# Joint Normalization and Dimensionality Reduction on Grassmannian: A Generalized Perspective

Tianci Liu, Zelin Shi, and Yunpeng Liu

***Abstract*—This letter proposes a generalized framework with joint normalization that learns lower dimensional subspaces with maximum discriminative power by using Riemannian geometry. We model the similarity/dissimilarity between subspaces using various metrics defined on Grassmannian and formulate dimensionality reduction as a nonlinear constraint optimization problem considering the orthogonalization. To obtain the linear mapping, we derive the components required to perform Riemannian optimization from the original Grassmannian through an orthonormal projection. We respect the Riemannian geometry of the Grassmann manifold and search for this projection directly from one Grassmann manifold to another face-to-face without any additional transformations. In this natural geometry-aware approach, any metric on the Grassmann manifold can theoretically reside in our model. We combine five metrics with our model, and the learning process is treated as an unconstrained optimization problem on a Grassmann manifold. Experiments on several datasets demonstrate that our approach leads to a significant accuracy gain over state-of-the-art methods.**

***Index Terms*—Dimensionality reduction, image-set recognition, Grassmann manifold, Grassmannian optimization.**

## I. Introduction

MODELING videos and image-sets by linear subspaces is beneficial in various visual recognition tasks. However, subspaces constructed from visual data are notoriously high dimensional, which in turn limits the applicability of existing techniques. Thus, the emergence of a dimensionality reduction method designed for subspaces to learn a low-dimensional and more discriminative space is extremely urgent. Furthermore, subspaces in visual data have a rigorous geometry that should be considered with the corresponding method of dimensionality reduction. Linear subspaces with the same dimensionality reside on a special type of Riemannian manifold, i.e., the Grassmann manifold, which has a nonlinear structure. Conventional methods, such as principal component analysis (PCA) [13] and linear discriminant analysis [16], are devised for vectors in the flat Euclidean space instead of the curved Riemannian space. Simply applying these methods to subspaces may create distortions in the geometry. In this context, a natural question arises: How can popular dimensionality reduction techniques be extended to subspaces with Riemannian geometry?

In response to this issue, this letter proposes a manifold-to-manifold method to learn a low-dimensional and more discriminative Grassmann manifold in a generalized framework, which can be regarded as a geometry-aware dimensionality reduction of the Grassmann manifold (as shown in Fig. 1). Noted that our framework is suitable for any metric on the Grassmann manifold rather than being limited to the projection framework as [14], [25], and [27]. The main contributions of this letter are listed as follows.

1) We propose a method with the orthonormal constraint that learns a low-dimensional space of the Grassmann manifold from the high dimensional one, which maximizes the discriminative power of the classification.
2) We propose a generalized Grassmannian framework that is more extended and complete than other similar models [14], [25]. Our model is more flexible and available for various metrics on the Grassmannian.
3) We model the dimensionality reduction as an optimization problem on Grassmannian with joint normalization considering the orthogonalization. This guarantees the reduced matrices endowed with the Grassmannian geometry in each iteration.

## II. Joint Normalization and Dimensionality Reduction on Grassmannian

### A. Proposed Method

We aim to learn a mapping *W*, which can map the Grassmann manifold in high dimensionality $G(n, D)$[1] to a lower one, $G(n, d)$, for better classification ($D \gg d$). Specifically, assume that $X \in G(n, D)$, $Y \in G(n, d)$ with ($D \gg d$). We aim to find a column full-rank matrix $W \in \mathbb{R}^{D \times d}$ such that a general mapping $f : G(n, D) \rightarrow G(n, d)$ can be learned

$$f(X) = W^T X = Y. \quad (1)$$

Manuscript received February 26, 2018; revised April 16, 2018; accepted April 17, 2018. Date of publication April 20, 2018; date of current version May 7, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marco Felipe Duarte. *(Corresponding author: Tianci Liu.)*

T. Liu is with the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Key Laboratory of Opto-Electronic Information Processing, CAS, Shenyang 110016, China (e-mail: liutianci@sia.cn).

Z. Shi and Y. Liu are with the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China, and also with the Key Laboratory of Opto-Electronic Information Processing, CAS, Shenyang 110016, China (e-mail: 582507583@qq.com; 1099149944@qq.com).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2829106

[1] A Grassmann manifold *G*(*n*, *D*), where $n < D$, is the set of *n*-dimensional linear subspaces in the *D*-dimensional Euclidean space $\mathbb{R}^D$, which is the compact Riemannian manifold with *n*(*D*-*n*) dimensionality. The defination of *G*(*n*, *d*) is similar.

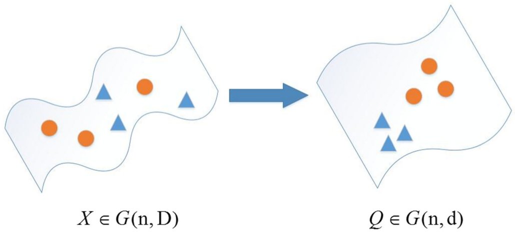

$X \in G(\mathrm{n}, \mathrm{D})$
$Q \in G(\mathrm{n}, \mathrm{d})$

Fig. 1. Conceptual illustration of the proposed supervised dimensionality reduction method on the Grassmann manifold. The new Grassmannian leads to a lower dimensional and more discriminative space.

This mapping can be regarded as a special type of dimensionality reduction of the Grassmann manifold, but we aim to find a new low-dimensional geometry in which two image-sets are close to each other if they belong to the same class and far apart if they do not. That is, given a set of image sets $\Gamma = \{X_1, X_2, \ldots, X_N\}$, this geometry is structured by the affinity matrix $G \in R^{N \times N}$, which is undirected and symmetric. Each element $G(i,j)$ reflects the pairwise relationship of the *i*th image set and the *j*th one. The detailed notation of *G* will be provided in Section III.

When the affinity matrix *G* is given, we use different metrics to encode this structure into the low-dimensional manifold. Several prevalent Grassmannian distances that have many applications in [6], [9], and [10] are introduced in Table I. For this purpose, the objective function has the corresponding formula

$$L(W) = \sum_{i,j} G(i,j) \cdot d(W^T X_1, W^T X_2) \tag{2}$$

where $d : M \times M \to R^+$ represents the metrics in Table I.

As the metrics in (2) are subspace distances, $W^T X$ should be on the Grassmannian accordingly. However, $W^T X$ is not guaranteed to be on the Grassmann manifold even if *W* is an orthogonal matrix. Note that only the linear subspaces spanned by an orthonormal basis matrix can form a valid Grassmann manifold. To solve this issue, we use *QR*-decomposition to obtain the orthonormal components of $W^T X$ s.t. $W^T X = QR$, where *Q* is the orthonormal matrix consisting of the first $d$ columns and *R* is the invertible upper-triangular matrix.

*1) Joint Normalization of Y:* As shown above, considering the orthogonalization, the cost function will result in the following formulation:

$$L(W) = \sum_{i,j} G(i,j) \cdot d(Q_i, Q_j) \tag{3}$$

where

$$Q_i = W^T X_i R_i^{-1}, Q_j = W^T X_j R_j^{-1}. \tag{4}$$

To avoid degeneracies of optimization and follow common practice in dimensionality reduction, we impose an orthogonality constraint on *W* such that $W^T W = I_d$. Finally, dimensionality reduction is written as the optimization problem that seeks the solution to *W* by minimizing the cost function (3)

$$W^* = \arg\min_W L(W) \quad \text{s.t. } W^T W = I_d. \tag{5}$$

From a mathematical perspective, the search space of *W* is on the Stiefel manifold if the minimization problem $L(W)$ has the orthogonality constraints, i.e., $W^T W = I_d$. Moreover, the objective function is invariant to the orthogonal group, i.e., $L(W) = L(WH)$ for any $H \in O(n)$, which means that (3) is independent of the choice of basis spanned by $W$. Thus, (5) can be solved as an unconstrained minimization problem on *G*(*d*, *D*).

### *B. Optimization on the Riemannian Manifold*

In practice, the solution to (5) can be sought through the conjugate gradient method [2], [6] on *G*(*d*, *D*), which is implemented in the ManOpt toolbox [3]. This nonlinear method essentially requires the gradient on the Riemannian manifold. The Riemannian gradient on *G*(*d*, *D*) can be computed as

$$R_W L(W) = (I_D - WW^T)\nabla_W L(W) \tag{6}$$

where $\nabla_W L(W)$ is the Euclidean gradient of $L(W)$ with respect to W, which is the Jacobian matrix of size *D* by *d*. In Section II-C, the detailed derivations of $\nabla_W L(W)$ under different metrics are described.

### *C. Computing the Gradient*

Let $X_1, X_2 \in G(n, D)$ be two points on the Grassmannian *G*(*n*, *D*) and $Y_1 = W^T X_1, Y_2 = W^T X_2$ are the corresponding transformed matrices with $Y_1, Y_2 \in \mathbb{R}^{d \times n}$, such that $Y_1 = Q_1 R_1, Y_2 = Q_2 R_2$ are obtained by *QR* decomposition. From the perspective of matrix backpropagation [15], the gradient computation of $L(W)$ can be split into four steps

$$W \to (Y_1, Y_2) \to (Q_1, Q_2) \to L.$$

We consider them in reverse order from the objective down to the inputs. In the first part, the derivative of the objective function $L$ (i.e., $\partial L/\partial Q_1, \partial L/\partial Q_2$) can be calculated by the matrix chain rule. Then, we focus on the part receiving $Y_1$ or $Y_2$ as inputs and producing the corresponding orthogonal components (i.e., $Q_1$ or $Q_2$). These derivatives can be obtained with the application of Proposition 1. Finally, for the part taking $W$ as an input and producing $Y_1$ or $Y_2$, the derivatives are computed by the matrix chain rule (i.e., $\frac{\partial L \circ f}{\partial W} = \frac{\partial L}{\partial Y_1} + \frac{\partial L}{\partial Y_2}$).

Subsequently, the derivatives $\nabla_W L(W)$ under different metrics are obtained as follows:

*Remark 1: (Variations under the projection metric).* The resulting partial derivative of $L(W)$ is

$$\begin{aligned} \frac{\partial L \circ f}{\partial W} = X_1 {R_1}^{-1} \Bigg( & \left(I - Q_1 {Q_1}^T\right)^T \frac{\partial L}{\partial Q_1} \\ & + Q_1 \left( {Q_1}^T \frac{\partial L}{\partial Q_1} \right)_{btril} \Bigg)^T \\ + X_2 {R_2}^{-1} \Bigg( & \left(I - Q_2 {Q_2}^T\right)^T \frac{\partial L}{\partial Q_2} \\ & + Q_2 \left( {Q_2}^T \frac{\partial L}{\partial Q_2} \right)_{btril} \Bigg)^T \end{aligned} \tag{7}$$

where $\frac{\partial L}{\partial Q_1} = 2(Q_1 - Q_2 Q_2^T Q_1)$, $\frac{\partial L}{\partial Q_2} = 2(Q_2 - Q_1 Q_1^T Q_2)$.

TABLE I
DIFFERENT MEASURES ON THE GRASSMANN MANIFOLD

| **Measure Name** | **Mathematical Expression** | **Metric/ Distance** | **Kernel** |
|---|---|---|---|
| projection F-norm [6] | $d_{pro}(X_1, X_2) = 2^{-\frac{1}{2}} \left\| X_1 {X_1}^T - X_2 {X_2}^T \right\|_F = \|\sin(\Theta)\|_2$ | √ | × |
| Fubini-Study [6] | $d_{FS}(X_1, X_2) = arccos \left| \det({X_1}^T X_2) \right| = \arccos(\prod_i \cos\theta_i)$ | √ | × |
| Binet-Cauchy distance [10] | $d^2_{\mathrm{BC}}(X_1, X_2) = 2 - 2 \left| \det({X_1}^T X_2) \right| = 2 - 2 \prod_i (1 - sin^2 \theta_i)$ | √ | × |
| projection kernel distance [10] | $d^2_{\mathrm{pk}}(X_1, X_2) = 2\mathrm{n} - 2 \left\| {X_1}^T X_2 \right\|^2_F = 2n - 2 \sum_i \cos^2(\theta_i)$ | √ | × |
| Binet-Cauchy kernel [9] | $d^2_{\mathrm{BCK}}(X_1, X_2) = \det({X_1}^T X_2 {X_2}^T X_1) = \prod_i \cos^2(\theta_i)$ | × | √ |

$X_1$, $X_2$ are two points on the Grassmannian $G(n, D)$.

TABLE II
AVERAGE RECOGNITION RATES AND PARAMETER SETTINGS ON DIFFERENT DATASETS

| **Method** | NN-P | NN-FS | NN-PK | NN-BC | NN-BCK | GGDA | GDL | PML | **P-DR** | **FS-DR** | **PK-DR** | **BC-DR** | **BCK-DR** | **GGDA-DR** | **GDL-DR** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | ETH-80: 80 samples (40 test samples and 40 training samples); Parameters: n=3, D=400, d=23 | | | | | | | | | | | | | | |
| **Results** | 90 | 85 | 90 | 85 | 85 | 90 | 92.5 | 92.5 | **97.5** | 95 | **97.5** | 95 | 92.5 | **97.5** | **97.5** |
| **Dataset** | Extended Yale B: 380 samples (190 test samples and 190 training samples); Parameters: n=4, D=400, d=60 | | | | | | | | | | | | | | |
| **Results** | 74.21 | 43.16 | 74.21 | 43.16 | 43.16 | 93.68 | 95.79 | 94.21 | **1** | 92.11 | **1** | 91.58 | 88.95 | **1** | **1** |
| **GPCA** | 60 | 38.95 | 60 | 38.95 | 38.95 | 86.32 | 90.53 | – | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **Dataset** | JHMDB: 2730 samples (1260 test samples and 1470 training samples); Parameters: n=6, D=4096, d=500 | | | | | | | | | | | | | | |
| **VGG** | 52.38 | 36.51 | 52.38 | 36.51 | 36.51 | 48.54 | 52.75 | 50.79 | 55.56 | 53.97 | 58.73 | 46.03 | 42.86 | 57.62 | **60.32** |
| **GPCA** | 46.03 | 34.92 | 46.03 | 34.92 | 34.92 | 45.16 | 47.62 | – | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **State-of-the-art** | Avg. Pooling [21]: 55.9 | | | Rank Pooling [7]: 55.2 | | | P-CNN [4]: 61.1 | | | Action Tubes [8]: 62.5 | | | IDT + FV [26]: 62.8 | | |

For the other metrics, the form of $\frac{\partial L \circ f}{\partial W}$ is the same as (7). Thus, these forms are omitted due to space constraints.

*Remark 2: (Variations under the Fubini-study metric).* The resulting partial derivative of $L(W)$ is

$$\frac{\partial L}{\partial Q_1} = Q_2 \left( \frac{\partial L}{\partial A} \right)^T, \frac{\partial L}{\partial Q_2} = Q_1 \frac{\partial L}{\partial A}$$

where $\frac{\partial L}{\partial A} = \frac{-1}{\sqrt{1 - |\det(Q_1^T Q_2)|^2}} |\det(Q_1^T Q_2)| (Q_2^T Q_1)^{-1}$.

*Remark 3: (Variations under the Binet–Cauchy distance).* The resulting partial derivative of $L(W)$ is

$$\frac{\partial L}{\partial Q_1} = Q_2 \left( \frac{\partial L}{\partial A} \right)^T, \frac{\partial L}{\partial Q_2} = Q_1 \frac{\partial L}{\partial A}$$

where $\frac{\partial L}{\partial A} = -2 \left| \det(Q_1^T Q_2) \right| \left( Q_2^T Q_1 \right)^{-1}$.

*Remark 4: (Variations under the projection kernel distance).* The resulting partial derivative of $L(W)$ is

$$\frac{\partial L}{\partial Q_1} = -4 Q_2 Q_2^T Q_1, \frac{\partial L}{\partial Q_2} = -4 Q_1 Q_1^T Q_2.$$

*Remark 5: (Variations under the Binet–Cauchy kernel).* The resulting partial derivative of $L(W)$ is

$$\frac{\partial L}{\partial Q_1} = Q_2 B^T \left( \frac{\partial L}{\partial A} + \left( \frac{\partial L}{\partial A} \right)^T \right),$$

$$\frac{\partial L}{\partial Q_2} = Q_1 \left( \frac{\partial L}{\partial A} + \left( \frac{\partial L}{\partial A} \right)^T \right) B$$

where

$$\frac{\partial L}{\partial A} = \det(B B^T) (B B^T)^{-1}, B = Q_1^T Q_2.$$

## III. DEFINING THE AFFINITY MATRIX

Let $y_i$ denote the class label of the image set $X_i$, with $1 \leq y_i \leq C$. Each element of the affinity matrix is expressed as

$$G(i,j) = G_w(i,j) - G_b(i,j) \tag{8}$$

where $G_w$ is the within-class similarity graph and $G_b$ is the graph to measure the between-class similarity. Equation (8) resembles the maximum margin criterion of [20].

$G_w$ and $G_b$ are defined as

$$G_w(i,j) = \begin{cases} 1, & \text{if } X_i \in N_w(X_j) \text{ or } X_j \in N_w(X_i) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$G_b(i,j) = \begin{cases} 1, & \text{if } X_i \in N_b(X_j) \text{ or } X_j \in N_b(X_i) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $N_w(X_i)$ consists of $k_w$ neighbors that belong to the same label as $X_i$ and $N_b(X_i)$ is the set of $k_b$ neighbors that own different labels from $X_i$. In practice, $k_w$ is defined as the minimum number of points in each class and the value of $k_b \leq k_w$ is set by cross validation to balance the relationship between $G_w$ and $G_b$.

## IV. EXPERIMENTS

### A. Experimental Settings

In our experiments, each image set is represented in the matrix form as $X_i = (x_1, x_2, x_3, \ldots, x_n)$, where $x_i \in R^D$ corresponds to the vectorized feature of the *i*th frame. We model the linear subspace of $X_i$ as an element on the $G(n, D)$ by preserving its first $n$ singular-vectors by SVD. In theory, $W$ can be initialized as a random $D \times d$ orthogonal matrix. In implementation, we simply initialize $W$ as the $D \times d$ identity matrix. All algorithms used in our experiments are referenced as follows.

1) NN-P/FS/PK/BC/BCK: NN classifier based on the $d_{\mathrm{pro}}$/$d_{\mathrm{FS}}$/$d_{\mathrm{pk}}$/$d_{\mathrm{BC}}$/$d_{\mathrm{BCK}}$ metric on the original manifold.

2) P/FS/PK/BC/BCK-DR: NN classifier with different metric on the low-dimensional Grassmann manifold.
3) GGDA [12]/GGDA-DR: Graph-embedding Grassmann discriminant analysis/on the learning manifold.
4) GDL [11]/GDL-DR: Grassmann dictionary learning/on the learning manifold with our method.
5) PML [14]: Projection metric learning on Grassmannian.
6) GPCA/VGG: Subspaces with low-dimensional/high-dimensional CNN features obtained by PCA/VGG.

### B. Datasets

*The ETH-80 dataset* [18] contains 3280 images of eight object categories. Each image is resized to the size of 20 × 20. We generate 80 points on $G(3{,}400)$ based on the gray feature.

*The Extended Yale B dataset* [19] contains 2432 face images of 38 human subjects under 64 different illumination conditions. All images used are front face images, which are manually aligned, cropped, and then resized to a size of 20×20. We generate Grassmann points based on the gray feature.

*JHMDB dataset* [17] is a challenging dataset on activity recognition with more variations in scenes and viewpoints, which can be used to examine the robustness of the proposed methods in noisy scenarios. We employ the Matconvnet [24] to extract CNN features on the FC6[2] layer of the VGG-16 Net [22]. The model is pretrained on Imagenet [5], and then fine-tuned on the data from the training sets of JHMDB and UCF101 [23] datasets. The performances of state-of-the-art methods presented in recent literature on this dataset are also provided Table II.

Experimental results for the classification tasks are shown in Table II. Noted that the blue numbers represent the results that obtain the best improvement by our method and the pink numbers represent the best classification accuracies in each dataset. In the original manifold, NN-P and NN-PK are better than the other three metrics by a large margin. However, after the dimensionality reduction, the results under five metrics reach a competitive level enhanced by our method. Both the GDL-DR and GGDA-DR improve the original methods and outperform the other competing methods.

## V. Discussion

The general convergence analysis of the Newton method on Grassmannian has been theoretically described in [1]. Fig. 2 illustrates the typical convergence behavior of our method in the ETH-80 dataset. In practice, it is observed that the algorithm generally converges speedily in less than 25 iterations. Because we aim to learn a discriminative low-dimensional manifold $G(n, d)$ from $G(n, D)$, the impact of the setting of the reduced dimensionality (i.e., $d$) should also be considered. For this reason, the performances of our method with different $d$ on the Extended Yale B and JHMDB datasets are reported in Fig. 3. It is gratifying that our method delivers favorable performance with a desirable low dimensionality, and the impact of $d$ tends to be mild when $d$ is sufficiently large.

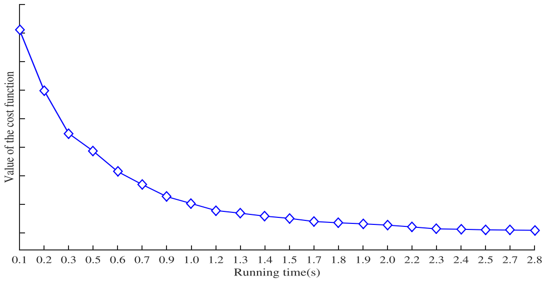
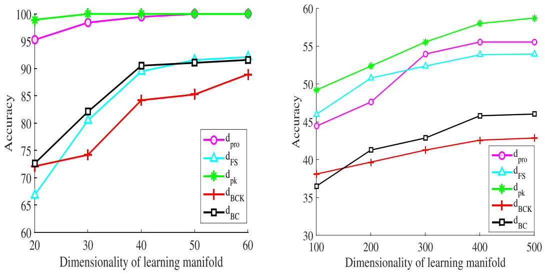


Fig. 2. Convergence behavior of our algorithm.



(a) (b)

Fig. 3. Averaged accuracies of the proposed method with different dimensionalities of the learning manifold. (a) Curves of five metrics for the Extended Yale B dataset. (b) Curves of five metrics for the JHMDB dataset.

TABLE III
Running Time (Seconds) Under Different Metrics

| **Metric** | $d_{pro}$ | $d_{FS}$ | $d_{pk}$ | $d_{BC}$ | $d_{BCK}$ |
|---|---|---|---|---|---|
| **NN** | 88.26 | 0.77 | 0.89 | 0.78 | 1.12 |
| **DR-NN** | 1.71 | 0.50 | 0.64 | 0.49 | 0.58 |

Finally, we compare the capabilities of the five metrics in Table I. From the experimental evaluations above, $d_{\text{pro}}$ and $d_{\text{pk}}$ yield better results than the remaining metrics possibly because these metrics compute the distances related to the cosine or sine of the principal angles between the subspaces. $d_{\text{pro}}$ and $d_{\text{pk}}$ are based on the accumulation operation, whereas the other metrics are based on multiplications. Luckily, these metrics lead to competitive results on the learning manifold by our method. We also compare the running time of our method under different metrics on the Extended Yale B dataset in Table III. The time cost of $d_{\text{pro}}$ is high because the computation complexity of $X_i X_j^T$ is greater than that of $X_i^T X_j$ when the subspace dimension is large. From the perspectives of the time cost and accuracy, $d_{\text{pk}}$ is a suitable choice for our model.

## VI. Conclusion and Future Work

We introduced a novel supervised algorithm to learn a low-dimensional and more discriminative Grassmann manifold from the original one under different metrics. Learning can be modeled as an optimization problem on a Grassmann manifold. Our new approach serves as not only a dimensionality reduction method but also a discriminant learning technique for the Grassmann manifold. Our experimental evaluation has demonstrated that the resulting low-dimensional Grassmann manifold leads to state-of-the-art recognition accuracies on several challenging datasets.

[2] We use the rectified output of the fully connected layer fc6 of the VGG-net, which are 4096 dimensional vectors, for all of the evaluations.

## REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, “Riemannian geometry of Grassmann manifolds with a view on algorithmic computation,” *Acta Applicandae Mathematicae*, vol. 80, no. 2, pp. 199–220, 2004.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[3] N. Boumal *et al.*, “Manopt, a matlab toolbox for optimization on manifolds,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1455–1459, 2014.

[4] G. Cheron, I. Laptev, and C. Schmid, “P-cnn: Pose-based CNN features for action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3218–3226.

[5] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[6] A. Edelman, T. A. Arias, and S. T. Smith, “The geometry of algorithms with orthogonality constraints,” *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

[7] B. Fernando, E. Gavves, M. J. Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 5378–5387.

[8] G. Gkioxari and J. Malik, “Finding action tubes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 759–768.

[9] J. Hamm and D. D. Lee, “Grassmann discriminant analysis: A unifying view on subspace-based learning,” in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 376–383.

[10] M. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li, “Expanding the family of Grassmannian kernels: An embedding perspective,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 408–423.

[11] M. Harandi, C. Sanderson, C. Shen, and B. C. Lovell, “Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3120–3127.

[12] M. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, “Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2705–2712.

[13] S. M. Holland, “Principal components analysis (PCA),” Dept. Geol., Univ. Georgia, Athens, GA, USA, pp. 30602–2501, 2008.

[14] Z. Huang, R. Wang, S. Shan, and X. Chen, “Projection metric learning on Grassmann manifold with application to video based face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 140–149.

[15] C. Ionescu, O. Vantzos, and C. Sminchisescu, “Matrix backpropagation for deep networks with structured layers,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2965–2973.

[16] A. J. Izenman, “Linear discriminant analysis,” in *Modern Multivariate Statistical Techniques*. New York, NY, USA: Springer, 2013, pp. 237–280.

[17] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, “Towards understanding action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3192–3199.

[18] T.-K. Kim and R. Cipolla, “Canonical correlation analysis of video volume tensors for action categorization and detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1415–1428, Aug. 2009.

[19] K. Lee, J. Ho, and D. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[20] X. R. Li, T. Jiang, and K. Zhang, “Efficient and robust feature extraction by maximum margin criterion,” *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.

[21] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Adv. Neural Inf. Process. Syst.*, vol. 1, no. 4, pp. 568–576, 2014.

[22] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Comput. Sci.*, 2014, arXiv:1409.1556.

[23] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *Comput. Sci.*, 2012, arXiv:1212.0402.

[24] A. Vedaldi and K. Lenc, “Matconvnet: Convolutional neural networks for matlab,” in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

[25] B. Wang *et al.*, “Locality preserving projections for Grassmann manifold,” in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2893–2900.

[26] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 3551–3558.

[27] Q. Wang, J. Gao, and H. Li, “Grassmannian manifold optimization assisted sparse spectral clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3145–3153.