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

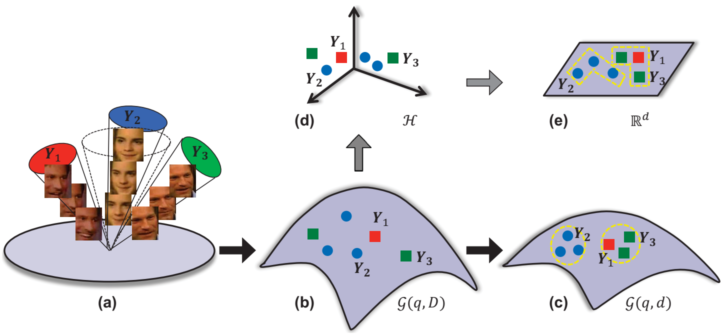

Figure 1: Conceptual illustration of the proposed Projection Metric Learning (PML) on the Grassmann Manifold. Traditional Grassmann discriminant analysis methods take the away (a)-(b)-(d)-(e) to first embed the original Grassmann manifold $\mathcal{G}(q,D)$ (b) into high dimensional Hilbert space $\mathcal{H}$ (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).

In video based face recognition, great success has been made by representing videos as linear subspaces, which typically reside on Grassmann manifold endowed with the well-studied projection metric. Under the projection metric framework, most of recent studies [1, 2, 3, 4, 5] exploited a series of positive definite kernel functions on Grassmann manifold to first embed the manifold into a high dimensional Hilbert space, and then map the flattened manifold into a lower-dimensional Euclidean space (see Fig.1 (a)-(b)-(d)-(e)). 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.

To overcome the limitations of existing Grassmann discriminant analysis methods, by endowing the well-studied Projection Metric with Grassmann manifold, this paper 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)).

Formally, assume m video sequences 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 in the original paper, we 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

$$f(\boldsymbol{Y}_i \boldsymbol{Y}_i^T) = \boldsymbol{W}^T \boldsymbol{Y}_i \boldsymbol{Y}_i^T \boldsymbol{W} = (\boldsymbol{W}^T \boldsymbol{Y}_i)(\boldsymbol{W}^T \boldsymbol{Y}_i)^T. \quad (1)$$

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 give more details in the original paper. Here, we briefly describe the formulation of 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:

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

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 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:

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} J(\boldsymbol{P}) = \arg\min_{\boldsymbol{P}} (J_w(\boldsymbol{P}) - \alpha J_b(\boldsymbol{P})). \quad (3)$$

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:

$$J_w(\boldsymbol{P}) = \frac{1}{N_w} \sum_{i=1}^{m} \sum_{j: C_i = C_j} 2^{-1/2} tr(\boldsymbol{P} \boldsymbol{A}_{ij} \boldsymbol{A}_{ij}^T \boldsymbol{P}). \quad (4)$$

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

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.

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

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

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

[4] Mehrtash Tafazzoli Harandi, Mathieu Salzmann, Sadeep Jayasumana, Richard Hartley, and Hongdong Li. Expanding the family of grassmannian kernels: An embedding perspective. In *ECCV*. 2014.

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.