Metric Learning with A-based Scalar Product for Image-set Recognition

Naoya Sogi[1] Lincon S. Souza[1] Bernardo B. Gatto[2] Kazuhiro Fukui[1,2]

[1]Graduate School of Science and Technology, University of Tsukuba

[2]Center for Artificial Intelligence Research (C-AIR), University of Tsukuba

{sogi,lincons,bernardo}@cvlab.cs.tsukuba.ac.jp kfukui@cs.tsukuba.ac.jp

Abstract

In this paper, we propose a metric learning method for image set recognition using subspace representation. The subspace representation is effective for image set recognition where each image set is compactly represented by a subspace in a high dimensional vector space. In this framework, the similarity between two given image sets is measured by the canonical angles between the two corresponding subspaces. Many types of methods utilizing the concept of canonical angles have been developed and studied extensively. However, there still remains large potential in improving the ability to measure canonical angles. Our key idea is to learn a general scalar product space (metric space) that produces more valid canonical angles between two subspaces. To realize this idea, we first introduce an $\mathbf{A}$*-based scalar product instead of the standard scalar product, where* $\mathbf{A}$ *is a symmetric positive definite matrix and the canonical angles between two subspaces are measured through the* $\mathbf{A}$*-based scalar product. We learn a discriminative metric space by optimizing metric* $\mathbf{A}$ *in terms of the Fisher ratio from local Fisher discriminant analysis. Besides, we introduce a mechanism to automatically reduce the dimension of the metric space by imposing a low-rank constraint on metric* $\mathbf{A}$*. The effectiveness of the proposed methods is validated through extensive classification experiments on three real-world datasets.*

1. Introduction

Image set recognition has become a fundamental approach for object recognition in the computer vision field, since it can work robustly against various variations, such as change of illumination condition or viewpoint [4, 13, 16, 18, 23, 27, 28, 33, 43, 42, 44, 47, 48]. An image set is usually represented by a model, such as a covariance matrix [20, 43], affine/convex hull [4], and subspace [13, 14, 15, 16, 23, 47], to simplify the problem of calculating the similarity between image sets to that of the distance between two models. Although various practical methods using each model have been proposed, we focus on the subspace representation, as it has a wide range of applications, and its effectiveness has been reported by many different studies from the theoretical and experimental side [3, 10, 11, 12, 19, 21, 32, 36, 39, 40, 47, 49].

In subspace-based methods, each image set is represented by a subspace in high dimensional vector space, where the subspace is generated by applying principal component analysis (PCA) to a set of images. Then, classification is performed by comparing input and reference subspaces. As fixed dimensional subspaces compose a Grassmann manifold, various classification methods utilizing the geometry of this manifold have been developed and studied extensively [13, 16, 19]. These Grassmann-based methods have achieved considerable results by designing a method to efficiently utilize similarity (or dissimilarity) defined by canonical angles [2, 17] between two subspaces measured in the vector space.

In these methods, canonical angles are calculated in a metric vector space with a simple scalar product, i.e., Euclidean space. Although this calculation is simple and easy to implement, there is still large room for the improvement of the representation ability in which an identity matrix is naively used as the metric matrix. To enhance the representation ability of canonical angles, we introduce a generalized concept of canonical angles with a scalar product space (also called a metric space) based on $\mathbf{A}$-based scalar product [24]. In our metric space, the canonical angles between two subspaces are measured through a scalar product defined by a symmetric positive definite matrix $\mathbf{A}$.

Then, we propose a method for learning a valid metric vector space for more discriminative canonical angles between two given subspaces by searching for a suitable $\mathbf{A}$. The conceptual diagram of the proposed methods is shown in Fig. 1. This learning method is equivalent to the deformation problem of the metric vector space, while fixing all the class subspaces. In this sense, the proposed method can be regarded as a kind of dual problem of learning subspace

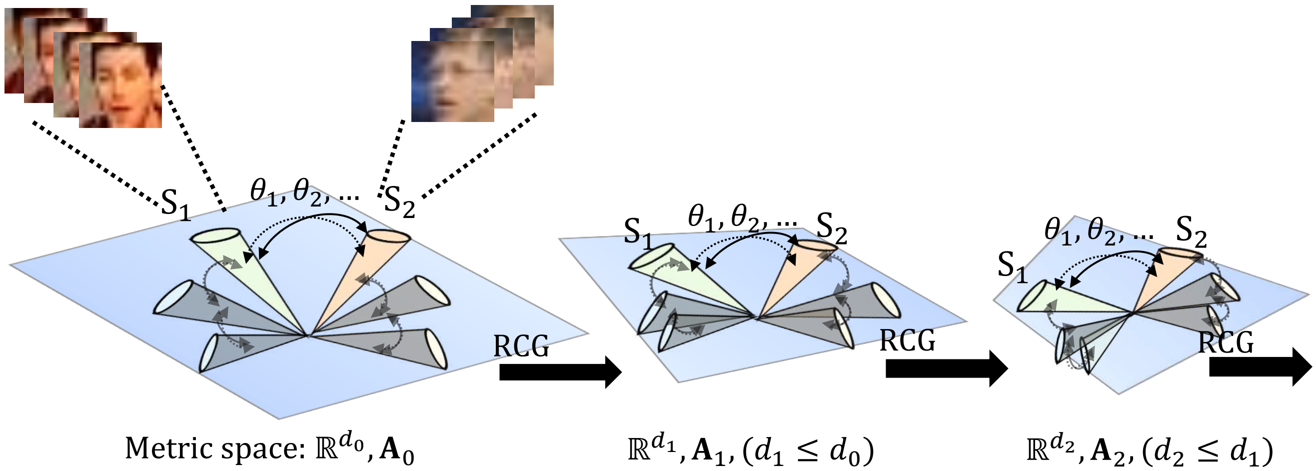


Figure 1: Conceptual diagram of the proposed metric learning method. The similarity between two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$ is calculated by the canonical angles $\{\theta_i\}$ in $\mathbf{A}_t$-based scalar product space (metric space). The metric space is updated step by step through the optimization of the metric $\mathbf{A}_t$, using Riemann conjugate gradient (RCG) method such that the different category subspaces are more separated while the same subspaces are more close, with reference to the local relationship between subspaces. The rank of $\mathbf{A}_t$ corresponds to a dimension of t-th metric space. In the optimization, the low-rank constraint on $\mathbf{A}_t$ has the effect of sequentially reducing the dimension of the metric space.

method [30], where the class subspaces are moved in reverse, while fixing the metric vector space.

Our idea of deforming a metric vector space is simple, yet effective. However, the optimization of the metric matrix $\mathbf{A}$ is not trivial, since $\mathbf{A}$ is required to be a symmetric positive definite matrix. This means that we need to solve an optimization problem on a Riemann manifold consisting of symmetric positive definite matrices. We solve our optimization problem by the Riemann conjugate gradient method (RCG) [1, 7]. Since RCG optimizes $\mathbf{A}$ by using the Euclidean gradient of the objective function, we first compute the gradient of the objective function, including the computation of canonical angles with respect to $\mathbf{A}$. Then, with this gradient, a metric space is sequentially updated by the RCG to provide efficient canonical angles for classification.

With the above technique, we construct a metric learning method, called $\mathbf{A}$-based Metric Learning for Subspace representation (AMLS), by formulating the optimization problem of finding $\mathbf{A}$ such that it minimizes the cost function designed based on local Fisher discriminant analysis (LFDA) [37, 38]. LFDA can efficiently incorporate local information of data and outperforms traditional feature extraction methods in classification. We embed local discriminative information into $\mathbf{A}$ under the framework of LFDA.

A few studies have proposed reliable metric learning methods using subspace representation [19, 49]. The proposed methods is essentially different from these methods in what is the object to be learned. The proposed method focuses on the metric space, in which class subspaces exist. We try to adjust the space itself as mentioned previously. In contrast, the conventional methods focus on each class subspace and rotate them by some orthogonal transformation.

Furthermore, our idea on learning a metric space leads us to the following fact: the rank of $\mathbf{A}$ corresponds to the dimension of a metric space. This characteristic enables us to reduce the dimension d of the metric space sequentially by decreasing the rank of $\mathbf{A}_0$ in optimization steps. We impose a low-rank constraint on $\mathbf{A}$ by adding a term of trace norm regularization to the cost function. This constraint induces sparseness on the singular values of $d \times d$ matrix $\mathbf{A}_0$. As a result, several singular values are set to zero so that only d' singular values are non-zero. This means that the original d-dimensional metric space has shrunk to a d'-dimensional metric space, according to the rank of $\mathbf{A}_0$, although the dimension of the metric space still appears to remain d. To extract the actual d' metric space, we perform the dimension reduction from d to d' by applying PCA to a set of learning data in the $\mathbf{A}_0$-based metric space. After that, we project the learning data onto the new d' metric space. A new metric $d' \times d'$ matrix $\mathbf{A}_1$ is optimized for the projected learning data in the same way mentioned above. The same set of steps can be sequentially repeated in the optimization process. Besides, our dimension reduction based on sparseness provides a more discriminative space, as demonstrated in the experiment section. Although the process in this idea may seem complicated, it is simple and scalable compared to previous dimension reduction methods.

The main contributions of this paper are as follow:

1. Introducing the generalized concept of canonical angles based on a metric space for subspace-based image set classification.
2. Deriving the gradient of the similarity computation using the generalized canonical angles, and then proposing a learning method, called **A**-based metric learning method for subspace representation (AMLS), for generating a discriminant metric space.
3. Incorporating the low-rank constraint on **A** into AMLS for sequential dimension reduction of metric space.
4. Demonstrating the fundamental performance of AMLS in tasks of image set based recognition using public databases, ETH-80, YTC, and UCF.

The paper is organized as follows. In Section 2, we describe the concept of the canonical angles in an **A**-based scalar product. In Section 3, we describe the proposed metric learning method, and discuss with the relationship between the proposed methods and previous methods. In Section 4, we present the details of the automatic dimension reduction method. In Section 5, we demonstrate the effectiveness of the proposed methods through face recognition and action recognition experiments using videos and an object recognition experiment using multi-view images. Section 6 concludes the paper.

2. Preliminaries

In this section, we provide the concepts of subspace representation and canonical angles and its natural generalization, canonical angles in an **A**-based inner product, which are the necessary fundamental techniques for constructing our method.

2.1. Subspace representation and canonical angles

Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ be an image set, where the image set has n images, each image is expressed as a d dimensional vector. The image set can be compactly and accurately represented by a low dimensional subspace. The orthonormal basis vectors $\mathbf{S} \in \mathbb{R}^{d \times m}$ of the m dimensional subspace can be obtained as eigenvectors corresponding to m largest eigenvalues of the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$.

One of the benefits of adopting the subspace representation is its matching flexibility, which is facilitated by the use of the canonical angles. Besides, a massive amount of data can be compared with low computational cost.

Given two m dimensional subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$ in d-dimensional vector space, the canonical angles $\{0 \leq \theta_1, \cdots, \theta_m \leq \frac{\pi}{2}\}$ between $\mathbf{S}_1$ and $\mathbf{S}_2$ are recursively defined as follows [2, 17]:

$$\cos\theta_i = \max_{\mathbf{u} \in \mathbf{S}_1} \max_{\mathbf{v} \in \mathbf{S}_2} \mathbf{u}_i^{\mathrm{T}} \mathbf{v}_i, \quad (1)$$
$$s.t.\ \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1,\ \mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = 0,\ i \neq j,$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the canonical vectors producing the i-th smallest canonical angle θ_i between $\mathbf{S}_1$ and $\mathbf{S}_2$. The j-th canonical angle θ_j is the smallest angle in the orthogonal direction to the canonical angles $\{\theta_k\}_{k=1}^{j-1}$. A conventional similarity applying the canonical angles is defined as follows [13].

$$f_p(\mathbf{S}_1, \mathbf{S}_2) = \sum_{i=1}^{m} \cos^2\theta_i. \quad (2)$$

This similarity can be easily obtained as $\|\mathbf{S}_1^{\mathrm{T}}\mathbf{S}_2\|_F^2$ without calculating individual angles [13].

2.2. Canonical angles with an A-based scalar product

The canonical angles described in the previous section are measured in a space equipped with the standard Euclidean scalar product. These angles can be generalized by introducing an **A**-based scalar product, i.e., scalar product is calculated on the metric space defined by the symmetric positive-definite matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ [24].

Let $(\mathbf{x}, \mathbf{y})_{\mathbf{A}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$ be an **A**-based scalar product, and $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{(\mathbf{x}, \mathbf{x})_{\mathbf{A}}}$ be the corresponding vector norm. Given two subspaces $\mathbf{S}_1$, $\mathbf{S}_2$, the canonical angles in the **A**-based scalar product between them are defined as follows:

$$\cos\theta_i = \max_{\mathbf{u} \in \mathbf{S}_1} \max_{\mathbf{v} \in \mathbf{S}_2} (\mathbf{u}_i, \mathbf{v}_i)_{\mathbf{A}}, \quad (3)$$
$$s.t.\ \|\mathbf{u}_i\|_{\mathbf{A}} = \|\mathbf{v}_i\|_{\mathbf{A}} = 1,\ (\mathbf{u}_i, \mathbf{u}_j)_{\mathbf{A}} = (\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{A}} = 0,\ i \neq j.$$

The corresponding similarity $f_{\mathbf{A}}$ and dissimilarity $d_{\mathbf{A}}$ are defined as follows:

$$f_{\mathbf{A}}(\mathbf{S}_1, \mathbf{S}_2) = \sum_{i=1}^{m} \cos^2\theta_i = \|\hat{\mathbf{S}}_1^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_2\|_F^2, \quad (4)$$

$$d_{\mathbf{A}}(\mathbf{S}_1, \mathbf{S}_2) = \sum_{i=1}^{m} \sin^2\theta_i = \sum_{i=1}^{m}(1 - \cos^2\theta_i),$$
$$= m - \|\hat{\mathbf{S}}_1^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_2\|_F^2 = m - f_{\mathbf{A}}(\mathbf{S}_1, \mathbf{S}_2), \quad (5)$$

where the orthonormal basis $\mathbf{S}_1$, $\mathbf{S}_2$ of each subspace are converted to **A**-orthonormal basis $\hat{\mathbf{S}}_1$, $\hat{\mathbf{S}}_2$, i.e., $\hat{\mathbf{S}}_1^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_1 = \hat{\mathbf{S}}_2^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_2 = \mathbf{I}$ to preserve the orthonormality of a subspace basis under the **A**-based metric by the following equation.

$$\hat{\mathbf{S}}_i = \mathbf{S}_i\mathbf{U}_i\Sigma_i^{-1/2}, \quad (6)$$

where the columns of $\mathbf{U}_i$ are the eigenvectors, and the diagonal elements of Σ_i are the eigenvalues of the matrix $\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i$. If $\mathbf{A}$ is the identity matrix, the above definition corresponds to the conventional canonical angles in the Euclidean scalar product.

Inspired by the above formulation, we propose a metric learning method based on this new scalar product space (also called metric space). To demonstrate its capabilities, we build a high-performance classification method for image sets in the next section.

3. A-based metric learning for subspace representation

In this section, we describe the details of the proposed **A**-based metric learning for subspace representation (AMLS). First, we formulate the optimization problem of $\mathbf{A}$ in order to learn a suitable metric space for calculating the canonical angles. Then, we derive the optimization method based on Riemann conjugate method with the gradient of the similarity, provided in the Eq. 4.

3.1. Problem formulation

Given N image sets $\{\mathbf{X}_i \in \mathbb{R}^{d\times n_i}\}_{i=1}^N$, where each image set $\mathbf{X}_i = [\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i]$ has n_i images, and each image is represented by a $d-$dimensional feature vector $\mathbf{x}_j^i$, our objective is to classify $\mathbf{X}_i$. We first generate subspaces $\{\mathbf{S}_i\}_{i=1}^N$ corresponding to each image set. Then, we learn the metric $\mathbf{A}$ by minimizing a discriminative cost function.

To design the cost function, we utilize the idea of local Fisher discriminant analysis (LFDA) [37, 38]. LFDA can work even if each class has a multimodal distribution by incorporating local relationships between data. First, we define two terms: the sum of the local similarity J_b between subspaces of different categories, and the sum of the local dissimilarity J_w between subspaces of the same category as follows:

$$J_b(\mathbf{A}) = \frac{1}{kN}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_k^i} f_{\mathbf{A}}(\mathbf{S}_i, \mathbf{S}_j), \tag{7}$$

$$\begin{aligned} J_w(\mathbf{A}) &= \frac{1}{kN}\sum_{i=1}^{N}\sum_{j\in\mathcal{P}_k^i} d_{\mathbf{A}}^2(\mathbf{S}_i, \mathbf{S_i}) \\ &= \frac{1}{kN}\sum_{i=1}^{N}\sum_{j\in\mathcal{P}_k^i} m - f_{\mathbf{A}}(\mathbf{S}_i, \mathbf{S}_i), \end{aligned} \tag{8}$$

where, $\mathcal{N}_k^i$ is an index set of k-neighbor subspaces with the i-th subspace in different categories from the i-th subspace, and $\mathcal{P}_k^i$ is an index set of k-neighbor subspaces with the i-th subspace in the same category from the i-th subspace. Finally, we define the discriminative cost function as follows:

$$J(\mathbf{A}) = J_b(\mathbf{A}) + J_w(\mathbf{A}) + \lambda(1 - \|\mathbf{A}\|_F^2)^2, \tag{9}$$

where, the last term is the regularization to prevent the norm of $\mathbf{A}$ from becoming too large, and $\lambda(\geq 0)$ is the weight parameter of the regularization. As the cost function decreases, we can obtain a more suitable metric space for classification.

In the next subsection, we describe the optimization method for the cost function with respect to $\mathbf{A}$.

3.2. Optimization

To minimize the cost function, we utilize the Riemann conjugate gradient method (RCG), since the metric $\mathbf{A}$ is on the manifold of symmetric positive definite matrices. The RCG updates $\mathbf{A}$ by searching a suitable matrix along the geodesic of the direction to the Riemann gradient of the cost function. The Riemann gradient is, intuitively, the closest vector to the Euclidean gradient that is also tangent to the manifold, in this case of symmetric positive definite matrices. Therefore, we need to show the Euclidean gradient of the cost function.

To this end, since the function has the terms of the similarity $f_{\mathbf{A}}$, in order to easily obtain the gradient, we first reformulate the similarity (Eq. 4) as follows:

$$\begin{aligned} f_{\mathbf{A}}(\mathbf{S}_i, \mathbf{S}_j) &= \|\hat{\mathbf{S}}_i^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_j\|_F^2 = tr(\hat{\mathbf{S}}_j^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_i\hat{\mathbf{S}}_i^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_j) \\ &= tr(\hat{\mathbf{S}}_j\hat{\mathbf{S}}_j^{\mathrm{T}}\mathbf{A}\hat{\mathbf{S}}_i\hat{\mathbf{S}}_i^{\mathrm{T}}\mathbf{A}) \\ &= tr(\mathbf{S}_j\mathbf{U}_j\Sigma_j^{-1/2}\Sigma_j^{-1/2}\mathbf{U}_j^{\mathrm{T}}\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_i\mathbf{U}_i\Sigma_i^{-1/2}\Sigma_i^{-1/2}\mathbf{U}_i^{\mathrm{T}}\mathbf{S}_i^{\mathrm{T}}\mathbf{A}) \\ &= tr(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}). \end{aligned} \tag{10}$$

In the above equation, we can use the cyclic property of the matrix trace, Eq. 6 and $\mathbf{U}_i\Sigma_i^{-1}\mathbf{U}_i^{\mathrm{T}} = (\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}$. This allows us to obtain the gradient $\nabla_{\mathbf{A}} f_{\mathbf{A}}(\mathbf{S}_i, \mathbf{S}_i)$ of the similarity with respect to $\mathbf{A}$ as follows:

$$\begin{aligned} \nabla_{\mathbf{A}} f_{\mathbf{A}}(\mathbf{S}_i, \mathbf{S}_i) &= \nabla_{\mathbf{A}} tr(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}) \\ &= \mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}} \\ &\quad + \mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}\mathbf{S}_i^{\mathrm{T}} \\ &\quad - \mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}} \\ &\quad - \mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_j(\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_j)^{-1}\mathbf{S}_j^{\mathrm{T}}\mathbf{A}\mathbf{S}_i(\mathbf{S}_i^{\mathrm{T}}\mathbf{A}\mathbf{S}_i)^{-1}\mathbf{S}_i^{\mathrm{T}}. \end{aligned} \tag{11}$$

With this gradient $\nabla_{\mathbf{A}} f_{\mathbf{A}}$, we can obtain the Euclidean gradient of the cost function as follows:

$$\begin{aligned}\nabla_{\mathbf{A}} J(\mathbf{A}) =& \frac{1}{kN}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_k^i} \nabla_A f_{\mathbf{A}}(\mathbf{S}_i,\mathbf{S}_j) \\ &- \frac{1}{kN}\sum_{i=1}^{N}\sum_{j\in\mathcal{P}_k^i} \nabla_A f_{\mathbf{A}}(\mathbf{S}_i,\mathbf{S}_j) \\ &- \lambda(1-\|\mathbf{A}\|_F^2)\mathbf{A}. \end{aligned} \tag{12}$$

$\mathbf{A}$ can be updated by the RCG with this Euclidean gradient, until the maximum number of iterations is reached.

The image-set classification can be executed by the nearest neighbor strategy by using the similarity (Eq. 4) by the optimized $\mathbf{A}$.

4. Dimension reduction of a metric space

In this section, we describe the details of the proposed method of sequentially reducing the dimension of the original metric space. The method consists of two steps, as mentioned in Sec. 1. First, we outline the algorithm of the optimization with a low-rank constraint. Next, we show how to apply principal component analysis (PCA) to our $\mathbf{A}$-based metric space to reduce its dimensionality.

4.1. Optimization with Low rank constraint

Our idea is based on the fact that the rank of metric $\mathbf{A}$ indicates the dimension of a metric space, as mentioned previously. This suggests that we can realize the dimension reduction effectively by decreasing the rank of a metric matrix $\mathbf{A}$. We impose a low-rank constraint on $\mathbf{A}$ by adding a term of trace norm regularization to our cost function $J(\mathbf{A})$. The cost function is then modified as follows:

$$J_{lr}(\mathbf{A}) = J(\mathbf{A}) + \eta\|\mathbf{A}\|_*, \tag{13}$$

where $\|\mathbf{A}\|_*$ indicates the trace norm of $\mathbf{A}$, and $\eta(>0)$ is the weight parameter of the regularization.

To minimize the above mentioned cost function, $J_{lr}(\mathbf{A})$, we utilize the proximal gradient method (PGM) [5, 6, 26]. This method is built as a combination of two processes: a regularization operation and the traditional gradient-based method without the trace norm regularization. In this experiment, we consider a combination of an operation for the trace norm regularization and the Riemann conjugate gradient (RCG) method for optimizing $J(\mathbf{A})$ described in the previous section. The first operation prox_η^{tr} is defined as follows [8]:

$$\mathrm{prox}_\eta^{tr}(\mathbf{A}) = \mathbf{U}\max(\Sigma - \eta\mathbf{I},\ 0)\mathbf{U}^{\mathrm{T}}, \tag{14}$$

where the columns of $\mathbf{U}$ are the eigenvectors $\mathbf{A}$, the Σ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$, $\mathbf{I}$ is the identity matrix, and max is the element-wise max operation. In PGM, the operation prox_η^{tr} is applied to $\mathbf{A}$ after each step of the RCG to minimize $J_{lr}(\mathbf{A})$.

After the rank has decreased to d' by this optimization, we move to the next process of PCA-based dimension reduction.

Algorithm 1: Algorithm of the proposed metric learning method.

Input: N image sets $\{\mathbf{X}_i \in \mathbb{R}^{d\times n_i}\}_{i=1}^N$, class labels $\{l_i\}_{i=1}^N$ and weight parameters λ, η, subspace dimension m, the number of neighborhood subspaces k, and the number of maximum iterations T

```
Randomly initialize A_0 ∈ R^{d×d}.
V_0 = I                 // Matrix for dimension reduction
Generate m-dimensional subspace S_i for each image set X_i.
for t = 1, ..., T do
    Compute the gradient ∇_{A_{t-1}} J(A_{t-1}) by Eq. 12.
    A_t = RCG(A_{t-1}, ∇_{A_{t-1}} J(A_{t-1}))   // Update A by RCG step
    A_t = prox_η^tr(A_t), (Eq. 14)       // Low-rank constraint
    if Rank of A_t is reduced then
        Calculate V by Eq. 18
        V_t = V^T V_{t-1}
        A_t = V^T A_t V                  // Dimension reduction
        for i = 1 ... N do
            X̂_i = V_t X_i
            Generate m-dimensional subspace S_i from X̂_i
        end
    else
        V_t = V_{t-1}
    end
end
return A_T and V_T
```

4.2. Dimension reduction based on A-based PCA

The dimension of the metric space still appears to remain d, after the rank of $\mathbf{A}$ is reduced to d' in the previous subsection. This means that we need to extract the actual d'-dimensional metric space. To this end, we apply PCA to a set of learning data, where PCA is required to be performed in the updated metric space based on the optimized metric $\mathbf{A}$ in the previous section.

In this subsection, we explain the technical details of how to obtain a transformation matrix $\mathbf{V} \in \mathbb{R}^{d\times d'}$ that maps data from the original d-dimensional metric space to the d'(= the reduced rank of $\mathbf{A}$)-dimensional metric space.

First of all, we describe how to perform PCA in the $\mathbf{A}$-based metric space. Given N image sets $\{\mathbf{X}_i \in \mathbb{R}^{d \times n_i}\}_{i=1}^N$, where each image set $\mathbf{X}_i = [\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i]$ has n_i images, and each image is represented by a d-dimensional vector $\mathbf{x}_j^i$. As well known, PCA is a method for maximizing the scalar product between image vectors $\{\mathbf{x}_j^i\}$ and a principal component vector $\mathbf{w}$. We define the objective function of PCA in the $\mathbf{A}$-based metric space as follows:

$$\arg\max_{\mathbf{w}} \sum_{i,j} (\mathbf{w}, \mathbf{x}_j^i)_{\mathbf{A}}^2 = \sum_{i,j} (\mathbf{w}^{\mathrm{T}} \mathbf{A} \mathbf{x}_j^i)^2, \quad s.t.\ \|\mathbf{w}\|_{\mathbf{A}}^2 = 1. \tag{15}$$

The above optimization problem can be rewritten by the Lagrange multipliers method as follows:

$$\arg\max_{\mathbf{w}} \sum_{i,j} (\mathbf{w}^{\mathrm{T}} \mathbf{A} \mathbf{x}_i^j)^2 + \alpha(\mathbf{w}^{\mathrm{T}} \mathbf{A} \mathbf{w} - 1)^2. \tag{16}$$

The differential of the Eq. 16 yields the following equation.

$$\begin{aligned} &\mathbf{ARAw} - \beta \mathbf{Aw} \\ &= \mathbf{ARv} - \beta \mathbf{v}, \end{aligned} \tag{17}$$

where $\mathbf{v} = \mathbf{Aw}$ and $\mathbf{R} = \sum_{i,j} (\mathbf{x}_j^i)(\mathbf{x}_j^{i\mathrm{T}})$. By assuming that Eq. 17 equals 0, the transformation matrix $\mathbf{V} \in \mathbb{R}^{d \times d'}$ can be obtained as a matrix whose columns are the $d'(= rank(\mathbf{A}))$ eigenvectors corresponding to the d' largest eigenvalues of the following equation:

$$\mathbf{ARV} = \beta \mathbf{V}. \tag{18}$$

With this matrix $\mathbf{V}$, we extract a new metric $\mathbf{A}_1$ of the actual d'-dimensional metric space as $\mathbf{V}^{\mathrm{T}}\mathbf{AV}$. Each image vector is projected onto this new metric space as $\mathbf{V}^{\mathrm{T}}\mathbf{x}_j^i$. After that, new subspaces $\{\mathbf{S}_i\}$ are generated from the projected vectors $\{\mathbf{V}^{\mathrm{T}}\mathbf{x}_j^i\}_j$.

AMLS with the Low-rank constraint is called AMLSL. This algorithm is summarized in Algorithm 1. If $\eta = 0$, the algorithm corresponds to AMLS explained in the Sec. 3.

5. Evaluation experiments

In this section, we demonstrate the effectiveness of the proposed methods through extensive experiments on three tasks: video-based face recognition, multi-view image-based object recognition, and action recognition. For each task, we used YouTube Celebrity dataset [22], ETH-80 dataset [25], and UCF sports dataset [31, 35], respectively.

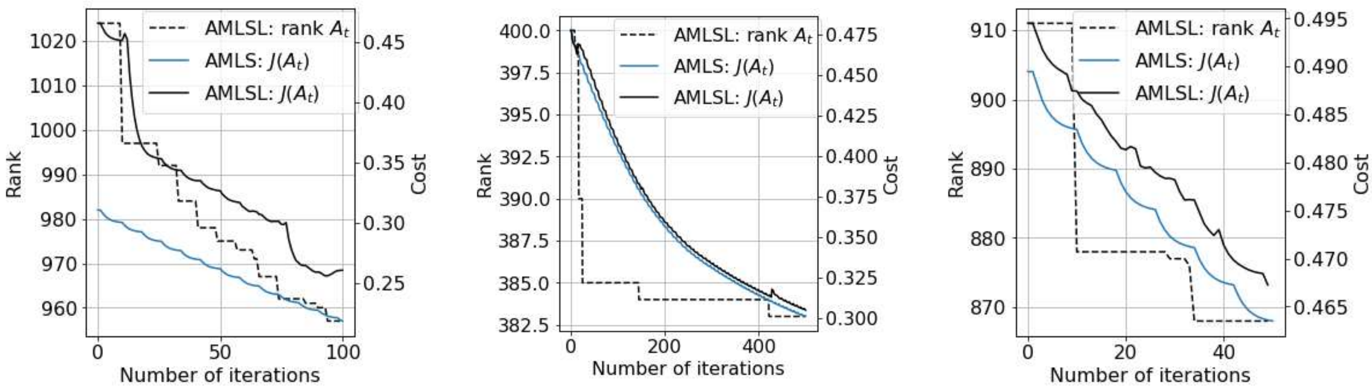
5.1. Experimental settings

The ETH-80 dataset consists of eight different categories, captured from 41 viewpoints, and there are 10 objects for each category. Five objects were randomly sampled from each category and used as a training data, and the remaining five objects were used as a testing data. As an input image set, we used 41 multi-view images for each object. We resized each image to 32×32 pixels and used a 1024-dimensional feature vector whose element is a pixel intensity of the corresponding image. We evaluated the classification performance of each method in terms of the average accuracy of ten trials using randomly divided datasets.

The YTC dataset [22] contains 1910 videos of 47 people. Face recognition using this dataset is still challenging since all of the face images are low-resolution and face images of the same person have extreme variations, such as change of face direction and emotion. We used a set of face images extracted from a video by the Viola and Jones detection algorithm [41], as an image set. All the extracted face images were resized to 20×20 pixels. After converting each image to grayscale, we applied a histogram equalization contrast adjustment method as a preprocessing. We used a 400-dimensional feature vector whose element is a pixel intensity of the corresponding image. We used three videos per each person randomly selected as training data, and six videos per each person randomly selected as test data. We evaluated the classification performance of each method in terms of the average accuracy of five trials using randomly divided datasets.

The UCF sports dataset contains 150 sequences of subjects practicing sports, with ten classes, namely: diving, golf swing, kicking, lifting, horse riding, running, skateboarding, swing-bench, swing-side, and walking. The action bounding box has been extracted, using annotations provided. Then, we resized the extracted images to 38×24 pixels and converted them to grayscale images. We used a 912-dimensional feature vector whose element is a pixel intensity of each image. We evaluated the classification performance of each method by a leave-one-out cross-validation scheme (LOOCV), a standard experiment setting for this data, i.e., 150 repetitions of training, with one video as testing and the remaining 149 videos as training data.

The proposed methods have five parameters: the maximum number of optimization steps, weight parameters λ and η for corresponding regularization, subspace dimension, and the number of neighbor subspaces k. The maximum numbers of optimization steps were fixed to 100, 500, and 50 for ETH-80, YTC and UCF datasets, respectively. For all experiments, λ was fixed to 1e-3. For the proposed method with the low-rank constraint, AMLSL, η was fixed at 1e-4. The subspace dimension was tuned at the range from 10 to 30 with the increments of 10 by the grid search



(a) ETH-80 (b) YTC (c) UCF

Figure 2: The curves of the cost function $J(\mathbf{A})$ and rank of $\mathbf{A}$ (the dimension of a metric space) on three datasets. The horizontal axis means the number of iterations of the optimization method. The blue and black solid lines are values of the cost function on AMLS and AMLSL respectively, and the black break lines indicate the rank of $\mathbf{A}$.

algorithm on the training data. The number of neighbors k was tuned at the range from 2 to 10 with the increments of 2 by the same strategy.

5.2. Comparison methods

To examine the effectiveness of the proposed methods, we compared them with various methods: classification methods based on models other than subspace: affine hull based method (AHISD) [4] and covariance based metric learning methods, Log-Euclidean Metric Learning (LEML) [20] and Riemann Manifold Metric Learning on Symmetric Positive Definite manifold (RMML-SPD) [49].

For comparison with subspace based classification method, we compared with traditional methods: Discriminative Canonical Correlations (DCC) [23], Grassmann Discriminant Analysis (GDA) [13], Graph-Embedding GDA (GGDA)[16], and and state-of-the-art metric learning method: Projection Metric Learning (PML) [19], and RMML on Grassmann manifold (RMML-GM) [49].

The parameters, other than the subspace dimension, of the above methods were tuned at the suggested range by the original papers under our experimental settings using the grid-search algorithm on the training data. The subspace dimension was tuned at the range from 10 to 30 with the increments 10 by the same strategy.

5.3. Results and discussion

We first discuss the characteristic of the proposed methods by using the results shown in Fig. 2. Figure 2 shows transitions of the cost function $J(\mathbf{A})$ (Eq. 9) and the rank of $\mathbf{A}$ (the dimension of a metric space) on the three datasets. Since the cost values decrease as the number of iterations of the optimization method increases, the basic ability of the proposed optimization method using the differential (Eq. 12) was validated. The low-rank constraint did not contribute to the improvement of the cost value unfortunately, as the cost values of AMLSL are larger than AMLS. However, it was confirmed that the optimization step automatically learns the dimension of the metric space. This may suppress the effect of overfitting in the testing phase. The more detailed discussion is in the later with the results of classification.

Table 1 shows the experimental results including those from various conventional methods. Overall, subspace-based methods show superior or the same results compared with the other models under our settings. The effectiveness of the proposed methods can be seen as it achieved better results compared with state of the art subspace-based metric learning methods, such as PML and RMML-GM, which learn classification methods based on canonical angles in the standard metric space. This supports the effectiveness of our key idea; to learn efficient metric space for calculating canonical angles.

AMLSL showed the superior or same results to AMLS, although its cost values are larger, as shown in Fig. 2. This suggests that by applying dimension reduction automatically, AMLSL can avoid overfitting, and efficiently extract essential information for the classification.

6. Conclusion

In this paper, we proposed $\mathbf{A}$-based Metric Learning for Subspace representation (AMLS). We further extended the idea of AMLS by adding the Low-rank constraint. This enhanced AMLSL is a powerful method with a function of dimension reduction and high discriminative ability. The core ideas behind the proposed methods are 1) measuring

Table 1: Experimental results (recognition rate (%), standard deviation) of the three tasks using the public datasets.

	ETH-80	YTC	UCF
AHISD [4]	70.25±4.78	56.31±4.61	53.33
LEML [20]	89.75±2.91	40.92±1.84	64.76
RMML-SPD [49]	83.75±4.90	39.57±4.45	53.33
DCC [23]	93.00±3.07	51.70±3.88	59.33
GDA [13]	92.50±4.25	45.74±7.27	70.00
GGDA [16]	93.75±4.75	47.45±8.06	47.33
PML [19]	93.75±3.17	54.82±4.63	72.67
RMML-GM [49]	90.25±4.16	54.96±4.96	59.33
AMLS	**95.28±3.21**	59.29±2.72	71.33
AMLSL	95.25±2.82	**59.85±3.53**	**74.00**

the canonical angles in a metric space equipped with metric $\mathbf{A}$, a symmetric positive definite matrix, and 2) designing suitable metric space for calculating canonical angles by optimizing metric matrix $\mathbf{A}$ in terms of the local relationship among subspaces. We formulated the optimization method of $\mathbf{A}$ based on the Riemann conjugate gradient method. To this end, we have rewritten the subspace similarity on a metric space and then derived the gradient of it. We verified the effectiveness of the proposed methods through the extensive classification experiments using the public datasets, ETH-80, TYC, and UCF.

To the best of our knowledge, this work is the first pursuit of designing a metric space for measuring canonical angles based on $\mathbf{A}$-based scalar product. As this approach is fundamentally different from conventional methods, which are designed to utilize standard canonical angles efficiently, we think that our methods can offer various possible future directions, e.g., the combination of the proposed method with the kernel-based subspace learning method [13, 16], or to incorporate the success of the metric learning for vector data [9, 29, 34, 46, 45].

Acknowledgments

This work was partly supported by JSPS KAKENHI grant number 19H04129, and a Grant-in-Aid for JSPS Fellows.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009. 2

[2] Sydney N Afriat. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 53, pages 800–816, 1957. 1, 3

[3] Peter N. Belhumeur and David J. Kriegman. What Is the Set of Images of an Object Under All Possible Illumination Conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998. 1

[4] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *Proceedings of Computer Vision and Pattern Recognition*, pages 2567–2573. IEEE, 2010. 1, 7, 8

[5] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011. 5

[6] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005. 5

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

[8] Maryam Fazel, Haitham Hindi, Stephen P Boyd, et al. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American control conference*, volume 6, pages 4734–4739. Citeseer, 2001. 5

[9] Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 8

[10] Kazuhiro Fukui and Atsuto Maki. Difference Subspace and Its Generalization for Subspace-Based Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2164–2177, 2015. 1

[11] Kazuhiro Fukui and Osamu Yamaguchi. Face Recognition Using Multi-viewpoint Patterns for Robot Vision. In *The Eleventh International Symposium of Robotics Research*, pages 192–201, 2005. 1

[12] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001. 1

[13] Jihun Hamm and Daniel D Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, pages 376–383. ACM, 2008. 1, 3, 7, 8

[14] Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C. Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 1

[15] Mehrtash Tafazzoli Harandi, Mathieu Salzmann, Sadeep Jayasumana, Richard I. Hartley, and Hongdong Li. Expanding the family of grassmannian kernels: An embedding perspective. In *ECCV*, 2014. 1

[16] Mehrtash T Harandi, Conrad Sanderson, Sareh Shirazi, and Brian C Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching.

In *Proceedings of Computer Vision and Pattern Recognition*, pages 2705–2712. IEEE, 2011. 1, 7, 8

[17] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 1, 3

[18] Yiqun Hu, Ajmal S Mian, and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1992–2004, 2012. 1

[19] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on Grassmann manifold with application to video based face recognition. In *Proceedings of Computer Vision and Pattern Recognition*, pages 140–149, 2015. 1, 2, 7, 8

[20] Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, and Xilin Chen. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *International conference on machine learning*, pages 720–729, 2015. 1, 7, 8

[21] Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on grassmann manifolds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[22] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 6

[23] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007. 1, 7, 8

[24] Andrew V Knyazev and Merico E Argentati. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002. 1, 3

[25] Bastian Leibe and Bernt Schiele. Analyzing appearance and contour based methods for object categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 409–415, 2003. 6

[26] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. 5

[27] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *Proceedings of Computer Vision and Pattern Recognition*, pages 1137–1145, 2015. 1

[28] Jiwen Lu, Gang Wang, and Jie Zhou. Simultaneous feature and dictionary learning for image set based face recognition. *IEEE Transactions on Image Processing*, 26(8):4042–4054, 2017. 1

[29] Yung-Kyun Noh, Byoung-Tak Zhang, and Daniel D Lee. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1822–1830, 2010. 8

[30] Erkki Oja and Maija Kuusela. The ALSM algorithm—an improved subspace method of classification. *pattern Recognition*, 16(4):421–427, 1983. 2

[31] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *Proceedings of Computer Vision and Pattern Recognition*, volume 1, page 6, 2008. 6

[32] Hitoshi Sakano and Naoki Mukawa. Kernel mutual subspace method for robust facial image recognition. In *International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies* , volume 1, pages 245–248, 2000. 1

[33] Syed Afaq Ali Shah, Uzair Nadeem, Mohammed Bennamoun, Ferdous Ahmed Sohel, and Roberto Togneri. Efficient image set classification using linear regression based image reconstruction. In *Computer Vision and Pattern Recognition Workshops*, pages 601–610, 2017. 1

[34] Yuan Shi, Aurélien Bellet, and Fei Sha. Sparse compositional metric learning. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014. 8

[35] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *Computer vision in sports*, pages 181–208. Springer, 2014. 6

[36] Lincon Souza, Hideitsu Hino, and Kazuhiro Fukui. 3d object recognition with enhanced grassmann discriminant analysis. In *ACCV 2016 Workshop (HIS 2016)*, 2016. 1

[37] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 905–912. ACM, 2006. 2, 4

[38] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of machine learning research*, 8(May):1027–1061, 2007. 2, 4

[39] Chendra Hadi Suryanto, Jing-Hao Xue, and Kazuhiro Fukui. Randomized time warping for motion recognition. *Image and Vision Computing*, 54:1–11, 2016. 1

[40] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011. 1

[41] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 6

[42] Ruiping Wang and Xilin Chen. Manifold discriminant analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 429–436. IEEE, 2009. 1

[43] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2496–2503. IEEE, 2012. 1

[44] Ruiping Wang, Shiguang Shan, Xilin Chen, and Wen Gao. Manifold-manifold distance with application to face recognition based on image set. In *Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[45] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009. 8

[46] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003. 8

[47] Osamu Yamaguchi, Kazuhiro Fukui, and Kenichi Maeda. Face recognition using temporal image sequence. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323, 1998. 1

[48] Zhong-Qiu Zhao, Shou-Tao Xu, Dian Liu, Wei-Dong Tian, and Zhi-Da Jiang. A review of image set classification. *Neurocomputing*, 335:251–260, 2019. 1

[49] Pengfei Zhu, Hao Cheng, Qinghua Hu, Qilong Wang, and Changqing Zhang. Towards Generalized and Efficient Metric Learning on Riemannian Manifold. In *IJCAI*, pages 3235–3241, 2018. 1, 2, 7, 8