This CVPR2015 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Projection Metric Learning on Grassmann Manifold with Application to Video based Face Recognition

Zhiwu Huang[1,2], Ruiping Wang[1], Shiguang Shan[1], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
zhiwu.huang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Abstract

In video based face recognition, great success has been made by representing videos as linear subspaces, which typically lie in a special type of non-Euclidean space known as Grassmann manifold. To leverage the kernel-based methods developed for Euclidean space, several recent methods have been proposed to embed the Grassmann manifold into a high dimensional Hilbert space by exploiting the well established Project Metric, which can approximate the Riemannian geometry of Grassmann manifold. Nevertheless, they inevitably introduce the drawbacks from traditional kernel-based methods such as implicit map and high computational cost to the Grassmann manifold. To overcome such limitations, we propose a novel method to learn the Projection Metric directly on Grassmann manifold rather than in Hilbert space. From the perspective of manifold learning, our method can be regarded as performing a geometry-aware dimensionality reduction from the original Grassmann manifold to a lower-dimensional, more discriminative Grassmann manifold where more favorable classification can be achieved. Experiments on several real-world video face datasets demonstrate that the proposed method yields competitive performance compared with the state-of-the-art algorithms.

1. Introduction

Nowadays, linear subspaces have proven a powerful representation for video based face recognition [43, 41, 10, 30, 27, 12, 11, 16, 7], where each video can be treated as a set of face images without considering temporal information. As well recognized, a set of face images of a single person can be well approximated by a low dimensional linear subspace [43]. The benefits of using subspaces lie in its much lower computational cost of comparing large data sets and its well justified capacity of modeling complex appearance variations in the set data [12]. However, such advantages come along with the challenge of representing and handling the subspaces appropriately, with their unique geometric structure being concerned.

As is mostly studied in [42, 28, 9, 32, 12, 34, 21, 22], linear subspaces with the same dimensionality reside on a special type of Riemannian manifold, i.e. Grassmann manifold, which has a nonlinear structure. As a consequence, popular techniques developed for Euclidean spaces are not directly eligible for the Grassmannian data. To tackle this problem, a number of works [42, 9, 32, 1, 12, 11, 5, 16, 15] studied and explored the Riemannian geometry of the Grassmann manifold. Among them, quite a few works [42, 9, 12, 11, 5, 16, 15] exploited the projection mapping to represent each element (i.e., linear subspace) on the Grassmann manifold with its projection operator. The resulting projection distance, i.e., Projection Metric developed in [9], is related to the true geodesic distance on the Grassmann manifold at an infinitesimal scale [15]. By encoding the geometry of the Grassmann manifold, traditional algorithms developed in Euclidean space can be extended to new versions on the nonlinear manifold. For example, Cetingul [5] *et al.* proposed a new clustering approach on the Grassmann manifold, and Harandi *et al.* [15] presented a Grassmannian Dictionary Leaning approach.

In this paper, we focus on the problem of conducting discriminant analysis on the Grassmann manifold for video based face recognition. Under the projection mapping framework, most of recent studies [12, 11, 16, 14, 36] exploited a series of positive definite kernel functions on Grassmann manifold to first embed the manifold into a high dimensional Hilbert space, which actually obeys Euclidean geometry. Then, the flattened manifold is mapped into a lower-dimensional Euclidean space (see Fig.1 (a)-(b)-(d)-(e)). The Grassmann kernels allow us to treat the Grassmann manifold as if it were a Euclidean vector space. As a result, kernel learning algorithms (e.g., kernel discriminant

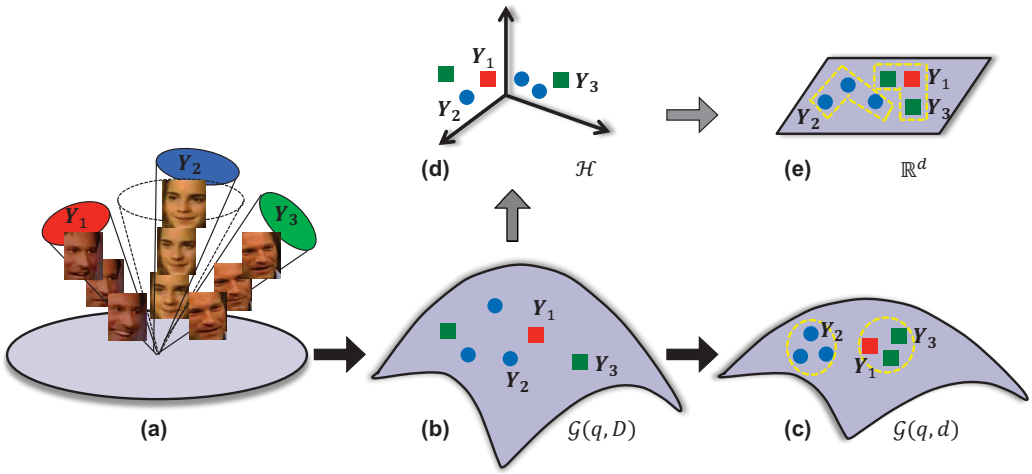

Figure 1. Conceptual illustration of the proposed Projection Metric Learning (PML) on the Grassmann Manifold. Traditional Grassmann discriminant analysis methods take the away (a)-(b)-(d)-(e) to first embed the original Grassmann manifold $\mathcal{G}(q, D)$ (b) into high dimensional Hilbert space $\mathcal{H}$ (d) and then learn a map from the Hilbert space to a lower-dimensional, optionally more discriminative space $\mathbb{R}^d$ (e). In contrast, the newly proposed approach goes the way (a)-(b)-(c) to learn the metric/mapping from the original Grassmann manifold $\mathcal{G}(q, D)$ (b) to a new more discriminant Grssmann manifold $\mathcal{G}(q, d)$ (c).

analysis [2]) in vector spaces can be extended to their counterparts on Grassmann manifold. However, several drawbacks of traditional kernel learning algorithms are also introduced to the Grassmann manifold, such as the derivation of kernel function typically involves a complex theoretical/technical problem for satisfying Mercer's theorem to generate valid Reproducing Kernel Hilbert Space (RKHS). Furthermore, the transformed data in Hilbert space are usually implicitly known[1] and only a measure of similarity between them is available through the derived kernel function. Last but not least, the computational burden of constructing the involved kernel matrix is considerably high scaling with the size of data samples.

To overcome the limitations of existing Grassmann discriminant analysis methods, by endowing the well-studied Projection Metric with Grassmann manifold, we attempt to learn a Mahalanobis-like matrix on the Grassmann manifold without resorting to kernel Hilbert space embedding. In contrast to the kernelization scheme, our approach directly works on the original manifold and exploits its geometry to learn a representation that still benefits from useful properties of the Grassmann manifold. Furthermore, the learned Mahalanobis-like matrix can be decomposed into the transformation for dimensionality reduction, which maps the original Grassmann manifold to a lower-dimensional, more discriminative Grassmann manifold (see Fig.1 (a)-(b)-(c)). While in the literature a couple of subspace-based dimensionality reduction techniques [10, 30, 27] have also been proposed, they attempt to pursue a transformation using the canonical correlation based distance, which is not a structured metric [12]. Accordingly, these techniques fail to explore the data structure of Grassmann manifold spanned by the linear subspaces. Different from them, by exploring the Grassmannian geometry, our method directly learns the Projection Metric which is eligible to induce a positive definite kernel. Consequently, it is qualified to serve as a pre-processing step for other kernel-based methods on Grassmann manifold by feeding them more discriminative manifold-valued data and thus further improve them.

[1]Although one can employ Nystrom approximation to obtain a vectorised representation of kernel data, this is only an approximation of explicit map.

2. Related work

In this section, we review in more details on several dimensionality reduction techniques for linear subspaces, existing kernel-based Grassmann discriminant analysis methods, and two manifold-to-manifold map learning works in previous literature.

In early years, there are several dimensionality reduction methods for linear subspaces such as Constrained Mutual Subspace Method (CMSM) [10, 30] and Discriminant Canonical Correlations (DCC) [27]. The CMSM approach exploits a constrained subspace where the subspace pairs from different classes have small canonical correlations. However, this method is very susceptible to the dimensionality of the constrained subspace. The DCC algorithm attempts to maximize the canonical correlations of within-class subspace pairs and minimizes the canonical correlations of between-class subspace pairs by transforming the original linear subspaces to low-dimensional ones. However, as noted in [12], when all subspaces are sharply con-

centrated on one point, the max correlation distance used in CMSM and DCC will be close to zero for most data. Furthermore, the max correlation distance is not a metric and may not be used together with more sophisticated algorithms. Lastly, these techniques fail to explore the specific data structure of linear subspaces, which typically reside on Grassmann manifold, and may thus learn an undesirable transformation for them.

More recently, there are several approaches [12, 11, 16, 14, 36] to treat linear subspaces on Grassmann manifold and learn kernel-based discriminant functions on this specific manifold. For example, by deriving Grassmann kernels based on Projection Metric, Grassmann Discriminant Analysis (GDA) [12] first embeds the Grassmann manifold into high dimensional Hilbert space and then learns a map to a lower-dimensional, more discriminative space under Fisher LDA criteria. Grassmannian Graph-embedding Discriminant Analysis (GGDA) [16] further improves GDA under a more general graph embedding discriminative learning framework. Although these methods can be employed for supervised classification, they are limited to the Mercer kernels which yields implicit projection, and thus restricted to use only kernel-based classifiers. Moreover, the computational complexity of these kernel-based methods increases with the number of training sample.

In this paper, by exploiting Riemannian geometry of the Grassmann manifold with Projection Metric, we alternatively learn a Mahalanobis-like (i.e., symmetric positive semidefinite (PSD)) matrix on the PSD manifold without relying on Hilbert space embedding. In the literature, we find two relevant works [31, 35] also optimize a parameterized matrix on a certain type of manifold. However, their ideas are totally different from ours, that is, they attempt to optimize a transformation matrix on Stiefel manifold for dimension reduction with data vectors as elements lying in Euclidean space. In contrast, our work learns a Mahalanobis-like matrix on PSD manifold for the problem of projection metric learning with linear subspaces as elements residing on Grassmann manifold. Additionally, the learned Mahalanobis-like matrix can be also regarded as a dimension reduction transformation from the original Grassmann manifold to a lower-dimensional, more discriminative Grassmann manifold. To our knowledge, there are only two similar works [24, 13] seeking to learn the mapping from manifold to manifold. However, the work [24] learns the mapping from high-dimensional spheres to submanifolds of decreasing dimensionality while the other one [13] seeks an embedding of high-dimensional SPD manifold into a low-dimensional SPD manifold.

3. Preliminaries

Before presenting our approach, let's begin with a brief summary of the basic Riemannian geometry of Grassmann manifold, which provides the grounding for the proposed algorithm. Details on Grassmann manifold and related topics can be found in [42, 28, 9, 32, 1, 18, 15].

A Grassmann manifold $\mathcal{G}(q, D)$ is the set of q-dimensional linear subspaces of the $\mathbb{R}^D$ and it is a $q(D-q)$ dimensional compact Riemannian manifold. An element of $\mathcal{G}(q, D)$ is a linear subspace span($\boldsymbol{Y}$), which is spanned by its orthonormal basis matrix $\boldsymbol{Y}$ of size $D \times q$ such that $\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is the identity matrix of size $q \times q$.

Under the projection mapping $\Phi(\boldsymbol{Y}) = \boldsymbol{Y}\boldsymbol{Y}^T$ framework, an alternative strategy proposed in [9] is to represent the elements on the Grassmann manifold with projection matrices $\boldsymbol{Y}\boldsymbol{Y}^T$. As noted in the work [18], the projection embedding is a diffeomorphism from a Grassmann manifold onto idempotent symmetric matrices of rank q, i.e., it is a one-to-one, continuous, differentiable mapping with a continuous, differentiable inverse. As a result, there exists one such unique projection matrix corresponding to each point on the Grassmann manifold.

Since the projection operator $\Phi(\boldsymbol{Y})$ is a $D \times D$ symmetric matrix, a natural choice of inner product is $\langle \boldsymbol{Y}_1, \boldsymbol{Y}_2 \rangle_\Phi = tr(\Phi(\boldsymbol{Y}_1)^T\Phi(\boldsymbol{Y}_2))$. The inner product is invariant to the specific realization of a subspace, and induces a distance:

$$d_p(\boldsymbol{Y}_1\boldsymbol{Y}_1^T, \boldsymbol{Y}_2\boldsymbol{Y}_2^T) = 2^{-1/2}\|\boldsymbol{Y}_1\boldsymbol{Y}_1^T - \boldsymbol{Y}_2\boldsymbol{Y}_2^T\|_F. \quad (1)$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. Since such distance satisfies the axioms of a metric, it is also called Projection Metric [12]. As proved in [15], the Projection Metric is able to approximate the true Grassmannian geodesic distance up to a scale of $\sqrt{2}$, and thus has become one of the most popular metrics on the Grassmann manifold.

4. Projection Metric Learning

In this section, we first formulate the problem of our proposed Projection Metric Learning (PML) on Grassmann manifold for video based face recognition. Then we describe the optimization of our problem.

4.1. Formulation

Assume m video sequences of face frames are given as $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m\}$, where $\boldsymbol{X}_i \in \mathbb{R}^{D\times n_i}$ describes a data matrix of the i-th video containing n_i frames, each frame being expressed as a D-dimensional feature vector. In these data, each video belongs to one of face classes denoted by C_i. The i-th video $\boldsymbol{X}_i$ is represented by a q-dimensional linear subspace spanned by an orthonormal basis matrix $\boldsymbol{Y}_i \in \mathbb{R}^{D\times q}$, s.t. $\boldsymbol{X}_i\boldsymbol{X}_i^T \simeq \boldsymbol{Y}_i\boldsymbol{\Lambda}_i\boldsymbol{Y}_i^T$, where $\boldsymbol{\Lambda}_i$, $\boldsymbol{Y}_i$ correspond to the matrices of the q largest eigenvalues and eigenvectors respectively.

Given a linear subspace span($\boldsymbol{Y}_i$) on Grassmann manifold (as discussed before, we interchangeably denote $\boldsymbol{Y}_i\boldsymbol{Y}_i^T$

as the elements on the manifold), we seek to learn a generic mapping $f : \mathcal{G}(q, D) \rightarrow \mathcal{G}(q, d)$ that is defined as

$$f(\boldsymbol{Y}_i\boldsymbol{Y}_i^T) = \boldsymbol{W}^T\boldsymbol{Y}_i\boldsymbol{Y}_i^T\boldsymbol{W} = (\boldsymbol{W}^T\boldsymbol{Y}_i)(\boldsymbol{W}^T\boldsymbol{Y}_i)^T. \quad (2)$$

where $\boldsymbol{W} \in \mathbb{R}^{D\times d}$ $(d \leq D)$, is a transformation matrix of column full rank. With this mapping, the original Grassmann manifold $\mathcal{G}(q, D)$ can be transformed into a lower-dimensional Grassmann manifold $\mathcal{G}(q, d)$. However, except the case $\boldsymbol{W}$ is an orthogonal matrix, $\boldsymbol{W}^T\boldsymbol{Y}_i$ is not generally an orthonormal basis matrix. Note that only the linear subspaces spanned by orthonormal basis matrix can form a valid Grassmann manifold. To tackle this problem, we temporarily use the orthonormal components of $\boldsymbol{W}^T\boldsymbol{Y}_i$ defined by $\boldsymbol{W}^T\boldsymbol{Y}_i^{'}$ to represent an orthonormal basis matrix of the transformed projection matrices. As for the approach to get the $\boldsymbol{W}^T\boldsymbol{Y}_i^{'}$, we will give more details in the next subsection. Now, we first study the Projection Metric on the new Grassmann manifold and the proposed objection function in the following.

Learned Projection Metric. The Projection Metric of any pair of transformed projection operators $\boldsymbol{W}^T\boldsymbol{Y}_i^{'}\boldsymbol{Y}_i^{'T}\boldsymbol{W}$, $\boldsymbol{W}^T\boldsymbol{Y}_j^{'}\boldsymbol{Y}_j^{'T}\boldsymbol{W}$ is defined by:

$$\begin{aligned} & d_p^2(\boldsymbol{W}^T\boldsymbol{Y}_i^{'}\boldsymbol{Y}_i^{'T}\boldsymbol{W}, \boldsymbol{W}^T\boldsymbol{Y}_j^{'}\boldsymbol{Y}_j^{'T}\boldsymbol{W}) \\ & = 2^{-1/2}\|\boldsymbol{W}^T\boldsymbol{Y}_i^{'}\boldsymbol{Y}_i^{'T}\boldsymbol{W} - \boldsymbol{W}^T\boldsymbol{Y}_j^{'}\boldsymbol{Y}_j^{'T}\boldsymbol{W}\|_F^2 \\ & = 2^{-1/2}tr(\boldsymbol{P}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\boldsymbol{P}). \end{aligned} \quad (3)$$

where $\boldsymbol{A}_{ij} = \boldsymbol{Y}_i^{'}\boldsymbol{Y}_i^{'T} - \boldsymbol{Y}_j^{'}\boldsymbol{Y}_j^{'T}$ and $\boldsymbol{P} = \boldsymbol{W}\boldsymbol{W}^T$. Since $\boldsymbol{W}$ is required to be a matrix with column full rank, $\boldsymbol{P}$ is a rank-d symmetric positive semidefinite (PSD) matrix of size $D \times D$, which has a similar form as Mahalanobis matrix.

Discriminant Function. The discriminant function is designed to minimize the projection distances of any within-class subspace pairs while to maximize the projection distances of between-class subspace pairs. The matrix $\boldsymbol{P}$ is thus achieved by the objective function $J(\boldsymbol{P})$ as:

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} J(\boldsymbol{P}) = \arg\min_{\boldsymbol{P}} (J_w(\boldsymbol{P}) - \alpha J_b(\boldsymbol{P})). \quad (4)$$

where α reflects the trade-off between the within-class compactness term $J_w(\boldsymbol{P})$ and between-class dispersion term $J_b(\boldsymbol{P})$, which are measured by average within-class scatter and average between-class scatter respectively as:

$$J_w(\boldsymbol{P}) = \frac{1}{N_w}\sum_{i=1}^{m}\sum_{j:C_i=C_j} 2^{-1/2}tr(\boldsymbol{P}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\boldsymbol{P}). \quad (5)$$

$$J_b(\boldsymbol{P}) = \frac{1}{N_b}\sum_{i=1}^{m}\sum_{j:C_i\neq C_j} 2^{-1/2}tr(\boldsymbol{P}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\boldsymbol{P}). \quad (6)$$

where N_w is the number of pairs of samples from the same class, N_b is the number of pairs of samples from different classes, $\boldsymbol{A}_{ij} = \boldsymbol{Y}_i^{'}\boldsymbol{Y}_i^{'T} - \boldsymbol{Y}_j^{'}\boldsymbol{Y}_j^{'T}$ and $\boldsymbol{P}$ is the PSD matrix that needs to be learned.

4.2. Optimization

The optimization problem Eq.4 includes the variable $\boldsymbol{P}$ as well as $\boldsymbol{Y}^{'}$. Since $\boldsymbol{Y}^{'}$ is not explicitly expressed by $\boldsymbol{P}$, it is hard to find a closed form solution for $\boldsymbol{P}$. We propose an iterative solution for one of the two variables at a time by fixing the other and repeating for a certain number of iterations. To make the columns of $\boldsymbol{W}^T\boldsymbol{Y}$ be orthonormal, one of the proposed iterative optimizations involves normalization of $\boldsymbol{Y}$. Since $\boldsymbol{P}$ is a rank-d PSD matrix of size $D \times D$ as discussed before, we exploit the nonlinear Riemannian Conjugate Gradient (RCG) method [9, 1] on the manifold of PSD matrices to seek the optimal $\boldsymbol{P}$ when fixing $\boldsymbol{Y}^{'}$.

Normalization of $\boldsymbol{Y}$. For all i, the matrix $\boldsymbol{Y}_i$ need to be normalized to $\boldsymbol{Y}_i^{'}$ for a fixed $\boldsymbol{P} = \boldsymbol{W}\boldsymbol{W}^T$ so that the columns of $\boldsymbol{W}^T\boldsymbol{Y}_i$ are orthonormal. Specifically, we perform QR-decomposition of $\boldsymbol{W}^T\boldsymbol{Y}_i$ s.t. $\boldsymbol{W}^T\boldsymbol{Y}_i = \boldsymbol{Q}_i\boldsymbol{R}_i$, where $\boldsymbol{Q}_i \in \mathbb{R}^{D\times q}$ is the orthonormal matrix composed by the first q columns and $\boldsymbol{R}_i \in \mathbb{R}^{q\times q}$ is the invertible upper-triangular matrix. Since $\boldsymbol{R}_i$ is invertible and $\boldsymbol{Q}_i$ is orthonormal, we can make $\boldsymbol{W}^T\boldsymbol{Y}_i^{'}$ become an orthonormal basis matrix by normalizing $\boldsymbol{Y}_i$ as:

$$\boldsymbol{Q}_i = \boldsymbol{W}^T(\boldsymbol{Y}_i\boldsymbol{R}_i^{-1}) \quad \rightarrow \quad \boldsymbol{Y}_i^{'} = \boldsymbol{Y}_i\boldsymbol{R}_i^{-1}. \quad (7)$$

Computation of $\boldsymbol{P}$. The optimal PSD matrix $\boldsymbol{P}$ is computed for a given $\boldsymbol{Y}_i$ by applying the nonlinear RCG algorithm on the manifold of rank-d PSD matrices of size $D \times D$. With $\boldsymbol{P}$ being on the outside of the trace in Eq.5 and Eq.6, the discriminative function $J(\boldsymbol{P})$ in Eq.4 is transformed as:

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} tr(\boldsymbol{P}\boldsymbol{S}_w\boldsymbol{P}) - \alpha tr(\boldsymbol{P}\boldsymbol{S}_w\boldsymbol{P}). \quad (8)$$

where $\boldsymbol{S}_w$ and $\boldsymbol{S}_b$ are defined as:

$$\boldsymbol{S}_w = \frac{1}{N_w}\sum_{i=1}^{m}\sum_{j:C_i=C_j} 2^{-1/2}tr(\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T). \quad (9)$$

$$\boldsymbol{S}_b = \frac{1}{N_b}\sum_{i=1}^{m}\sum_{j:C_i\neq C_j} 2^{-1/2}tr(\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T). \quad (10)$$

As the Conjugate Gradient (CG) algorithm developed in Euclidean space, the RCG algorithm on the Riemannian manifold also runs in an iterative procedure (see Algorithm 2). An outline for the iterative part of this algorithm goes as follows: at the k-th iteration, find $\boldsymbol{P}_k$ by searching the minimum of J along the geodesic γ in the direction $\boldsymbol{H}_{k-1}$ from $\boldsymbol{P}_{k-1} = \gamma(k-1)$, compute the Riemannian gradient $\nabla_{\boldsymbol{P}}J(\boldsymbol{P}_k)$ at this point, choose the new search direction to be a combination of the old search direction and the new gradient, i.e., $\boldsymbol{H}_k \leftarrow -\nabla_{\boldsymbol{P}}J(\boldsymbol{P}_k) + \eta\tau(\boldsymbol{H}_{k-1}, \boldsymbol{P}_{k-1}, \boldsymbol{P}_k)$, and iterate until convergence. In the procedure, the Riemannian gradient $\nabla_{\boldsymbol{P}}J(\boldsymbol{P}_k)$ can be approximated from its corresponding Euclidean gradient $D_{\boldsymbol{P}}J(\boldsymbol{P}_k)$ by $\nabla_{\boldsymbol{P}}J(\boldsymbol{P}_k) =$

Algorithm 1 Projection Metric Learning (PML)

Input: All linear subspaces $span(\boldsymbol{Y}_i) \in \mathcal{G}(q, D)$
1. $\boldsymbol{W} \leftarrow \boldsymbol{I}_{D\times d}, \boldsymbol{P} \leftarrow \boldsymbol{I}_D$.
2. **Do iterate the following:**
3. Normalize $\boldsymbol{Y}_i$ by using Eq.7 for all i.
4. Compute $\boldsymbol{S_w}$ and $\boldsymbol{S_b}$ by using Eq.9 and Eq.10.
5. Optimize $\boldsymbol{P}$ in Eq.8 by using Algorithm 2.
6. Update $\boldsymbol{W}$ by computing the matrix square root of $\boldsymbol{P}$.
7. **End**
Output: The PSD matrix $\boldsymbol{P}$

Algorithm 2 Riemannian Conjugate Gradient (RCG)

Input: The initial PSD matrix $\boldsymbol{P}_0$
1. $\boldsymbol{H}_0 \leftarrow 0, \boldsymbol{P} \leftarrow \boldsymbol{P}_0$.
2. **Repeat**
3. $\boldsymbol{H}_k \leftarrow -\nabla_{\boldsymbol{P}} J(\boldsymbol{P}_k) + \eta\tau(\boldsymbol{H}_{k-1}, \boldsymbol{P}_{k-1}, \boldsymbol{P}_k)$.
4. Line search along the geodesic γ with the direction $\boldsymbol{H}_k$ from $\boldsymbol{P}_{k-1} = \gamma(k-1)$ to find $\boldsymbol{P}_k = \arg\min_{\boldsymbol{P}} J(\boldsymbol{P})$.
5. $\boldsymbol{H}_{k-1} \leftarrow \boldsymbol{H}_k$, $\boldsymbol{P}_{k-1} \leftarrow \boldsymbol{P}_k$.
8. **Until** convergence
Output: The PSD matrix $\boldsymbol{P}$

$D_{\boldsymbol{P}} J(\boldsymbol{P}_k) - \boldsymbol{P}_k \boldsymbol{P}_k^T D_{\boldsymbol{P}} J(\boldsymbol{P}_k)$, and $\tau(\boldsymbol{H}_{k-1}, \boldsymbol{P}_{k-1}, \boldsymbol{P}_k)$ denotes the parallel transport of tangent vector $\boldsymbol{H}_{k-1}$ from $\boldsymbol{P}_{k-1}$ to $\boldsymbol{P}_k$. For a more detailed treatment, we refer the reader to [9, 1]. As for now, we just need to compute the Euclidean gradient $D_{\boldsymbol{P}} J(\boldsymbol{P}_k)$ of Eq.8 as:

$$D_{\boldsymbol{P}} J(\boldsymbol{P}_k) = 2(\boldsymbol{S_w} - \alpha \boldsymbol{S_b})\boldsymbol{P}_k. \tag{11}$$

The main procedure for our Projection Metric Learning (PML) is given in Algorithm1. Once the optimal PSD matrix $\boldsymbol{P}$ is found, a comparison of any two linear subspaces is achieved by using Eq.3. Although providing a theoretical proof of convergence of this optimization algorithm is hard, we find it can generally make the objective function Eq.8 converge to a stable and desirable solution after a few iterations, which will be shown in our experiments.

5. Experiments

In this section, we present extensive experiments to evaluate our proposed PML method on two video based face recognition tasks: video based face identification and video based face verification. We used YouTube Celebrities (YTC) [26] for video based face identification, YouTube Face (YTF) [40] and Point-and-Shoot Face Recognition Challenge (PaSC) [3] for video based face verification task. We will first briefly overview these datasets used in the experiments, followed by a description and discussion of the experiments.

In all experiments, each video was treated as an image set with the data matrix $\boldsymbol{X}_i = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_i}]$, where $\boldsymbol{x}_j \in \mathbb{R}^D$ is the vectorised descriptor of the j-th frame. Through the singular value decomposition (SVD) of $\boldsymbol{X}_i$, the image set can be modeled as a linear subspace. Specifically, we use the leading q left singular-vectors as the orthonormal basis matrix $\boldsymbol{Y}_i$ to represent a q-dimensional linear subspace for $\boldsymbol{X}_i$, which be treated as a point on the Grassmann manifold $\mathcal{G}(q, D)$. In the following experiments, the setting of q is determined by cross-validation.

To study the effectiveness of the proposed PML method, we compare four unsupervised subspace-based methods including Projection Metric (PM) [9], Mutual Subspace Method (MSM) [43], Affine Hull based Image Set Distance (AHISD) [6] and Convex Hull based Image Set Distance (CHISD) [6]. In addition, we also test several state-of-the-art supervised subspace-based learning methods including Constrained Mutual Subspace Method (CMSM) [10], Set-to-set distance metric learning (SSDML) [45], Discriminative Canonical Correlations (DCC) [27], Grassmann Discriminant Analysis (GDA) [12] and Grassmannian Graph-Embedding Discriminant Analysis (GGDA) [16]. For fair comparison, the key parameters of each method are empirically tuned according to the recommendations in the original works. For MSM/AHISD, the first canonical correlation or leading component is exploited when comparing two subspaces. For CMSM/DCC, the dimensionality of the resulting discriminant subspace is tuned from 1 to 10. For SSDML, its key parameters are tuned and empirically set as: $\lambda_1 = 0.001, \lambda_2 = 0.5$, the numbers of positive and negative pairs per sample are 10 and 20 respectively. For GDA/GGDA, the final dimensionality is set $c-1$ (c is the number of face classes in training). In GGDA, the other parameter β is tuned at the range of $\{1e^2, 1e^3, 1e^4, 1e^5, 1e^6\}$. For our PML, the parameter α is set to 0.2.

5.1. Video based Face Identification

The YouTube Celebrities (YTC) [26] is a quite challenging and widely used video face dataset. It has 1,910 video clips of 47 subjects collected from YouTube. Most clips contain hundreds of frames, which are often low resolution and highly compressed with noise and low quality (see Fig.2). Each face in YTC is resized to a 20×20 image as [38, 29] and pre-processed by the histogram equalization to eliminate lighting effects. Then we extract gray feature for each face image. Following the prior works [37, 38, 29], we conduct ten-fold cross validation experiments, i.e., 10 randomly selected gallery/probe combinations. In each fold, one person has 3 randomly chosen videos for the gallery and 6 for probes. In this experiment, each video is represented by a linear subspace of order 10. Finally, the average recognition rates of different methods are reported.

Figure 2. Examples of YouTube Celebrities (YTC) dataset.

Method	YTC
MSM [43]	60.25 ± 3.05
PM [9]	62.17 ± 3.65
AHISD [6]	63.70 ± 2.89
CHISD [6]	66.62 ± 2.79
CMSM [10]	63.81 ± 3.70
SSDML [45]	68.85 ± 2.32
DCC [27]	65.48 ± 3.51
GDA [12]	65.02 ± 2.91
GGDA [16]	66.37 ± 3.52
PML	66.69 ± 3.54
PML-GDA	68.08 ± 3.78
PML-GGDA	70.32 ± 3.69

Table 1. Video based face identification results on YTC dataset. Here, the results are mean rank-1 face recognition rates with standard deviation.

We report the performances of the state-of-the-art methods on this dataset in Tab.1. The results show that our proposed method PML outperforms the baseline methods by learning the Projection Metric on the Grassmann manifold. The performances of our method is comparable to the state-of-the-art methods. Since the learned PSD matrix in our method can be decomposed into the transformations to yield a low-dimensional manifold, this manifold can be fed into the methods GDA and GGDA which explore the same Riemannain metric (i.e., Projection Metric). As shown in Tab.1, both the PML-GDA and the PML-GGDA improve the original methods (i.e., GDA and GGDA) and outperform the other competing methods.

On this dataset, there are several state-of-the-art methods [39, 37, 38, 29, 17, 20] with other kinds of set modeling or other kinds of classifiers. Among them, we implement PLS-based Covariance Discriminant Learning (CDL) and Localized Multi-Kernel Metric Learning (LMKML) methods. Their performances are 70.21% and 70.30% respectively,

Figure 3. Examples of YouTube Face (YTF) dataset.

Figure 4. Examples of Point-and-Shoot Challenge (PaSC) dataset.

which demonstrate our PML-GGDA method can achieve the state-of-the-art.

5.2. Video based Face Verification

For video face verification task, we conduct experiments on two challenging large-scale datasets: YouTube Face (YTF) [40] and Point-and-Shoot Face Recognition Challenge (PaSC) [3]. The YTF [40] contains 3,425 videos of 1,595 different persons collected from the YouTube website. In this database, there exist large variations in pose, illumination, and expression in each video sequence. The PaSC [3] includes 2,802 videos of 265 people carrying out simple actions. Every action was filmed by two cameras: a high quality, 1920×1080 pixel, camera on a tripod and one of five alternative handheld video cameras. The tripod-based data serves as a control. The handheld cameras have resolutions ranging from 640×480 up to 1280×720. As shown in Fig.3 and Fig.4, there are some examples of YTF and PaSC datasets.

On YTF, we follow the standard evaluation protocol [40] to perform standard, ten-fold, cross validation, pair-matching tests. Specifically, we use the officially provided 5,000 video pairs, which are equally divided into 10 folds. Each fold contains 250 intra-personal pairs and 250 inter-personal pairs. On PaSC, there are two video face verification experiments: control-to-control and handheld-to-handheld experiments. In both of the two experiments, the

Method	YTF
MSM [43]	65.20 ± 1.97
PM [9]	65.12 ± 2.00
AHISD [6]	64.80 ± 1.54
CHISD [6]	66.30 ± 1.21
CMSM [10]	66.46 ± 1.54
SSDML [45]	65.38 ± 1.86
DCC [27]	68.28 ± 2.21
GDA [12]	67.00 ± 1.62
GGDA [16]	66.56 ± 2.07
PML	67.30 ± 1.76
PML-GDA	70.88 ± 1.69
PML-GGDA	70.04 ± 2.19

Table 2. Video based face verification results on YTF dataset. Here, the results represent the mean accuracies with standard deviations.

Method	PaSC-control	PaSC-handheld
MSM [43]	35.80	34.56
PM [9]	35.65	33.60
AHISD [6]	21.96	14.29
CHISD [6]	26.12	20.97
CMSM [10]	36.67	36.22
SSDML [45]	29.19	22.89
DCC [27]	38.87	37.53
GDA [12]	41.88	43.25
GGDA [16]	43.35	43.09
PML	37.25	37.23
PML-GDA	42.93	43.64
PML-GGDA	43.63	43.95

Table 3. Video based face verification results when false accept rate is 0.01 on PaSC dataset. Here, the PaSC-control/handheld indicates the experiments with control/handheld videos.

target and query sigsets contains the same set of videos. The task was to verify a claimed identity in the query video by comparing with the associated target video. Since the same 1,401 videos served as both the target and query sets,'same video' comparisons were excluded.

In our experiments, we directly crop the face images according to the provided data and then resize them into 24×40 pixels for YTF as [8] and 256×256 pixels for PaSC. On YTF dataset, we extract the raw intensity feature of resized video frames. On PaSC dataset, we had also conducted experiments using gray features, but found the highest performance is extremely low (around 10%). Therefore, we employ the Caffe [23] to extract the state-of-the-art Deep Convolutional Neural Network (DCNN) feature of the video frames. The DCNN model is pretrained on CFW [44], and then fine-tuned on the data from the training sets of PaSC and COX [19] datasets. In the experiments on YTF and PaSC, each video sequence is modeled as a linear subspace of order 10.

Tab.2 lists the mean accuracies and standard deviations on YTF, and Fig.5 shows the ROC for the video based face verification on YTF. Tab.3 tabulates the verification rates of different methods on PaSC when the false accept rate is 0.01. On YTF dataset, since DCC [27], GDA [12] and GGDA [16] are not specifically designed for pair-wise based face verification, we modify their original LDA-like part as its pairwise version (like the two works [33, 25]) by constructing the within-class scatter matrix from intra-class pairs and the between-class scatter matrix from inter-class pairs. The results in Tab.2 and Tab.3 show that our PML

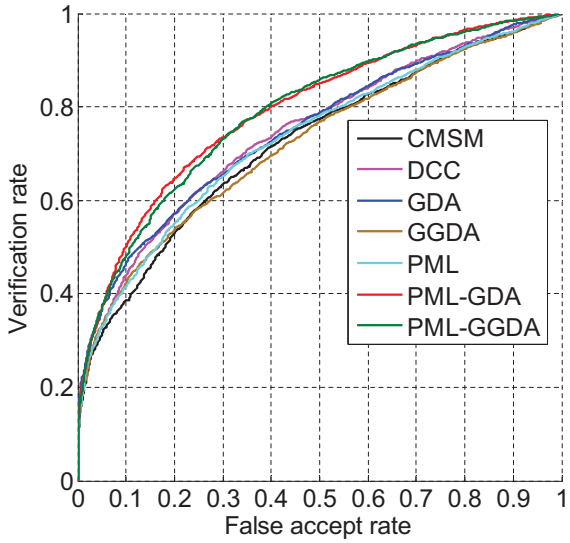

Figure 5. ROC curve for video based face verification on YTF dataset.

has achieved comparable performances as the state-of-the-art methods. After feeding the new transformed Grassmann manifold learned by PML, GDA and GGDA consistently gain improvement for video based face verification on the two datasets. Note that on PaSC, we extract the state-of-the-art DCNN feature and find most of the comparative set-based methods significantly outperform (our method has a significant gain of around 18% above) the state-of-the-art methods that were evaluated in [4], where the best performance is 26% in the handheld experiment.

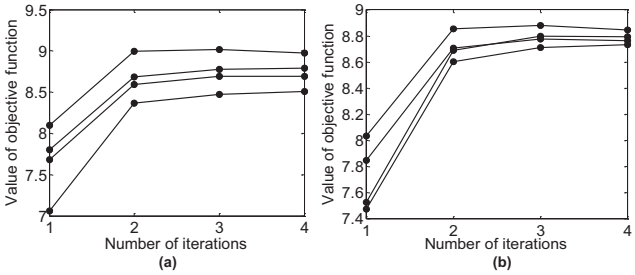

Figure 6. Convergence characteristics of the optimization algorithm of the proposed method: (a) depicts the values of our objective function varying with different number of iterations on 4 folds of YTF. (b) shows the convergence to a close maximum with 4 different random initialization on one of 10 folds of YTF.

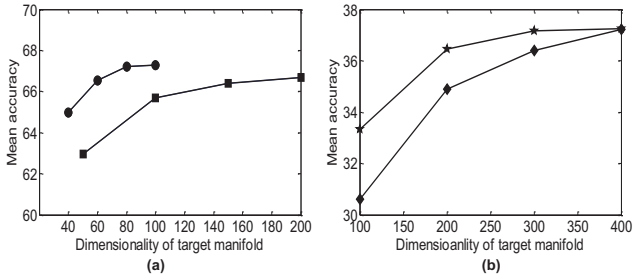

Figure 7. Mean accuracies of the proposed method with different dimensionalities of the target manifold: (a) The curve with squares is for YTC while the one with circles is for YTF. (b) The curve with stars is for PaSC-Control while the one with diamonds is for PaSC-handheld.

Methods	CMSM	SSDML	DCC	GDA	GGDA	**PML**
Train	22.25	30.03	23.15	974.14	1306.13	33.71
Test	9.25	23.15	8.87	278.71	420.17	9.44

Table 4. Average computation time (seconds) of different methods on the YTF dataset for training and testing (verification for one pair of comparing videos).

5.3. Discussion

Although we do not offer a proof of convergence or uniqueness of the proposed optimization algorithm, its convergence to a global maximum is confirmed experimentally. Fig.6 (a) shows some examples of the learning iteration. The examples are for the learning using 4 folds on the YTF dataset. The value of objective function J for all cases converges to a stable value after a few iterations starting with the initial value $\boldsymbol{P} = \boldsymbol{I}$. On one of the four folds (see Fig.6 (a)), the values of cost function from 1 to 10 iterations and the value at 100 iteration are: 7.69, 8.47, 8.57, 8.58, 8.59, 8.59, 8.58, 8.58, 8.59, 8.58 and 8.58, which shows our method can converge to a stable value with more iterations. In addition, we also test our method after only 1 iteration, and the results are 64.27% (YTC); 66.34% (YTF); 35.47%, 35.98% (PaSC), which shows the initial solution has not reached the best performance. Furthermore, as shown in Fig.6 (b), we observe that the proposed optimization algorithm converges to very close values irrespective of the initial value of $\boldsymbol{P}$.

As our PML method can be considered as a dimensionality reduction technique on Grassmann manifold, we also care about the impact of the setting of the dimensionality (i.e., d in this paper) of the target Grassmann manifold. Therefore, we compare the results of our methods with different d on YTC, YTF and PaSC. As shown in Fig.7, we find that the impact of d on our method tends to be mild when it is large enough. The accuracies with the last setting of d on each dataset are all the best and are used as the final results respectively reported in Tab.1, Tab.2, Tab.3.

Lastly, we also compare the running time of several competing methods on the YTF dataset. Tab. 4 lists their training time and testing time on a 3.40GHz PC. As can be seen from this table, our proposed method and CMSM, SSDML, DCC are much faster than GDA and GGDA, which both need to calculate the kernel matrices and thus are time expensive. We also provide the running time of PML-GDA (Train: 932.04, Test: 195.10) and PML-GGDA (Train: 1076.33, Test: 226.58), which shows our proposed method can speed up the original methods.

6. Conclusion

We have introduced a novel discriminant analysis algorithm on the Grassmann manifold to tackle the problem of video based face recognition. Specifically, we exploited a Fisher LDA-like framework to learn the Projection Metric by mapping data from the original Grassmann manifold to a new more discriminant one. Our new approach can not only serve as a metric learning method but also a dimensionality reduction technique for the Grassmann manifold. Our experimental evaluation has demonstrated that the new technique and its coupling with other methods lead to state-of-the-art recognition accuracies on several challenging datasets for video based face identification/verification.

We believe our work is among the first attempts towards showing the importance of preserving the Riemannian structure of the Grassmann manifold when performing metric learning or dimensionality reduction. In the future, we plan to study how to improve this framework with other metrics such as Binet-Cauchy metric on Grassmann manifold. It may be difficult for exploiting Binet-Cauchy metric in our proposed framework because the dimensionality of its embedding space grows exponentially. Nevertheless, it would be very interesting to further explore this field.

Acknowledgements

This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61222211, 61379083, and 61390511.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[3] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *BTAS*, 2013.

[4] J. R. Beveridge, H. Zhang, P. J. Flynn, Y. Lee, V. E. Liong, J. Lu, M. d. A. Angeloni, T. d. F. Pereira, et al. The IJCB 2014 PaSC video face and person recognition competition. In *IJCB*, 2014.

[5] H. E. Cetingul and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *CVPR*, 2009.

[6] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, 2010.

[7] S. Chen, C. Sanderson, M. T. Harandi, and B. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *CVPR*, 2013.

[8] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *CVPR*, 2013.

[9] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

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

[11] J. Hamm and D. D. Lee. Extended grassmann kernels for subspace-based learning. In *NIPS*, 2008.

[12] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *ICML*, 2008.

[13] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices. In *ECCV*. 2014.

[14] M. T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li. Expanding the family of grassmannian kernels: An embedding perspective. In *ECCV*. 2014.

[15] M. T. Harandi, C. Sanderson, C. Shen, and B. Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *ICCV*, 2013.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, 2011.

[17] M. Hayat, M. Bennamoun, and S. An. Reverse training: An efficient approach for image set classification. In *ECCV*, 2014.

[18] U. Helmke, K. Hüper, and J. Trumpf. Newton's method on grassmann manifolds. *arXiv preprint arXiv:0709.2205*, 2007.

[19] Z. Huang, R. Wang, S. Shan, and X. Chen. Learning euclidean-to-riemannian metric for point-to-set classification. In *CVPR*, 2014.

[20] Z. Huang, R. Wang, S. Shan, and X. Chen. Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning. *Pattern Recognition*, 2015.

[21] S. Jain and V. Govindu. Efficient higher-order clustering on the grassmann manifold. In *ICCV*, 2013.

[22] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. T. Harandi. Optimizing over radial kernels on compact manifolds. In *CVPR*, 2014.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv: 1408.5093*, 2014.

[24] S. Jung, I. Dryden, and J. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, 2012.

[25] M. Kan, S. Shan, D. Xu, and X. Chen. Side-information based linear discriminant analysis for face recognition. In *BMVC*, 2011.

[26] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008.

[27] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE T-PAMI*, 29(6):1005–1018, 2007.

[28] H. Le. On geodesics in euclidean shape spaces. *J. Lond. Math. Soc.*, pages 360–372, 1991.

[29] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, 2013.

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

[31] D. Pham and S. Venkatesh. Robust learning of discriminative projection for multicategory classification on the stiefel manifold. In *CVPR*, 2008.

[32] A. Srivastava and E. Klassen. Bayesian and geometric subspace tracking. *Advances in Applied Probability*, pages 43–56, 2004.

[33] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *JMLR*, pages 1027–1061, 2007.

[34] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE T-PAMI*, 33(11):2273–2286, 2011.

[35] K. Varshney and A. Willsky. Learning dimensionality-reduced classifiers for information fusion. In *ICIF*, 2009.

[36] R. Vemulapalli, J. Pillai, and R. Chellappa. Kernel learning for extrinsic classification of manifold features. In *CVPR*, 2013.

[37] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, 2009.

[38] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, 2012.

[39] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifold-Manifold distance and its application to face recognition with image sets. *IEEE T-IP*, 21(10):4466–4479, 2012.

[40] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[41] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *JMLR*, 4:913–931, 2003.

[42] Y.-C. Wong. Differential geometry of grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 57(3):589, 1967.

[43] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *FG*, 1998.

[44] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *IEEE Trans. on Multimedia*, 14(4):995–107, 2012.

[45] P. Zhu, L. Zhang, W. Zuo, and D. Zhang. From point to set: Extend the learning of distance metrics. In *ICCV*, 2013.