LEARNING DISCRIMINANT GRASSMANN KERNELS FOR IMAGE-SET CLASSIFICATION

Lei Zhang[1,2]*, Wenhui Liu*[2]*, Xuezhi Xiang*[2]*, Yan Sun*[2]*, Xiantong Zhen*[3]

[1] Guangdong University of Petrochemical Technology, Guangdong, China

[2] College of Information and Communication Engineering, Harbin Engineering University, Harbin, China

[3] The department of Computer Science and Engineering,The University of Texas at Arlington, Texas, USA

ABSTRACT

Image-set classification has recently generated great popularity due to widespread application to challenging tasks in computer vision. The great challenges arise from measuring the similarity between image sets which usually exhibit huge inter-class ambiguity and intra-class variation. In this paper, based on the assumption that each image set as a linear subspace can be treated as a point on a Grassmann manifold, we propose discriminant Grassmann kernels (DGK) of principal angles between subspaces. To tackle the ambiguity and variation, we propose learning the DGK via kernel target alignment, which achieves kernels of great discrimination by maximizing correlations with class labels. The proposed DGK has been evaluated on two challenging datasets including the ETH-80 and UCSD datasets for object recognition and video-based traffic congestion recognition, respectively. Extensive experiments have shown that the proposed DGKs achieves state-of-the-art performance and surpasses most of previous methods, which demonstrates the great effectiveness of the DGKs for image-set classification.

Index Terms— Grassmann manifold, Partial Grassmann kernel, Discriminant Grassmann kernel, Kernel alignment, Image-set classification

1. INTRODUCTION

Recently, image-set classification has been widely researched in computer vision due to its widespread applications. Compared with traditional single image classification tasks, image-set classification can preferably handle the conditions with multi-view cameras or larger within-class divergence tasks. In practical applications, illumination changing, viewpoint variations and background distinction can introduce a wide range of appearance variations within images. [1] embeds structure information or [2] provides supervised descriptor learning to handle this problem. However, image set classification tackles this problem by treating a set of images with different appearances as one sample.

Thanks to National Science Foundation of China (61571147,61401113), National Science Foundation of Heilongjiang (F2015027,LC201426)

The central issue of image-set classification lie in the measurement of the similarity between image sets, which is the major difference from single-image classification, while the similarity usually relies on the representation of image sets. Recently, a large family of image-set representation methods model image sets as Gaussian distributions [3], Gaussian mixture models [4] or other kind of probability distributions [5] to depict property of image sets. Correspondingly, the similarity between image sets can be measured by the Kullback-Leibler (KL) distance or other similarity metrics between distributions. However, these methods can guarantee soundable performance only when training and test sets have significant statical relationship [6]. Another kind of image-set representation is obtained based on second-order statistics, i.e., covariance matrix, which is a point lying on a Riemannian manifold spanned by symmetric positive definite (SPD) matrices [7, 8]. Thus the similarity between image sets turns into how to compute the distance between two points on the Riemannian manifold, e.g., log-Euclidean distance (LED) [8] or a kernel function mapping the covariance matrix from Riemannian manifold to a Euclidean space [7]. One thing needs to note is that in covariance matrix representation approach, most approaches adopt intensity of original images as feature vectors, and these vectors from the same image set compose a matrix. In fact, there are other ways to compose this matrix for image set; for instance, the LBP and HOG feature can be extracted for each image in an image set [9] and the matrices with the same size are treated as points on the Grassmannian manifold, which is a special case of the Riemannian manifold. Other extended research on the Grassmann manifold [10, 11] aims to find a discriminative projection to improve the performance or extend the Grassmann manifold by the kernel trick [12]. Recently, with the great success of deep learning [13, 14], there are also research on deep construction model for image-set classification [15].

Most of those previous methods were developed under assumptions of specific distributions of image sets, which however would not be able to properly generalize to wider application. In this paper, we propose learning discriminant Grassmann kernels (DGK) for image-set classification, which directly matches image sets without relying on any specific assumptions. Based on partial kernels of principal angels,

the global kernels between image sets aggregate the partial kernels, which are learned by kernel alignment in a supervised learning framework, which achieves highly discriminative kernels between image sets. Different from [10, 11] which add discriminant analysis into Grassmann manifolds or the graph-embedding framework on Grassmann manifolds, the proposed DGK avoids complicated discriminant learning on Grassmann manifolds and directly learns discriminant kernels between image sets, which achieves improved performance to the state of the arts [10, 11] on for object recognition and traffic congestion classification on representative widely-used benchmark datasets.

2. GRASSMANN MANIFOLDS

The Grassmann manifold has been extensively studied and shown great effectiveness in representation of image sets. Principal angles between linear subspaces which are points on the Grassmann manifold has been commonly used for the similarity metric. Bases on the metrics, Grassmann kernels can be constructed for classification. The Grassmann kernels allow us to flexibly adopt the usual kernel based algorithms.

2.1. Linear subspaces on Grassmann manifolds

A subset of vectors $V \in \mathbb{R}^D$ is a linear subspace if and only if it satisfies three following conditions.

- Condition 1: V contains the zero vector.
- Condition 2: V is closure under scalar multiplication, that is, if $\mathbf{x} \in V$, then for any scalar c, $c\mathbf{x} \in V$.
- Condition 3: V is closure under addition, that is, if $\forall \mathbf{x} \in V$ and $\forall \mathbf{y} \in V$, then $\mathbf{x} + \mathbf{y} \in V$.

The above definition is easy to satisfy, thus for image set after feature extraction, the matrix, where each vector corresponds one image in the set, can be viewed as a linear subspace which corresponds to a point on the Grassmann manifold defined as follows.

Defination 1 *[10] The Grassmann manifold $\mathcal{G}(m, D)$ is the set of m-dimensional linear subspaces of the $\mathbb{R}^D$.*

In fact, if there is no restriction for the matrix representation, the points in the Grassmann manifold $G(m, D)$ may be not unique since two matrices M_1 and M_2 can be the different representations even from the same subspace. Two subspaces are considered to be the same if only if $\mathrm{span}(M_1) = \mathrm{span}(M_2)$, where $\mathrm{span}(\cdot)$ denotes the subspace spanned by the column vectors. In this work, the matrix representation is constructed based on orthogonalization and normalization, which guarantees its uniqueness.

2.2. Principal Angles

The Riemannian distance between two subspaces is the length of the shortest geodesic connecting the two points on the Grassmann manifold. However, principal angles are the intuitive and computationally efficient way of defining the distances [10]. After orthogonalization and normalization, the subspace M_1 is represented by a set of basis vector as $\{U : \mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m\}$ and the subspace M_2 as $\{V : \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}$. The principal angle between two subspaces is defined as below.

Defination 2 *[10] Principal angles $0 \leq \theta_1 \leq ... \leq \theta_m \leq \pi/2$ between two subspaces M_1 and M_2 in Grassmann manifold are defined recursively by*

$$\cos\theta_k = \max_{\mathbf{u}_k} \max_{\mathbf{v}_k} \mathbf{u}_k{}^\top \mathbf{v}_k \tag{1}$$

It is known that the principal angles are related to the geodesic distance by $d_G^2(M_1, M_2) = \sum_i \theta_i^2$. Based on principal angles, many distance metrics have been developed including the projection metric $\sum_i \sin(\theta_i^2)$ and the Binet-Cauchy metric $(1 - \prod_i \cos(\theta_i)^2)$ in [10].

In practice, for the matrix representations M_1 and M_2, the principal angles can be computed by singular value decomposition (SVD) as follow.

$$M_1^\top M_2 = U(\cos\Theta)V^\top \tag{2}$$

where $\cos\Theta = \mathrm{diag}(\cos\theta_1, ..., \cos\theta_m)$.

It can be seen from Eq. (2) that the cosines of the principal angles $\cos\theta_1, ..., \cos\theta_m$ can be viewed as canonical correlations between U and V. Furthermore, [16] reveals the cosines of the principal angles related to the classification error under moderate SNR and low SNR conditions. Since in real applications, high SNR is hard to satisfy, we adopt the cosines of the principal angles as the Grassmann manifold distance in our kernel learning. Therefore, the principal angles provide an effective and computationally efficient measurement of similarity between subspaces.

In what follows, we introduce our discriminant Grassmann kernel (DGK) based on the principal angles, which is learned in supervised way via kernel alignment.

3. NEW GRASSMANN KERNEL LEARNING

We first define the partial kernels between points on the Grassmann manifold with respect to the principal angles according to Definition 2 and then we construct the global kernels by aggregating the partial Grassman kernels, which are learned in a supervised way via kernel alignment.

3.1. Partial Grassmann kernels

Assume that we have two points: $\mathbf{z}_i$ and $\mathbf{z}_j$, i.e., two linear subspaces on the Grassmann manifold and we can compute

P principal angles between $\mathbf{z}_i$ and $\mathbf{z}_j$. We now define the partial Grassmann kernel $K^p \in \mathbb{R}^{N\times N}$ with respect to the p-th principal angle as follows:

$$[K^p]_{ij} = k^p(\mathbf{z}_i, \mathbf{z}_j) \tag{3}$$

where N is the total number of points on the Grassmann manifold and the kernel k^p is defined as

$$k^p(z_i, z_j) = \cos^2\theta_p \tag{4}$$

Since K^p only reflects partial relations between two points based on the p-th principal angles, we refer it as the partial Grassmann kernel. A straightforward way to combine the discriminant Grassmann kernels as a global kernel between two samples is to sum all the partial kernels together with equal weights, that is,

$$[K]_{ij} = \sum_{p=1}^{P} k^p \tag{5}$$

The obtained global kernel in Eq. (5) however does not distinguishes the different roles of principal angles which therefore tends to be less discriminative for classification tasks.

3.2. Discriminant learning by kernel alignment

To achieve highly discriminative global kernels between two subspaces, instead of adopting unit weight in Eq. (5), we propose learning the weight coefficients w in a supervised way by kernel alignment, which has shown great effectiveness in learning optimal combination of multiple kernels [17]. Then the final kernel function changes from Eq. (5) into Eq. (6).

$$[K_w]_{ij} = \sum_{p=1}^{P} w_p k^p \tag{6}$$

The core idea of kernel alignment aims to adjust the parameters of the kernel K_w between image sets in Eq. (6) to achieve the maximum correlations between target kernel and K_w, where the target kernel is $\mathbf{y}\mathbf{y}^\top$ with $\mathbf{y}$ as label information. The alignment between the global kernel and the target kernel can be written as the following optimization problem:

$$\max_{\mathbf{w}} \rho(\bar{K}_w, \mathbf{y}\mathbf{y}^\top) = \max_{\mathbf{w}} \frac{< \bar{K}_w, \mathbf{y}\mathbf{y}^\top >_F}{\| \bar{K}_w \|} \tag{7}$$

where $\bar{K}_w$ is centering kernel of K_w as

$$\begin{aligned}[\bar{K}_w]_{ij} =& [K_w]_{ij} - \frac{1}{m}\sum_{i=1}^{m}[K_w]_{ij} - \frac{1}{m}\sum_{j=1}^{m}[K_w]_{ij} \\ &+ \frac{1}{m^2}\sum_{i,j=1}^{m}[K_w]_{ij}\end{aligned} \tag{8}$$

which can be simply computed as

$$\bar{K}_w = U_m K_w U_m \tag{9}$$

where $U_m = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/m$.

The optimization problem in Eq. (7) can be reduced into solving a simple QP problem as follow:

$$\min_{\mathbf{v}\geq 0} \mathbf{v}^\top M\mathbf{v} - 2\mathbf{v}^\top\mathbf{a} \tag{10}$$

where $\mathbf{a} = (\langle\bar{K}_1, \mathbf{y}\mathbf{y}^\top\rangle_F, ..., \langle\bar{K}_m, \mathbf{y}\mathbf{y}^\top\rangle_F)^\top$ and $[M]_{kl} = \langle\bar{K}_k, \bar{K}_l\rangle_F$. Furthermore, the solution $\mathbf{w}^*$ of the alignment maximization problem in Eq. (7) is given by $\mathbf{w}^* = \mathbf{v}/\|\mathbf{v}\|$.

4. EXPERIMENTS AND RESULTS

We have evaluated the proposed discriminant Gaussian kernels (DGK) on two representative computer vision tasks for object recognition [18] and traffic congestion classification [19]. The proposed DGK can deliver state-of-the-art performance on both datasets, which demonstrates its effectiveness for image-set classification.

4.1. Datasets and settings

ETH dataset [18] contains 80 objects from 8 carefully chosen categories with high-resolution color images. These 8 categories covers both natural and artificial objects. Among them, apples, pears, tomatoes for 'fruits and vegetables' area; cows, dogs and horses for 'animal'; cups for 'graspable' and cars for the 'vehicles' supercategory. Each category contain 10 objects with 41 views per object and total images number is 3280. For each object, 5 instances are selected as gallery and the remaining five are used for probes. We average 5 times experiment results with randomly selections of the gallery and probes sets.

USCD traffic dataset [19] contains 254 video sequences of highway traffic in Seattle, collected from a single stationary traffic camera over two days. The database contains a variety of traffic patterns with different weather conditions(e.g. overcast, raining, sunny rain drops on the camera lens). Videos were recorded in a outdoor environment and each one lasted for $4 \sim 5$ seconds (with 10 color frames of 320×240 pixels per second). The database was labeled by hand with respect to the amount of traffic congestion in each sequence. Among them, 44 sequences of heavy traffic, 45 of medium traffic and 165 of light traffic. We follow the training and test settings shared in [5] for evaluation.

For the image representations, we adopt the histogram of oriented gradients (HOG) with 31-dimension as basic descriptors. Among these 31 dimensions, it covers 9 contrast insensitive orientation bins ($0 \sim \pi$) and 18 contrast sensitive orientation bins ($0 \sim 2\pi$) and 4 dimensions capturing the overall gradient energy in square blocks of four cells around the central pixel. We set the cell size to be 7×7. Note that

Discriminant Grassmann kernels (DGK)	**95.5%**
Grassmann kernels (Eq. (5))	**94.0%**
Kernel Fisher Discriminant (KFD) [21]	81.1%
Marginal Fisher Analysis (MFA) [22]	80.1%
Manifold-Manifold Distance (MMD) [23]	85.0%
Mutual Subspace Method (MSM) [24]	83.3%
Manifold Discriminant Analysis (MDA) [25]	89.0%
Discriminant Canonical Correlations (DCC) [6]	91.7%
Log-Euclidean metric learning (LEML) [8]	94.8%
Graph embedding discriminant analysis (GEDA) [11]	92.3%
Localized multi-kernel metric learning (LMKML) [26]	94.5%

Table 1. Performance comparison on the ETH dataset

Discriminant Grassmann kernel (DGK)	**92.1%**
Grassmann kernels (Eq. (5))	**90.5%**
Linear dynamical system (LDS) [27]	87.5%
Compressive sensing LDS (CS-LDS) [27]	89.1%
Grassmann discriminant analysis (GDA) [10]	92.5%
Covariance discriminative learning (CDL)[7]	91.7%
NN classifier on Hellinger distance [5]	91.3%
NN classifier on J-divergence [5]	91.0%
Discriminant analytic stationary subspace analysis (DASSA) [28]	91.7%
Discriminant non-linear stationary subspace analysis (DNLSSA+RBF kernel) [28]	94.5%

Table 2. Performance comparison on the UCSD dataset.

our DGK does not rely on any specific descriptors and other advanced descriptors could also be used. Once the kernel after alignment of both training and test sets are constructed, we adopt the support vector machine (SVM) [20] with the setting of pre-computed kernels to conduct training and test.

4.2. Set-based object recognition

The comparison results for object categorization results on the ETH dataset are reported in Table 1. The first block corresponds the proposed approach results. The upper block shows the comparison between the proposed discriminant Grassmann kernels (DGK) learned via kernel alignment and Grassmann kernels without learning. As can be seen, the performance has been largely improved from 94.0% to 95.5% by the proposed discriminant learning.

The proposed DGK achieves state-of-the-art performance which is better than most of the compared methods in Table 1. [25] combines local linear models with linear discriminative analysis, and proposed MDA approach. In [25], other approaches as Kernel Fisher Discriminant (KFD) [21], Marginal Fisher Analysis (MFA) [22], Manifold-Manifold Distance (MMD) [23], Mutual Subspace Method (MSM) [24] were also evaluated and the corresponding performance are also listed in Table 1. By maximizing the canonical correlations of within-class sets and minimizing canonical correlations of between-class sets, Discriminant Canonical Correlations (DCC) can achieve over 91% performance. [11] and [10] are similar to our approach, which suppose descriptors in image set as a point on the Grassmann manifold, while [10] adopted the project metric instead of metric on principal angles and then combined the idea of LDA into learning a mapping function on the Grassmann manifold; [11] further apply the idea of LDA into a graph-embedding framework on Grassmann manifold. The major difference of our approach from those two approaches lies in that we propose discriminant Grassmann kernels on cosine function of each principal angle and further combine them by kernel alignment, which avoids the complicated supervised discriminant learning and achieve improved performance. By combining different kernel matrix, [26] learns an adaptive weight to each local region in the kernel space, which produces competitive while lower performance than ours.

4.3. Video-based traffic congestion classification

The comparison results on video-based traffic congestion recognition are reported in Table 2. As also can be seen in the first block, the proposed DGK has largely improved the Grassmann kernels from 90.5% to 92.1%, which again demonstrates the great effectiveness of the proposed supervised learning framework via kernel alignment for image-set classification.

The proposed DGK also achieves state-of-the-art performance on this dataset. The NGK outperforms three representative methods including Grassmann discriminant analysis (GDA) [10], ovariance discriminative learning (CDL) [7] and NN classifiers on the Hellinger distance/J-divergence [5] by large margins. The results on this dataset again demonstrates the effectiveness of the DGK for image-set classification.

5. CONCLUSION

In this paper, we have presented discriminant Grassmann kernel (DGK) learning for image-set classification. The kernels between image sets are constructed on principal angles of linear subspaces. To achieve highly discriminative kernels, we propose discriminant learning of the Grasssmann kernels by kernel alignment. The obtained global kernel is directly aligned to the target kernel constructed by labels, which achieves kernels of great discrimination for improved performance. We have evaluated the DGK on two representative computer vision tasks for object recognition on the ETH dataset and traffic congestion classification on the UCSD dataset. The experimental results show that the proposed DGK achieves state-of-the-art performance and surpasses most of previous methods on both the datasets, which demonstrates its effectiveness for image-set classification.

References

[1] X.T. Zhen, L. Shao, D.C. Tao, and X.L. Li, "Embedding motion and structure features for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1182–1190, 2013.

[2] X.T. Zhen, Z.J. Wang, M.Y. Yu, and S. Li, "Supervised descriptor learning for multi-output regression," in *CVPR*, 2015, pp. 1211–1218.

[3] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: extend the learning of distance metrics," in *ICCV*, 2013, pp. 2664–2671.

[4] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, "Discriminant analysis on riemannian manifold of faussian distributions for face recognition with image set," in *CVPR*, 2015, pp. 2048–2057.

[5] M. Harandi, M. Salzmann, and M. Baktashmotlagh, "Beyond gauss: Image-set matching on the riemannian manifold of pdfs," in *ICCV*, 2015, pp. 4112–4120.

[6] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1005–1018, 2007.

[7] R. Wang, H. Guo, L.S. Davis, and Q. Dai, "Covariance discriminative learning: a natural and efficient approach to image set classification," in *CVPR*, 2012, pp. 2496–2503.

[8] Z. Huang, R. Wang, S.Shan, X. Li, and X. Chen, "Log-euclidean matric learning on symmetric positive define manifold with application to image set classification," in *ICML*, 2015, pp. 720–729.

[9] Arif Mahmood, Ajmal Mian, and Robyn Owens, "Semi-supervised spectral clustering for image set classification," in *CVPR*, 2014, pp. 121–128.

[10] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *ICML*, 2008, pp. 1–8.

[11] 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*, 2010, pp. 2705–2712.

[12] J. Hamm and D. D. Lee, "Extend grassmann kernels for subspace-based learning," in *NIPS*, 2008, pp. 1–8.

[13] X.T. Zhen, M.Y. Yu, F. Zheng, I. B. Nachum, M. Bhaduri, D. Laidley, and S. Li, "Multitarget sparse latent regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. DOI: 10.1109/TNNLS.2017.2651068, pp. 1–12, 2017.

[14] L. Zhang, Y. Feng, J.Q. Han, and X.T. Zhen, "Realistic human action recognition: When deep learning meets vlad," in *ICASSP*, 2016, pp. 1352–1356.

[15] Munawar Hayat, Mohammed Bennamoun, and Senjian An, "Deep reconstruction models for image set classification," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 37, no. 4, pp. 713–727, 2015.

[16] J. Huang, Q. Qiu, and R. Calderbank, "The role of principal angles in subspace classification," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1933–1945, 2015.

[17] C. Cortes, M. Mohri, and A. Rostamizadeh, "Two-stage learning kernel algoritnms," in *ICML*, 2010, pp. 239–246.

[18] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *CVPR*, 2003, pp. 409–415.

[19] Antoni B. Chan and Nuno Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," in *CVPR*, 2005, pp. 1–8.

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[21] S. Mika, G. Ratsch, and K.-R. Muller, "A mathematical programming approach to the kernel fisher algorithm," in *NIPS*, 2000, pp. 591–597.

[22] S. Yan, D. Xu, B. Zhang, and H.-J Zhang, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[23] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application based on image set," in *CVPR*, 2008, pp. 1–8.

[24] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *FG*, 1998, pp. 318–323.

[25] R. Wang and X. Chen, "Manifold discriminant analysis," in *CVPR*, 2009, pp. 429–436.

[26] 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, pp. 329–336.

[27] Aswin C. Sankaranarayanan, Pavan K. Turaga, Richard G. Baraniuk, and Rama Chellappa, "Compressive acquisition of dynamic scenes," in *Springer Berlin Heidelberg*, 2013, pp. 129–142.

[28] M Baktashmotlagh, M Harandi, B. C Lovell, and M Salzmann, "Discriminative non-linear stationary subspace analysis for video classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2353–2366, 2014.