these characteristics as an objective function. To this end, we connected this optimization criterion on the Grassmannian to the idea of *Fisher Linear Discriminant Analysis* on general image sets used in [50]. Numerical solutions obtained using the *Steepest Descent Method* and a *Quasi-Newton (BFGS) Method* showed promising improvements on the separability criterion. That is, data sets will be less sensitive to perturbations (e.g., registration errors and noise in data collection) if points are transformed into spaces where distances among subjects of the same class are minimized while distances among subjects of different classes are maximized.

The thesis is concluded by suggesting a blueprint for extending the Grassmann framework in the general object classification problem. As an initial step, we demonstrated how the use of a geometric concept, Karcher mean, is able to provide robust prototype representations in object classification problems. After briefly setting up the definition of Karcher mean on the Grassmannian, we reviewed a SVD-based numerical technique for obtaining the Karcher mean on this manifold. Finally, a novel algorithm that computes subject prototypical points using the Karcher mean on the Grassmannian was presented based on the success of the framework applied on images of extremely low resolutions. It is worth noting that this calculation would not have been computationally tractable if not for the development of the Grassmann framework on images of extremely low resolutions (e.g., feature patches).

In this dissertation, a lot of new ideas for geometric data analysis are generated through studies of old ideas. In particular, we set up a novel geometric framework for understanding neighborhood relationships for large data sets. Whether a particular metric suits better than other metrics for a particular data set as well as the optimal number of principal angles are needed to construct a particular type of metric for a particular data set are open questions. By studying matrix perturbation theory and Karcher mean, we

found a way to improve robustness of the classifier and reduce computational complexity by solving appropriate objective functions in two optimization problems. The objective function obtained from the perturbation theory can potentially be more effective upon the exploration of the perturbation bounds while the objective function obtained from considering specific types of metric on the Grassmann manifold will provide new insights to the compression technique if different metrics are used. These two topics should serve as subjects of future research. Furthermore, does resolution of the data affect the robustness of the Grassmann framework? Namely, is high resolution data set less sensitive to perturbation and noise than their low resolution counterparts? We envision other parameter spaces such as Stiefel and flag manifolds also present opportunities for extension of these ideas. Additionally, although we focus on illumination as the source of state variation, we remark that other variations in data state, such as those obtained by multi-spectral cameras, also fit into this framework. It is our hope that the suite of these frameworks and algorithms can collectively provide useful insights in studying geometric aspects of large data sets.

Bibliography

[1] P.-A. Absil, R. Mahoney, and R. Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Arta Applicandae Mathematicae*, 80:199–220, 2004.

[2] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, volume 1, pages 581–588. IEEE Computer Society, June 2005.

[3] A. Barg and D. Nogin. Bounds on packings of spheres in the Grassmann manifold. *IEEE Trans. Information Theory*, 48(9):2450–2454, 2002.

[4] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003.

[5] E. Begelfor and M. Werman. Affine invariance revisited. In *CVPR*, volume 2, pages 2087–2094, 2006.

[6] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI*, 19(7):711–720, 1997.

[7] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions. *IJCV*, 28(3):245–260, July 1998.

[8] M. Berger. *A Panoramic View of Riemannian Geometry*. Springer, Berlin, 2003.

[9] A. Björck and G. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[10] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *PAMI*, 25(9):1063–1074, 2003.

[11] R. Brunelli and T. Poggio. Face recognition: feature versus templates. *PAMI*, 15(10):1042–1052, October 1993.

[12] J.-M. Chang. A feature-invariant classification model with applications in face recognition - a solution to varying viewing conditions. Master Paper, July 2004.

[13] J.-M. Chang, J.R. Beveridge, B. Draper, M. Kirby, H. Kley, and C. Peterson. Illumination face spaces are idiosyncratic. In *Int'l Conf. on Image Procesing & Computer Vision*, volume 2, pages 390–396, June 2006.

[14] J.-M. Chang, M. Kirby, H. Kley, J.R. Beveridge, C. Peterson, and B. Draper. Examples of set-to-set image classification. In *Seventh International Conference on Mathematics in Signal Processing Conference Digest*, pages 102–105, The Royal Agricultural College, Cirencester, December 2006. Institute for Mathematics and its Applications.

[15] J.-M. Chang, M. Kirby, H. Kley, C. Peterson, B. Draper, and J. R. Beveridge. Recognition of digital images of the human face at ultra low resolution via illumination spaces. In *Computer Vision – ACCV 2007*, volume 4844 of *LNCS*, pages 733–743. Springer, 2007.

[16] J.-M. Chang, M. Kirby, and C. Peterson. Set-to-set face recognition under variations in pose and illumination. In *2007 Biometrics Symposium*, Baltimore, MD, U.S.A., September 2007.

[17] J.-M. Chang, M. Kirby, and C. Peterson. Feature patch illumination spaces and karcher compression for face recognition via Grassmannians. under preparation, April 2008.

[18] R. Chellapa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[19] Y. Cheng, A. O'Toole, and H. Abdi. Computational approaches to sex classification of adults' and children's faces. *Cognitive Science*, 25:819–838, 2001.

[20] J. Conway, R. Hardin, and N. Sloane. Packing lines, planes, etc.: Packings in Grassmannian spaces. *Experimental Mathematics*, 5:139–159, 1996.

[21] C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[22] D. Dreisigmeyer. Direct search algorithms over Riemannian manifolds. under review, January 2007.

[23] Z. Drmač. On principal angles between subspaces of Euclidean space. *SIAM J. Matrix Anal. Appl.*, 22(1):173–194, 2000.

[24] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[25] H. Ekenel and B. Sankur. Multiresolution face recognition. *Image Vision Computing*, 23(5):469–477, 2005.

[26] K. Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proc. Natl. Acad. Sci.*, 37(11):760–766, 1951.

[27] G. Feng, P. Yuen, and D. Dai. Human face recognition using PCA on wavelet subband. *SPIE J. Electronic Imaging*, 9(2):226–233, April 2000.

[28] A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. In *Proc. ECCV*. Springer-Verlag, 2002.

[29] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *CVPR*, pages 26–36. IEEE Computer Society, 2003.

[30] R. Foltyniewicz. Automatic face recognition via wavelets and mathematical morphology. In *Proc. of the 13th Int'l Conf. on Pattern Recognition*, volume 2, pages 13–17, 1996.

[31] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310, 1948.

[32] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the 2nd European Conference on Computational Learning Theory*, pages 23–37, 1995.

[33] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *11th Intl. Symposium of Robotics Research (ISRR2003)*, pages 192–201, 2003.

[34] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *PAMI*, 23(6):643–660, 2001.

[35] A. Goldstein, L. Harmon, and A. Lesk. Identification of human faces. In *Proc. IEEE*, volume 59, pages 748–760, 1971.

[36] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, third edition, 1996.

[37] G. Golub and H. Zha. Perturbation analysis of the canonical correlations of matrix pairs. *Linear Algebra and its Applications*, 210:3–28, 1994.

[38] P. Griffiths and J. Harris. *Principles of Algebraic Geometry*. Wiley & Sons, 1978.

[39] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. In *AFGR*, pages 1–7, May 2002.

[40] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *PAMI*, 26(4):449–465, April 2004.

[41] J. Harris. *Algebraic Geometry: A First Course*. Springer, 1992.

[42] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–372, 1936.

[43] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hal, Englewood Cliffs, NJ, 1988.

[44] C. Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique*, 3:103–174, 1875.

[45] T. Kanade. Picture processing system by computer complex and recognition of human faces. PhD Thesis, 1973. Department of Information Science.

[46] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30:509–541, 1977.

[47] Y. Kaya and K. Kobayasaki. A basic study on human face recognition. In *Frontiers of Pattern Recognition*, pages 265–290, 1972.

[48] W. Kendall. Probability, convexity and harmonic maps with small image I: Uniqueness and fine existence. In *Proceedings of the London Mathematical Society*, volume 61, pages 371–406, 1990.

[49] T-K. Kim, O. Arandjelović, and R. Cipolla. Learning over sets using boosted manifold principal angles (BoMPA). In *IAPR British Machine Vision Conference*, pages 779–788, 2005.

[50] T-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *ECCV*, pages 251–262, 2006.

[51] M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for the characterization of human faces. *PAMI*, 12(1):103–108, 1990.

[52] A. Knyazev and M. Argentati. Principal angles between subspaces in an a-based scalar product: Algorithms and perturbation estimates. *SIAM J. Sci. Comput.*, 23(6), 2002.

[53] A. Kouzani, F. He, and K. Sammut. Wavelet packet face representation and recognition. In *IEEE Int'l Conf. on Systems, Man and Cybernetics*, volume 2, pages 1614–1619, Orlando, October 1997.

[54] K-C Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *PAMI*, 27(5):684–698, 2005.

[55] D. Luenberger. *Linear and Nonlinear Programming*. Springer, second edition, 2003.

[56] A. Mansfield and J. Wayman. Best practices in testing and reporting of biometric devices: Version 2.01. Technical Report NPL Report CMSC 14/02, Centre for Mathematics and Scientific Computing, National Physical Laboratory, UK, August 2002.

[57] J. Manton. A globally convergent numerical algorithm for computing the center of mass on compact Lie groups. In *Eighth Intl. Conf. on Control, Automation, Robotics and Vision*, volume 3, pages 2211–2216, 2004.

[58] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11:50–59, 1960.

[59] C. Nastar. The image shape spectrum for image retrieval. Technical Report RR-3206, INRIA, 1997.

[60] C. Nastar, B. Moghaddam, and A. Pentland. Flexible images: Matching and recognition using learned deformations. *Computer Vision and Image Understanding*, 65(2):179–191, 1997.

[61] National Science and Technology Council Subcommittee on Biometrics Washington DC. The national biometrics challenge. Special Publication, August 2006.

[62] J. Von Neumann. Some matrix-inequalities and metrization of matrix-space. *Tomsk Univ. Rev.*, 1, 1937.

[63] M. Nishiyama, O. Yamaguchi, and K. Fukui. Face recognition with the multiple constrained mutual subspace method. In *5th International Conference on Audio- and Video-based Biometric Person Authentication, (AVBPA2005)*, pages 71–80, 2005.

[64] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, and O. Yamaguchi. Recognizing faces of moving people by hierarchical image-set matching.

[65] E. Oja. *Subspace Methods of Pattern Recognition*. Research Studies Press LTD., 1983.

[66] J. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recogition grand challenge. In *CVPR*, pages 947–954. IEEE Computer Society, 2005.

[67] L. Qiu, Y. Zhang, and C.-K. Li. Unitarily invariant metrics on the Grassmann space. *SIAM J. Matrix Anal. Appl.*, 27(2):507–531, 2005.

[68] J.P. Costeira V. Barroso R. Ferreira, J. Xavier. Newton method for riemannian centroid computation in naturally reductive homogeneous spaces. In *Proceedings of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages III–III, 2006.

[69] I. Rahman, I. Drori, V. Stodden, D. Donoho, and P. Schroeder. Multiscale representations for manifold-valued data. *Multiscale Modeling & Simulation*, 4(4):1201–1232, 2005.

[70] R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object. *PAMI*, 24(10):1322–1333, 2002.

[71] T. Riklin-Raviv and A. Shashua. The quotient image: Class based re-rendering and recognition with varying illuminations. *PAMI*, 23(2):129–139, 2001.

[72] S. Romdhani, J. Ho, T. Vetter, and D. Kriegman. Face recognition using 3-D models: Pose and illumination. *Proceedings of the IEEE*, 94(11):1977–1999, 2006.

[73] G. Shakhnarovich, J. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. of European Conf. on Computer Vision*, pages 851–865. Copenhagen, Denmark, May 2002.

[74] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *PAMI*, 25(12):1615–1618, 2003.

[75] P. Simard, Y. Cun, J. Denker, and B. Victorri. Transformation invariance in pattern recognition - tangent distance and tangent propagation. *IJIST*, 11:181–194, 2001.

[76] L. Sirovich and M. Kirby. A low-dimensional procedure for the characterization of human faces. *J. of the Optical Society of America A*, 4(3):529–524, 1987.

[77] G. Stewart. On the perturbation of pseudo-inverses, projections, and linear least squares problems. *SIAM Review*, 19:634–662, 1977.

[78] G. Stewart. Computing the CS decomposition of a partitioned orthogonal matrix. *Numerische Mathematik*, 40:297–306, 1982.

[79] G. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Acamedic Press, 1990.

[80] J.-G. Sun. The stability of orthogonal projections. *J. Graduate School*, 1:123–133, 1984.

[81] J.-G. Sun. Perturbation of angles between linear subspaces. *Journal of Computational Mathematics*, 5(1):58–61, 1987.

[82] Q.-S. Sun, P-A. Heng, Z. Jin, and D-S. Xia. Face recognition based on generalized canonical correlation analysis. In *International Conference on Intelligent Computing*, pages 958–967. Hefei, China, 2005.

[83] Q.-S. Sun, Z.-D. Liu, P.-A. Heng, and D.-S. Xia. A theorem on the generalized canonical projective vectors. *Pattern Recognition*, 38(3):449–452, March 2005.

[84] M. Turk and A. Penland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[85] N. Vasconcelos and A. Lippman. Multiresolution tangent distance for affine-invariant classification. *Advances in Neural Info. Proc. Sys. (NIPS)*, 10:843–849, 1998.

[86] N. Vasconcelos and A. Lippman. A multiresolution manifold distance for invariant image similarity. *IEEE Trans. Multimedia*, 7(1):127–142, 2005.

[87] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.

[88] P. Wedin. On angles between subspaces of a finite dimensional inner product space. In *In B.Kågström and A. Ruthe, editors, Matrix Pencil Proceedings, volume 973 of Lecture Notes in Mathematics*, pages 263 – 285, Berlin Heidelberg New York, 1983. Springer-Verlag.

[89] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In Gerald Sommer, Kostas Daniilidis, and Josef Pauli, editors, *Proc. 7th Intl. Conf. on Computer Analysis of Images and Patterns, CAIP'97, Kiel*, number 1296, pages 456–463, Heidelberg, 1997. Springer-Verlag.

[90] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *JMLR*, 4(10):913–931, 2003.

[91] Y.-C. Wong. Differential geometry of Grassmann manifolds. *Proc. Natl. Acad. Sci. USA*, 57:589–594, 1967.

[92] Y.-C. Wong. Sectional geometry of grassmann manifolds. In *Proceedings of the National Academy of Sciences (U.S.A.)*, volume 60, pages 75–79, 1968.

[93] R. Woods. Characterizing volume and surface deformations in an atlas framework: theory, applications and implementation. *NeuroImage*, 18:769–788, 2003.

[94] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *AFGR*, pages 318–323, 1998.

[95] M. Yang, N. Ahuja, and D. Kriegman. Face recognition using kernel eigenfaces. In *Proceeding of IEEE, ICIP*, pages 37–40, 2000.

[96] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.

[97] S. Zhou and R. Chellappa. Image-based face recognition under illumination and pose variations. *J. Opt. Soc. Am. A*, 22(2):217–229, Feb 2005.

Appendix A

PROOFS

A.1 Karcher's Local Test

Let A be a measure space of volume 1. (Mostly A will be a compact Riemannian manifold or a finite set of points.) Let M be a complete Riemannian manifold and $B_\rho(m)$ a convex open ball of radium ρ around m in M ("Convex": For any $p, q \in B$ we require that the shortest geodesic from p to q is unique in M and lies in B; sufficiently small Riemannian balls are convex.) We call any measurable map $f : A \to B_\rho$ a "mass distribution" on B_ρ and define, as in the Euclidean situation,

$$P_f : \bar{B}_\rho \to \mathbb{R}, \quad P_f(m) = \frac{1}{2}\int_A d(m, f(a))^2\, da, \tag{A.1}$$

where $d(\cdot,\cdot)$ is the Riemannian distance of M and $<,>$ will be the Riemannian scalar product. We use now the Riemannian exponential map.

Theorem A.1.1. *[46]*

$$grad P_f(m) = -\int_A \exp_m^{-1} f(a) da. \tag{A.2}$$

Proof. [46] Let $\gamma : I \to B_\rho$ be a geodesic and consider the family of geodesic from $f(a)$ to $\gamma(t)$: $c_a(s,t) = \exp_{f(a)}(s \cdot \exp_{f(a)}^{-1} \gamma(t))$. ($s$ parameterizes points $f(a)$ in $B_\rho(m)$, t parameterizes points in the geodesic, $\exp_{f(a)}^{-1} \gamma(t)$ is a point in $T_{f(a)}A$). Denote $c_a^{'} = \frac{d}{ds} c_a(s,t)$, $\dot{c}_a = \frac{d}{dt} c(s,t)$. Since d is a Riemannian metric, $d(m, f(a)) = |c_a^{'}(s,t)|$ and is independent of s and that $s \to \dot{c}_a(s,t)$ is a family of Jacobi fields. Now,

$$
\begin{aligned}
\frac{d}{dt}P_f\left(\gamma(t)\right) &= \frac{1}{2}\frac{d}{dt}\int_A d^2\left\langle\gamma(t), f(a)\right\rangle da = \frac{1}{2}\frac{d}{dt}\int_A \left\langle c_a^{'}(s,t), c_a^{'}(s,t)\right\rangle \\
&\underset{productrule}{=} \frac{1}{2}\int_A \left[(c_a^{'}(s,t))^T\frac{D}{dt}c_a^{'}(s,t) + \frac{D}{dt}(c_a^{'}(s,t))^T c_a^{'}(s,t)\right] da \\
&= \frac{1}{2}\int_A 2\left(\frac{D}{dt}(c_a^{'})^T\cdot c_a^{'}\right) da = \int_A \left\langle \frac{D}{dt}c_a^{'}, c_a^{'}\right\rangle da \\
&= \int_A \left\langle \frac{D}{ds}\frac{D}{dt}c_a, \frac{D}{ds}c_a\right\rangle da \quad \text{(by continuity)} \\
&= \int_A \left\langle \frac{D}{ds}\dot{c}_a, c_a^{'}\right\rangle da = \int_A\left(\int_0^1 \left\langle \frac{D}{ds}\dot{c}_a, c_a^{'}\right\rangle\right) da \\
&\quad \text{(since } |c_a^{'}| \text{ is independent of } s\text{)} \\
&= \int_A\int_0^1 \frac{D}{ds}\left\langle \dot{c}_a, c_a^{'}\right\rangle ds\, da \\
&\quad \left(\frac{d}{ds}\left\langle \dot{c}_a, c_a^{'}\right\rangle = \left(\frac{d}{ds}\dot{c}_a\right)c_a^{'} + \overbrace{\dot{c}_a\left(\frac{d}{ds}c_a^{'}\right)}^{=0}, \text{property of geodesic}\right) \\
&= \int_A \frac{D}{ds}\int_0^1 \left\langle \dot{c}_a, c_a^{'}\right\rangle ds\, da \\
&\underset{FTC}{=} \int_A \left\langle \dot{c}_a(1,t), c_a^{'}(1,t)\right\rangle - \left\langle \overbrace{\dot{c}_a(0,t)}^{=0}, c_a^{'}(0,t)\right\rangle da \\
&= \int_A \left\langle \dot{c}_a(1,t), c_a^{'}(1,t)\right\rangle da.
\end{aligned}
$$

Now, $\dot{c}_a(1,t) = \frac{d}{dt}c_a(1,t) = \dot{\gamma}(t)$ is independent of a, and $c_a^{'}(1,t) = \frac{d}{ds}c_a(1,t) =$ tangent vector of the geodesic from $f(a)$ to $\gamma(t) = -\exp_{\gamma(t)}^{-1} f(a)$. Thus,

$$
\operatorname{grad} P_f(\gamma(t)) = \frac{d}{dt}P_f(\gamma(t)) = \int_A \left\langle\ \dot{\gamma}(t), -\exp_{\gamma(t)}^{-1} f(a)\right\rangle da.
$$

Replace $\gamma(t)$ with m:

$$
\operatorname{grad} P_f(m) = \int_A \left\langle \overbrace{\dot{\gamma}(t)}^{=1}, -\exp_m^{-1} f(a)\right\rangle da = -\int_A \exp_m^{-1} f(a) da.
$$

□

A.2 Lemma 3.2.1

Lemma 3.2.1 Suppose that $\sigma(U_{\mathcal{X}}^H U_{\mathcal{Y}}) = \{c_k\}_{k=1}^q$, $c_k = \cos\theta_k$, $\frac{\pi}{2} \geq \theta_1 \geq \ldots \geq \theta_q \geq 0$. If $(U_{\mathcal{X}}, W_{\mathcal{X}})$ forms an $n \times n$ unitary matrix and $\sigma(W_{\mathcal{X}}^H U_{\mathcal{Y}}) = \{s_k\}_{k=1}^q$, $s_1 \geq \ldots \geq s_q$, then

$$
s_k = \sin\theta_k, \quad k = 1, \ldots, q.
$$

Proof. [81] Consider the identity

$$\left(U_{\mathcal{X}}^H U_{\mathcal{Y}}\right)^H \left(U_{\mathcal{X}}^H U_{\mathcal{Y}}\right) + \left(W_{\mathcal{X}}^H U_{\mathcal{Y}}\right)^H \left(W_{\mathcal{X}}^H U_{\mathcal{Y}}\right) = I_q. \tag{A.3}$$

First notice that since $(U_{\mathcal{X}}, W_{\mathcal{X}})$ forms a $n \times n$ unitary matrix, thus

$$\left(U_{\mathcal{X}} W_{\mathcal{X}}\right)\left(U_{\mathcal{X}} W_{\mathcal{X}}\right)^H = I_n \Rightarrow \left(U_{\mathcal{X}} W_{\mathcal{X}}\right)\begin{pmatrix} U_{\mathcal{X}}^H \\ W_{\mathcal{X}}^H \end{pmatrix} = U_{\mathcal{X}} U_{\mathcal{X}}^H + W_{\mathcal{X}} W_{\mathcal{X}}^H = I_n.$$

Now, to verify the identity, we see that

$$U_{\mathcal{Y}}^H U_{\mathcal{X}} U_{\mathcal{X}}^H U_{\mathcal{Y}} + U_{\mathcal{Y}}^H W_{\mathcal{X}} W_{\mathcal{X}}^H U_{\mathcal{Y}} = U_{\mathcal{Y}}^H \left(U_{\mathcal{X}} U_{\mathcal{X}}^H + W_{\mathcal{X}} W_{\mathcal{X}}^H\right) U_{\mathcal{Y}} = U_{\mathcal{Y}}^H I_n U_{\mathcal{Y}} = I_q.$$

On the other hand, from the singular value decomposition of $U_{\mathcal{X}}^H U_{\mathcal{Y}}$, we get

$$\begin{aligned} \left(U_{\mathcal{X}}^H U_{\mathcal{Y}}\right)^H \left(U_{\mathcal{X}}^H U_{\mathcal{Y}}\right) &= V C^H U^H U C V^H = V C^H C V^H \\ &= V \operatorname{diag}(c_1^2, \ldots, c_q^2) V^H = \operatorname{diag}(c_1^2, \ldots, c_q^2). \end{aligned}$$

Similarly,

$$\left(W_{\mathcal{X}}^H U_{\mathcal{Y}}\right)^H \left(W_{\mathcal{X}}^H U_{\mathcal{Y}}\right) = \operatorname{diag}(s_1^2, \ldots, s_q^2).$$

It follows from the identity (A.3) that

$$c_k^2 + s_k^2 = 1, \quad k = 1, \ldots, q.$$

Thus, the relations $s_k = \sin\theta_k$, $k = 1, \ldots, q$. □

A.3 Lemma 3.2.2

Lemma 3.2.2 Assume the notations above for $\mathcal{X}$, $\mathcal{Y}$, $U_{\mathcal{X}}$, $U_{\mathcal{Y}}$, and $W_{\mathcal{X}}$, we have

$$\sigma_+(U_{\mathcal{X}}^H U_{\mathcal{Y}}) = \sigma_+\left(P_{\mathcal{X}} P_{\mathcal{Y}}\right) \tag{A.4}$$

and

$$\sigma_+(W_{\mathcal{X}}^H U_{\mathcal{Y}}) = \sigma_+\left(\left(I - P_{\mathcal{X}}\right) P_{\mathcal{Y}}\right). \tag{A.5}$$

Proof. [81] Suppose the SVD of $U_{\mathcal{X}}^H U_{\mathcal{Y}}$ is

$$U_{\mathcal{X}}^H U_{\mathcal{Y}} = U_1 \begin{pmatrix} C_1 & 0 \\ 0 & 0 \end{pmatrix} V_1^H, \tag{A.6}$$

where $U_1^H U_1 = V_1^H V_1 = V_1 V_1^H = I_q$, $C_1 = \operatorname{diag}(c_1, \ldots, c_r)$ contains the nonzero singular values in descending order, $r \le q$. Clearly,

$$\sigma_+(U_{\mathcal{X}}^H U_{\mathcal{Y}}) = \{c_k\}_{k=1}^r. \tag{A.7}$$

On the other hand, (A.6) implies

$$\begin{aligned} P_{\mathcal{X}} P_{\mathcal{Y}} &= U_{\mathcal{X}} (U_{\mathcal{X}}^H U_{\mathcal{Y}}) U_{\mathcal{Y}}^H \\ &= U_{\mathcal{X}} U_1 \begin{pmatrix} C_1 & 0 \\ 0 & 0 \end{pmatrix} V_1^H U_{\mathcal{Y}}^H = U_2 \begin{pmatrix} C_1 & 0 \\ 0 & 0 \end{pmatrix} V_2^H, \end{aligned}$$

where $U_2 = U_{\mathcal{X}} U_1 \in \mathbb{C}^{n\times q}$ and $V_2 = U_{\mathcal{Y}} V_1 \in \mathbb{C}^{n\times q}$ satisfying $U_2^H U_2 = V_2^H V_2 = I_q$. This decomposition means that

$$\sigma_+(P_{\mathcal{X}} P_{\mathcal{Y}}) = \{c_k\}_{k=1}^r. \tag{A.8}$$

Comparison of (A.7) and (A.8) gives (A.4).

For (A.5), observe that

$$W_{\mathcal{X}} W_{\mathcal{X}}^H = I_n - U_{\mathcal{X}} U_{\mathcal{X}}^H = I - P_{\mathcal{X}}.$$

If

$$W_{\mathcal{X}}^H U_{\mathcal{Y}} = U_3 \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} V_3^H,$$

where $U_3^H U_3 = V_3^H V_3 = V_3 V_3^H = I_q$, then

$$\sigma_+(W_{\mathcal{X}}^H U_{\mathcal{Y}}) = \{s_k\}_{k=1}^r \tag{A.9}$$

and

$$\begin{aligned} (I - P_{\mathcal{X}}) P_{\mathcal{Y}} &= W_{\mathcal{X}} (W_{\mathcal{X}}^H U_{\mathcal{Y}}) U_{\mathcal{Y}}^H \\ &= W_{\mathcal{X}} U_3 \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} V_3^H U_B^H = U_4 \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} V_4^H. \end{aligned}$$

This implies that

$$\sigma_+((I - P_{\mathcal{X}}) P_{\mathcal{Y}}) = \{s_k\}_{k=1}^r. \tag{A.10}$$

Comparison of (A.9) and (A.10) gives (A.5). □

A.4 Derivation of chordal F-norm

Let X and Y be two unitary matrices that span the range of the subspaces $\mathcal{X}, \mathcal{Y} \in G(k, n)$, respectively. The the chordal F-norm between $\mathcal{X}$ and $\mathcal{Y}$ is given by

$$d_{cF}(\mathcal{X}, \mathcal{Y}) := \min_{U,V \in O_k} ||XU - YV||_F = ||2 \sin \frac{1}{2}\theta||_2.$$

Proof. [79] Assume $2k \le n$. Without loss of generality, assume X and Y are the generator matrices for $\mathcal{X}$ and $\mathcal{Y}$, respectively. i.e.,

$$X = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}, \quad Y = \begin{bmatrix} C \\ S \\ 0 \end{bmatrix}$$

by CS-decomposition. One must find unitary matrices U and V that minimize

$$\left\| \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} U - \begin{bmatrix} C \\ S \\ 0 \end{bmatrix} V \right\|_F^2 = \|U - CV\|_F^2 + \|SV\|_F^2 = \|U - CV\|_F^2 + \|S\|_F^2 .$$

The second term of the right hand side is independent of U and V. Thus, U and V must minimize $\|U - CV\|_F^2$. But

$$\begin{aligned} \|U - CV\|_F^2 &= \text{trace}\left((U - CV)^H (U - CV) \right) \\ &= \text{trace}\left(\left(U^H - CV^H\right)(U - CV) \right) \\ &= \text{trace}\left(U^H U - CU^H V - CU^H V - CV^H U \right) \\ &= \text{trace}\left(I + C^2 - CJ^H V - CV^H U \right). \end{aligned}$$

This quantity is minimized when the diagonals of U and V are one. But U and V are unitary imply that $U = V = I$. So,

$$\begin{aligned}\|U - CV\|_F^2 + \|S\|_F^2 &= \operatorname{trace}\left(I + C^2 - 2C + S^2\right) \\ &= 2\operatorname{trace}(I - C) = 2\sum_{i=1}^{k}(1 - \cos\theta_i) \\ &= 4\sum_{i=1}^{k}\sin^2\frac{1}{2}\theta_i = \left\|2\sin\frac{1}{2}\theta\right\|_2 .\end{aligned}$$

□

A.5 Derivation of chordal 2-norm

Let X and Y be two unitary matrices that span the range of the subspaces $\mathcal{X}, \mathcal{Y} \in G(k, n)$, respectively. The the chordal 2-norm between $\mathcal{X}$ and $\mathcal{Y}$ is given by

$$d_{c2}(\mathcal{X}, \mathcal{Y}) := \min_{U,V \in O_k} ||XU - YV||_2 = ||2\sin\frac{1}{2}\theta||_F.$$

Proof. [79] Without loss of generality, assume X and Y are the generator matrices for $\mathcal{X}$ and $\mathcal{Y}$, respectively. i.e.,

$$X = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}, \quad Y = \begin{bmatrix} C \\ S \\ 0 \end{bmatrix}$$

by CS-decomposition. Notice

$$\min_{U,V}\|XU - YV\|_2 \Rightarrow \min_{U,V}\sigma(XU - YV) = \min_{U,V}\lambda_{\max}\left((XU - YV)^H(XU - YV)\right),$$

where

$$\begin{aligned}\left((XU - YV)^H(XU - YV)\right) &= \left(\begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}U - \begin{bmatrix} C \\ S \\ 0 \end{bmatrix}V\right)^H\left(\begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}U - \begin{bmatrix} C \\ S \\ 0 \end{bmatrix}V\right) \\ &= \begin{bmatrix} U - CV \\ -SV \\ 0 \end{bmatrix}^H\begin{bmatrix} U - CV \\ -SV \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} U^H - CV^H - SV^H \end{bmatrix}\begin{bmatrix} U - CV \\ -SV \\ 0 \end{bmatrix} \\ &= \left(U^H - CV^H\right)(U - CV) + \left(-SV^H\right)(-SH) \\ &= U^HU - CU^HV - CV^HU + C^2V^HV + S^2V^HV \\ &= I + C^2 + S^2 - CU^HV - CV^HU \\ &= 2I - CU^HV - CV^HU\end{aligned}$$

The expression $\lambda_{\max}\left(2I - CU^H V - CV^H U\right)$ is minimized when $U = V = I$, thus

$$\begin{aligned}
\min_{U,V} \|XU - YV\|_2 &= \lambda_{\max}\left(2I - CU^H V - CV^H U\right) \\
&= \lambda_{\max}\left(2I - 2C\right) = 2\lambda_{\max}\left(I - C\right) \\
&= \lambda_{\max} \begin{pmatrix} 2(1-\cos\theta_1) & & \\ & \ddots & \\ & & 2(1-\cos\theta_k) \end{pmatrix} \\
&= \lambda_{\max} \begin{pmatrix} 4\sin^2\frac{1}{2}\theta_1 & & \\ & \ddots & \\ & & 4\sin^2\frac{1}{2}\theta_k \end{pmatrix} \\
&= \lambda_{\max} \begin{pmatrix} 2\sin\frac{1}{2}\theta_1 & & \\ & \ddots & \\ & & 2\sin\frac{1}{2}\theta_k \end{pmatrix} \\
&= \left\| 2\sin\frac{1}{2}\theta \right\|_F .
\end{aligned}$$

□

Appendix B

MATLAB CODES

B.1 Code for Algorithm 3.4.1

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% quick algorithm for calculating the principal angles   %
% between the two subspaces spanned by the columns of    %
% A and B when angles are larger than 10^(-6).           %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [C,angles] = prinAngles(A,B)

[Qa,Ra] = qr(A,0);
[Qb,Rb] = qr(B,0);

C = svd(Qa'*Qb,0);
angles = acos(C);
```

B.2 Code for Algorithm 3.4.2

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% a modified algorithm for calculating the principal angles %
% between the two subspaces spanned by the columns of       %
% A and B. Good for small (<10^(-6)) and large angles.      %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [angles] = mPrinAngles(A,B)

[Qa,Ra] = qr(A,0);
[Qb,Rb] = qr(B,0);

C = svd((Qa')*Qb,0);

rkA = rank(Qa);
rkB = rank(Qb);

if rkA >= rkB
    B = Qb - Qa*(Qa'*Qb);
else
    B = Qa - Qb*(Qb'*Qa);
```

```
end
S = svd(B,0);
S = sort(S);

for i = 1:min(rkA,rkB)
    if (C(i))^2 < 0.5
        angles(i) = acos(C(i));
    elseif (S(i))^2 <= 0.5
        angles(i) = asin(S(i));
    end
end
angles = angles';
```

B.3 Code for Algorithm 11.4.1

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% calculate the Log_p(q) map for Grassmann manifold, where%
% p = point of tangency. p,q in G(k,n) with p'p = q'q = I.%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [X] = logMap(p,q)

A = p'*q;
[n,k] = size(p);
B = null(p')'*q;

[V,W,Z,C,S] = csdecomp(A,B);

if n > 2*k
    S = S(1:k,:);
    W = W(:,1:k);
elseif n < 2*k
    S = [S ; zeros(2*k-n,k)];
    W = [W zeros(n-k,2*k-n)];
end
C = diag(1./diag(C));
U = null(p')*W;
T = atan(S*C);

X = U*T*V';
```

B.4 Code for Algorithm 11.4.2

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% P = set of points p1,p2,...,pm,       %
% m = number of points, eps = machine 0 %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
function [pbar] = karcherMean(P,m,eps)

m = size(P,1);
k = size(P,2)/m;

% initialization: set pbar = p1;
pbar = P(:,1:k);
nw = Inf;

while nw >= eps
    [U,S,V] = svd(w,0);
    pbar = pbar*V*funm(S,@cos)+ U*funm(S,@sin); % Exp map
    w = zeros(n,k);
    for i = 1:m
        w = w + Logmap(pbar,P(:,k*(i-1)+1:k*i));
    end
    w = w/m;
    nw = norm(w,'fro');
end
pbar = orth(pbar); %to make pbar orthonormal
```

B.5 Subroutines for B.3

B.5.1 csdecomp.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Given Q_1 and Q_2 such that Q_1^T Q_1 + Q_2^T Q_2 = I,    %
% the CS decomposition is a joint factorization of the      %
% form Q_1 = UCZ^T and Q_2=VSZ^T where U, V, and Z are      %
% orthogonal matrices and C and S are diagonal matrices     %
% (not necessarily square) satisfying C^TC + S^TS = I.      %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [U,V,Z,C,S] = csdecomp(Q1,Q2)

[m,p] = size(Q1);
[n,pb] = size(Q2);
if pb ~= p
   error('MATLAB:gsvd:MatrixColMismatch',
         'Matrices must have the same number of columns.')
end

if m < n
   [V,U,Z,S,C] = csdecomp(Q2,Q1);
   j = p:-1:1; C = C(:,j); S = S(:,j); Z = Z(:,j);
   m = min(m,p);
   i = m:-1:1;
   C(1:m,:) = C(i,:); U(:,1:m) = U(:,i);
   n = min(n,p); i = n:-1:1;
   S(1:n,:) = S(i,:); V(:,1:n) = V(:,i);
```

```
   return
end
% Henceforth, n <= m.

[U,C,Z] = svd(Q1);

q = min(m,p);
i = 1:q;
j = q:-1:1;
C(i,i) = C(j,j);
U(:,i) = U(:,j);
Z(:,i) = Z(:,j);
S = Q2*Z;

if q == 1
   k = 0;
elseif m < p
   k = n;
else
   k = max([0; find(diag(C) <= 1/sqrt(2))]);
end
[V,R] = qr(S(:,1:k));
S = V'*S;
r = min(k,m);
S(:,1:r) = diagf(S(:,1:r));
if m == 1 && p > 1, S(1,1) = 0; end

if k < min(n,p)
   r = min(n,p);
   i = k+1:n;
   j = k+1:r;
   [UT,ST,VT] = svd(S(i,j));
   if k > 0,
      S(1:k,j) = 0;
   end
   S(i,j) = ST;
   C(:,j) = C(:,j)*VT;
   V(:,i) = V(:,i)*UT;
   Z(:,j) = Z(:,j)*VT;
   i = k+1:q;
   [Q,R] = qr(C(i,j));
   C(i,j) = diagf(R);
   U(:,i) = U(:,i)*Q;
end

if m < p
   % Diagonalize final block of S and permute blocks.
   q = min(nnz(abs(diagk(C,0))>10*m*eps(class(C))), ...
       nnz(abs(diagk(S,0))>10*n*eps(class(C))));
   i = q+1:n;
   j = m+1:p;
```

```
  % At this point, S(i,j) should have orthogonal columns
  % and the elements of S(:,q+1:p) outside of S(i,j)
  % should be negligible.

[Q,R] = qr(S(i,j));
   S(:,q+1:p) = 0;
   S(i,j) = diagf(R);
   V(:,i) = V(:,i)*Q;
   if n > 1
      i = [q+1:q+p-m 1:q q+p-m+1:n];
   else
      i = 1;
   end
   j = [m+1:p 1:m];
   C = C(:,j);
   S = S(i,j);
   Z = Z(:,j);
   V = V(:,i);
end

if n < p
   % Final block of S is negligible.
   S(:,n+1:p) = 0;
end

% Make sure C and S are real and positive.
[U,C] = diagp(U,C,max(0,p-m));
C = real(C);
[V,S] = diagp(V,S,0);
S = real(S);
```

B.5.2 diagk.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% diagk(X,k) gives the k-th diagonal of X,   %
% even if X is a vector.                      %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function D = diagk(X,k)

if min(size(X)) > 1
   D = diag(X,k);
elseif 0 <= k && 1+k <= size(X,2)
   D = X(1+k);
elseif k < 0 && 1-k <= size(X,1)
   D = X(1-k);
else
   D = [];
end
```

B.5.3 diagf.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% diagf(X) zeros all the elements off the  %
% main diagonal of X                       %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function X = diagf(X)

X = triu(tril(X));
```

B.5.4 diagp.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%diagp(Y,X,k) scales the columns of Y and the rows %
% of X by unimodular factors to make the k-th      %
% diagonal of X real and positive.                 %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [Y,X] = diagp(Y,X,k)

D = diagk(X,k);
j = find(real(D) < 0 | imag(D) ~= 0);
D = diag(conj(D(j))./abs(D(j)));
Y(:,j) = Y(:,j)*D';
X(j,:) = D*X(j,:);
```

B.6 Numerical Gradient

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this code calculates the gradient of the objective function   %
% E(L) in Grassmann potential optimization problem at a given L %
% with three-point approximation that uses secant line through  %
% the points (x-h1,f(x-h1)) and (x+h2,f(x+h2)).                 %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [gradF] = gradDesGPObjFn(Z,k,T)
%% Z = data set
%% k = subspace dimension
%% T = current linear transformation

[r,c] = size(T);

h1 = 0.0001;
h2 = 0.0001;
E1 = T; E2 = T;
for i = 1:r
    for j = 1:c
        E1(i,j) = T(i,j) - h1;
        E2(i,j) = T(i,j) + h2;
        m = calGPObjFn(Z,k,E2) - calGPObjFn(Z,k,E1);
        gradF(i,j) = m/(h1+h2);
```

```
    end
end
```

B.7 Calculation of Objective Function

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this code calculates the function value of the Grassmann   %
% potential objective function for a given transformation L. %
% distance = minimum principal angle.                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function f = calGPObjFn(Z,k,L)

Nsubj = size(Z,3);

for i = 1:Nsubj
    TT(:,:,i) = L'*Z(:,1:k,i);
    TQ(:,:,i) = L'*Z(:,k+1:2*k,i);
end

Dw = []; Db = []; Dbeta = [];
for i = 1:Nsubj
    a = norm(pinv(TT(:,:,i)),2);
    for j = 1:Nsubj
        theta = prin_angles(TT(:,:,i),TQ(:,:,j));
        b = norm(pinv(TQ(:,:,j)),2);
        Dbeta = [Dbeta  (a + b)];
        d = min(theta);
        if j == i %% within-class distances
            Dw = [Dw d];
        else %% between-class distances
            Db = [Db d];
        end
    end
end
Sw = sum(Dw);
Sb = sum(Db);
beta = sum(Dbeta);

f = -1*Sb/(Sw*beta);
```

B.8 Update of Objective Function

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this code updates the Grassmann potential objective function %
% using linear transformation obtained from either the         %
% descent method of BFGS method.                               %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
function f = funUpdate(Z,k,Lold,d,x)
%% Lold = linear transformation obtained in the kth step
%% L = linear transformation obtained in the (k+1)th step
%% f = objective function at the (k+1)th step
%% d = search direction

Nsubj = size(Z,3);
L = Lold - x*d;
f = calGPObjFn(Z,k,L);
```

B.9 Calculation of Separation Gap

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this code calculates the separation gap of a given data  %
% set Z and a given linear transformation L.               %
% distance = minimum principal angle.                      %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function sg = calSepGap(Z,k,L)

Nsubj = size(Z,3);

for i = 1:Nsubj
    TT(:,:,i) = L'*Z(:,1:k,i);
    TQ(:,:,i) = L'*Z(:,k+1:2*k,i);
end

for i = 1:Nsubj
    for j = 1:Nsubj
        theta(i,j,:) = mprin_angles(TT(:,:,i),TQ(:,:,j));
    end
end
D = theta(:,:,1);
N = max(diag(D));

for q = 1:Nsubj
    if q == 1
        off_diag(q,:) = D(q,q+1:Nsubj);
    elseif q == Nsubj
        off_diag(q,:) = D(q,1:Nsubj-1);
    else
        off_diag(q,:) = [D(q,1:(q-1)) D(q,(q+1):Nsubj)];
    end
end
M = min(min(off_diag));
sg = M - N;
```

B.10 Optimizing Grassmann Potential with Algorithm 10.3.2

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this code uses steepest descent method to calculate a   %
% local min of the Grassmann potential objective function %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% preamble %%
load LL5_illum %% images of 25 pixels
X = LL5_illum; clear LL5_illum5
n = length(X(:,1));
k = 10;
Nsubj = 67; Nvar = 21;
epsilon = 10^(-1); % threshold for stopping
max_count = 100; x1 = 0; x2 = 100;
set(0,'RecursionLimit',100)
for i = 1:Nsubj
    Z(:,:,i) = X(:,(i-1)*Nvar+1:i*Nvar);
end

%% step 1: initialize L and find step size
Lold = rand(n,n);
d = gradDesGPObjFn(Z,k,Lold);
[x,objfval(1),exitflag,output] = ...
fminbnd(@(x) funUpdate(Z,k,Lold,d,x), x1, x2)
sep_gap(1) = calSepGap(Z,k,Lold)
Lold = Lold - x*d;

%% iterate the following %%

w = inf;
counter = 2;
while (w > epsilon) & (counter < max_count)
    %% step 2: get gradident of F at current L
        d = gradDesGPObjFn(Z,k,Lold);

    %% step 3: optimize step size
    [x,objfval(counter),exitflag,output] = ...
    fminbnd(@(x) funUpdate(Z,k,Lold,d,x),x1,x2)

    %% step 4: update L using the new gamma (= x)
    sep_gap(counter) = calSepGap(Z,k,Lold)
    L = Lold - x.*d;
    w = norm(L-Lold,'fro')
    Lold = L;
    cd results/transformed_results/
    save OptTranSteepest_k10_x100 L sep_gap objfval
    cd ../../
    counter = counter + 1
end
```

B.11 Optimizing Grassmann Potential with Algorithm 10.3.3

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this code uses BFGS method to calculate a                             %
% local min of the Grassmann potential objective function               %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% preamble %%

load LL5_illum
X = LL5_illum; clear LL5_illum
n = length(X(:,1));
m = n^2; k = 10;
Nsubj = 67; Nvar = 21;
epsilon = 10^(-2);
max_count = 20; x1 = 0; x2 = 100;
set(0,'RecursionLimit',100)
for i = 1:Nsubj
    Z(:,:,i) = X(:,(i-1)*Nvar+1:i*Nvar);
end

%% step 1: initialize L and H
cd results/transformed_results/
load OptTranSteepest_k10_x100_good
cd ../../
Lold = L;
clear objfval sep_gap
l = reshape(Lold,m,1);
I = eye(m); H = eye(m);
Gold = gradDesGPObjFn(Z,k,Lold);
gold = reshape(Gold,m,1);
dold = H*gold;
Dold = reshape(dold,n,n);
[x,objfval(1),exitflag,output] = ...
fminbnd(@(x) funUpdate(Z,k,Lold,Dold,x), x1, x2)
sep_gap(1) = calSepGap(Z,k,Lold)
L = Lold + x*Dold;

%% iterate the following %%

w = inf;
counter = 2;
while (w > epsilon) & (counter <= max_count)
    G = gradDesGPObjFn(Z,k,L);
    g = reshape(G,m,1);
    Q = G - Gold;
    q = reshape(Q,m,1);
    p = x*dold;

    %% step 2: approximate Hessian and update search direction
    H = I - (q*p'+ p*q')./(p'*q) + (1 + (q'*q)/(p'*q))*(p*p')*(1/(p'*q));
    dold = H*g;
    Dold = reshape(dold,n,n);
```

```
    %% step 3: optimize step size
    [x,objfval(counter),exitflag,output] = ...
    fminbnd(@(x) funUpdate(Z,k,L,Dold,x),x1,x2)

    %% step 4: update L using the new gamma (= x)
    sep_gap(counter) = calSepGap(Z,k,L)
    Lold = L;
    L = Lold - x.*Dold;
    w = norm(L-Lold,'fro')
    Gold = G;
    cd results/transformed_results/
    save OptBFGS_k10_x100 L sep_gap objfval
    cd ../../
    counter = counter + 1;
end
```