Hybrid Euclidean-and-Riemannian Metric Learning for Image Set Classification

Zhiwu Huang[1,2], Ruiping Wang[1](✉), Shiguang Shan[1], and Xilin Chen[1]

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

Abstract. We propose a novel hybrid metric learning approach to combine multiple heterogenous statistics for robust image set classification. Specifically, we represent each set with multiple statistics – mean, covariance matrix and Gaussian distribution, which generally complement each other for set modeling. However, it is not trivial to fuse them since the mean vector with d-dimension often lies in Euclidean space $\mathbb{R}^d$, whereas the covariance matrix typically resides on Riemannian manifold Sym_d^+. Besides, according to information geometry, the space of Gaussian distribution can be embedded into another Riemannian manifold Sym_{d+1}^+. To fuse these statistics from heterogeneous spaces, we propose a Hybrid Euclidean-and-Riemannian Metric Learning (HERML) method to exploit both Euclidean and Riemannian metrics for embedding their original spaces into high dimensional Hilbert spaces and then jointly learn hybrid metrics with discriminant constraint. The proposed method is evaluated on two tasks: set-based object categorization and video-based face recognition. Extensive experimental results demonstrate that our method has a clear superiority over the state-of-the-art methods.

1 Introduction

Learning problems of classifying image sets is commonly encountered in many branches of computer vision community. In video-based face recognition, for example, each face video can be considered as an image set, which may cover large variations in a subject's appearance due to camera pose changes, nonrigid deformations, or different illumination conditions. The objective of image set classification task is to classify an unknown image set to one of the gallery image sets. Generally speaking, existing image set classification methods mainly focus on the key issues of how to quantify the degree of match between two sets and how to learn discriminant function from training image sets [1].

In the aspect of how to quantify the degree of match, image set classification methods can be broadly partitioned into sample-based methods [2–7], subspace-based methods [1,8–13] and distribution-based methods [14,15]. Sample-based methods compare sets based on matching their sample-based statistics (SAS)

© Springer International Publishing Switzerland 2015
D. Cremers et al. (Eds.): ACCV 2014, Part III, LNCS 9005, pp. 562–577, 2015.
DOI: 10.1007/978-3-319-16811-1_37

Table 1. Three major challenges for set modeling: arbitrary data distribution, large data variation and small set size. Here, the tick (/cross) indicates the corresponding set statistics, i.e., sample-based (SAS), subspace-based (SUS) or distribution-based (DIS) statistics, is (/not) qualified to handle the challenge in that column. The last row represents the combination of ALL above three statistics in our proposed method.

Statistics	Arbitrary distribution	Large variation	Small size
SAS	✓	×	✓
SUS	✓	×	✓
DIS	×	✓	✓
ALL	✓	✓	✓

such as sample mean and affine (convex) combination of samples. This kind of methods include Maximum Mean Discrepancy (MMD)[2], Affine (Convex) Hull based Image Set Distance (AHISD, CHISD)[3] and Sparse Approximated Nearest Point (SANP) [4] etc. Subspace-based methods typically apply subspace-based statistics (SUS) to model sets and classify them with given similarity function. For example, Mutual Subspace Method (MSM) [8] represent sets as linear subspaces and match them using canonical correlations [16]. The distribution-based methods, e.g., Single Gaussian Model (SGM) [14] and Gaussian Mixture Models (GMM) [15], model each set with distribution-based statistics (DIS) (i.e., Gaussian distribution), and then measure the similarity between two distributions in terms of the Kullback-Leibler Divergence (KLD) [17].

In the real-world scenario, image sets are often of arbitrary data distribution or large data variation or small set size. As shown in Table 1, however, SAS performs poorly when sets are of large variation while SUS is not good at dealing with the challenge of small set size, though both have no assumption of data distribution. Different from them, DIS requires the set data to follow Gaussian distribution. Fortunately, the three kinds of statistics are complementary for each other: when sets contain small variation, SAS is qualified to model sets with any size and in arbitrary distribution. As a complement, SUS is able to tackle the problem of large variation but requires the set size to be large enough. In addition to the above situations, the last challenge of large variation meanwhile small set size can be overcame by DIS to some extent. This is because DIS is usually obtained by jointly estimating the mean and the covariance, which are capable of adapting to the scenario of small set size and characterizing large data variation respectively.

The other important problem in set classification is how to learn discriminant function from training image sets, which generally are sets of single vectors. The first kind of methods [1,7,11,13] is to learn the discriminant function in Euclidean space. For instance, Discriminative Canonical Correlations (DCC) [1] seeks a discriminant projection of single vectors in Euclidean space to maximizes (minimizes) the canonical correlations of within-class (between-class) sets. Set-to-Set Distance Metric Learning (SSDML) [7] learns a proper metric between pairs of single vectors in Euclidean space to get more accurate set-to-set affine hull based distance for classification. Localized Multi-Kernel Metric Learning (LMKML) [13] treats three order statistics of each set as single vectors again in Euclidean spaces and attempts to learn one metric for them by embedding Euclidean spaces into Reproducing Kernel Hilbert Spaces (RKHS). However, the higher order statistics they used such as the tensors typically lie in non-Euclidean space, which does not adhere to Euclidean geometry. Therefore, in this method, applying the kernel function induced by Euclidean metric to the higher order statistics does not always preserve the original set data structure. In contrast, the second kind of learning methods [10,12,18] treat each subspace-based statistics as a point in a specific non-Euclidean space, and perform metric learning in the same space. For example, Grassmann Discriminant Analysis (GDA) [10] and Covariance Discriminative Learning (CDL) [12] represent each linear subspace or covariance matrix as a point on a Riemannian manifold and learn discriminant Riemannian metrics on that manifold.

In this paper, we propose a new approach to combine multiple statistics for more robust image set classification. From a view of probability statistics, we model each set as sample mean, covariance matrix and Gaussian distribution, which are the corresponding instances of SAS, SUS and DIS. As discussed above, the three kinds of statistics complement each other especially in the real-world settings. Therefore, we attempt to fuse them to simultaneously deal with the challenges of arbitrary distribution, large variation and small set size, which is shown in Table 1. However, combining these multiple statistics is not an easy job because they lie in multiple heterogeneous spaces: the mean is a d-dimension vector lying in Euclidean space $\mathbb{R}^d$. As studied in [19–21], the covariance matrix is regarded as a Symmetric Positive Definite (SPD) matrix residing on a Sym_d^+ manifold. In comparison, the space of Gaussian distribution can be embedded into another Riemannian manifold Sym_{d+1}^+ by employing information geometry [22]. To fuse these multiple statistics from heterogeneous spaces, inspired by our previous work [23], we propose a Hybrid Euclidean-and-Riemannian Metric Learning (HERML) method to exploit the Euclidean and Riemmannian metrics for embedding these spaces into high dimension Hilbert spaces, and jointly learn corresponding metrics of multiple statistics for discriminant objective.

2 Background

In this section, we first review the Riemannian metric of SPD matrices. This metric derives the Riemannian kernel function, which can be used to embed the Riemannian manifold into RKHS. Then, we introduce the Information-Theoretic Metric Learning method and its kernelized version.

2.1 Riemannian Metric of Symmetric Positive Definite Matrices

As mostly studied in [19–21], the space of SPD matrices is a specific Riemannian manifold Sym^+ when equipping Riemannian metric. The two most widely used Riemannain metric are the Affine-Invariant Distance (AID) [19] and the Log-Euclidean Distance (LED) [21]. In this work, we focus on the LED, which is a true geodesic distance on Sym^+ and yields a positive definite kernel as studied in [12, 24].

By exploiting the Lie group structure of Sym^+, the LED for Sym^+ manifold is derived under the operation $\boldsymbol{X}_i \odot \boldsymbol{X}_j :\, = exp(log(\boldsymbol{X}_i) + log(\boldsymbol{X}_j))$ for $\boldsymbol{X}_i, \boldsymbol{X}_j \in Sym^+$, where $exp(\cdot)$ and $log(\cdot)$ denote the common matrix exponential and logarithm operators. Under the log-Euclidean framework, a geodesic between $\boldsymbol{X}_i, \boldsymbol{X}_j \in Sym^+$ is defined as $\varsigma(t) = exp((1-t)log(\boldsymbol{X}_i) + tlog(\boldsymbol{X}_j))$. The geodesic distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is then expressed by classical Euclidean computations in the domain of matrix logarithms:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \|log(\boldsymbol{X}_i) - log(\boldsymbol{X}_j)\|_F. \tag{1}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius form. As studied in [12], a Riemannian kernel function on the Sym^+ manifold can be derived by computing the corresponding inner product in the space:

$$\kappa_x(\boldsymbol{X}_i, \boldsymbol{X}_j) = tr(log(\boldsymbol{X}_i) \cdot log(\boldsymbol{X}_j)) \tag{2}$$

2.2 Information-Theoretic Metric Learning

Information-Theoretic Metric Learning (ITML) [25] method formulates the problem of metric learning as a particular Bregman optimization, which aims to minimize the LogDet divergence subject to linear constraints:

$$\begin{aligned} \min_{\boldsymbol{A}\succeq 0, \boldsymbol{\xi}} \quad & D_{\ell d}(\boldsymbol{A}, \boldsymbol{A}_0) + \gamma D_{\ell d}(diag(\boldsymbol{\xi}), diag(\boldsymbol{\xi}_0)) \\ s.t. \quad & tr(\boldsymbol{A}(\boldsymbol{x}_i - x_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T) \leq \boldsymbol{\xi}_{ij}, \quad (i,j) \in S \\ & tr(\boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T) \geq \boldsymbol{\xi}_{ij}, \quad (i,j) \in D \end{aligned} \tag{3}$$

where $\boldsymbol{A}, \boldsymbol{A}_0 \in \mathbb{R}^{d \times d}$, $D_{\ell d}(\boldsymbol{A}, \boldsymbol{A}_0) = tr(\boldsymbol{A}\boldsymbol{A}_0^{-1}) - logdet(\boldsymbol{A}\boldsymbol{A}_0^{-1}) - d$, d is the dimensionality of the data. $(i, j) \in S(D)$ indicates the pair of samples $\boldsymbol{x}_i, \boldsymbol{x}_j$ is in similar (dissimilar) class. $\boldsymbol{\xi}$ is a vector of slack variables and is initialized to $\boldsymbol{\xi}_0$, whose components equal to a upper bound of distances for similarity constraints and a lower bound of distances for dissimilarity constraints.

Meanwhile, ITML method can be extended to a kernel learning one. Let $\boldsymbol{K}_0$ denote the initial kernel matrix, that is, $\boldsymbol{K}_0(i, j) = \phi(\boldsymbol{x}_i)^T \boldsymbol{A}_0 \phi(\boldsymbol{x}_j)$, where ϕ is an implicit mapping from original space to high dimensional kernel space. Note that the Euclidean distance in kernel space may be written as $\boldsymbol{K}(i, i) + \boldsymbol{K}(j, j) - 2\boldsymbol{K}(i, j) = tr(\boldsymbol{K}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^T)$, where $\boldsymbol{K}(i, j) = \phi(\boldsymbol{x}_i)^T \boldsymbol{A} \phi(\boldsymbol{x}_j)$ is the learned kernel matrix, $\boldsymbol{A}$ represents an operator in the RKHS, whose size can be potentially infinite, and $\boldsymbol{e}_i$ is the i-th canonical basis vector. Then the kernelized version of ITML can be formulated as:

$$\begin{aligned} \min_{\boldsymbol{K}\succeq 0, \boldsymbol{\xi}} \quad & D_{\ell d}(\boldsymbol{K}, \boldsymbol{K}_0) + \gamma D_{\ell d}(diag(\boldsymbol{\xi}), diag(\boldsymbol{\xi}_0)) \\ s.t. \quad & tr(\boldsymbol{K}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^T) \leq \boldsymbol{\xi}_{ij}, \quad (i,j) \in S \\ & tr(\boldsymbol{K}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^T) \geq \boldsymbol{\xi}_{ij}, \quad (i,j) \in D \end{aligned} \tag{4}$$

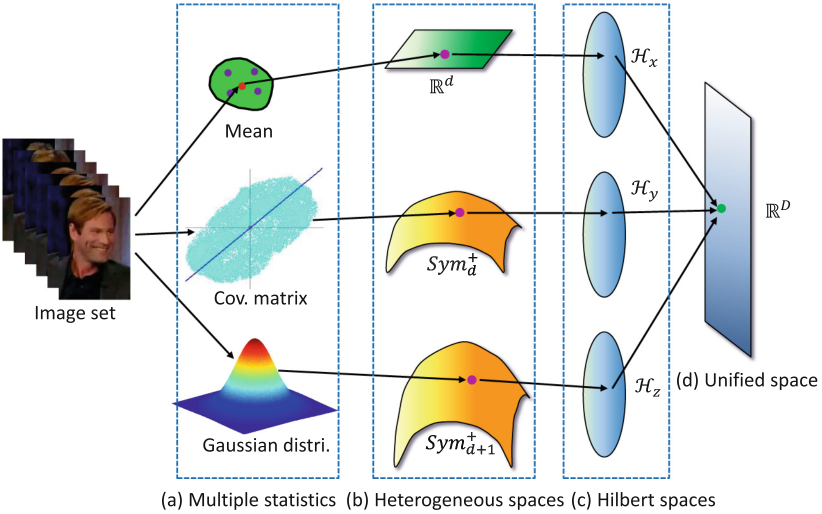

Fig. 1. Conceptual illustration of the proposed Hybrid Euclidean-and-Riemannian Metric Learning framework for image set classification. (a) We first model each image set by its sample mean, covariance matrix and Gaussian distribution. (b) Then we embed the space of them into one Euclidean space $\mathbb{R}^d$ and two Riemannian manifolds Sym_d^+, Sym_{d+1}^+ respectively. Finally, by further embedding such heterogeneous spaces into Hilbert spaces (c), the hybrid points are unified in a common subspace (d) by our proposed hybrid metric learning framework.

3 Proposed Method

In this section, we first describe an overview of our proposed approach for image set classification. Then, we introduce multiple statistics for set modeling from a view of probability statistics, followed by embedding them into multiple heterogeneous spaces, i.e., one Euclidean space and two different Riemannian manifolds. Subsequently, we present the Hybrid Euclidean-and-Riemannian Metric Learning (HERML) for fusing such statistics lying in heterogeneous spaces. Finally, we give a discussion about other related work.

3.1 Overview

This paper proposes a novel Hybrid Euclidean-and-Riemannian Metric Learning (HERML) approach for more robust image set classification. As discussed in the prior sections, simultaneously exploiting the multiple statistics may improve the performance of image set classification. With this in mind, we represent each image set with multiple statistics– mean, covariance matrix and Gaussian distribution. For such different statistics, we study their spanned heterogeneous spaces: one Euclidean space $\mathbb{R}^d$ and two Riemannian manifolds Sym_d^+, Sym_{d+1}^+ respectively. Therefore, we then formulate the problem as fusing points in such three heterogeneous spaces spanned by our employed multiple statistics. Since classical multiple kernel learning algorithms cannot take hybrid Euclidean-and-Riemannian points as their direct inputs, we explore an efficient hybrid metric learning framework to fuse the multiple Euclidean-and-Riemannian points by employing the classical Euclidean and Riemannian kernel. A conceptual illustration of our approach is shown in Fig. 1.

3.2 Multiple Statistics Modeling

Let $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ be the data matrix of an image set with n samples, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes the i-th image sample with d-dimensional feature representation. From a naive probability statistics perspective, we model each set as the following three statistics with different properties: sample-based, subspace-based and distribution-based statistics.

Sample-based statistics (SAS): Given a set of samples characterized by certain probability distribution, mean value is often used as the sample-based statistics to measure the central tendency of the set of samples. Specifically, the mean vector $\boldsymbol{m}$ of one set containing n samples shows the averaged position of the set in the high dimensional space and is computed as:

$$\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \tag{5}$$

Subspace-based statistics (SUS): Since the covariance matrix can be eigen-decomposed into the subspace spanned by the set of samples, it can be considered as the subspace-based statistics, which models the variations of the set data and makes no assumption about the data distribution. Given one set with n samples, the covariance matrix is calculated as:

$$\boldsymbol{C} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T \tag{6}$$

Distribution-based statistics (DIS): In probability theory, the Gaussian (or normal) distribution is a very commonly occurring probability distribution, which is a continuous distribution with the maximum entropy for a given mean and variance. Therefore, we can model the data distribution of set as a Single Gaussian Model (SGM) with estimated mean $\tilde{\boldsymbol{m}}$ and covariance matrix $\tilde{\boldsymbol{C}}$:

$$\boldsymbol{x} \sim \mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}}) \tag{7}$$

3.3 Heterogeneous Space Embedding

As well known, the mean vector lies in Euclidean space $\mathbb{R}^d$, where d is the dimension of the samples. Nevertheless, as studied in [19,21], the covariance matrix resides on Riemannian manifold Sym_d^+. Based on the information geometry [22,26] theory, we can embed the space of Gaussian distribution into a

Riemannian manifold Sym^+_{d+1}. In this case, our defined multiple statistics are in one Euclidean space and two different dimensional Riemannian manifolds respectively.

In the information geometry, if the random vector $\boldsymbol{s}$ follows $\mathcal{N}(0, \boldsymbol{I})$, then its affine transformation $\boldsymbol{Qx} + \tilde{\boldsymbol{m}}$ follows $\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}})$, where $\tilde{\boldsymbol{C}}$ has a decomposition $\tilde{\boldsymbol{C}} = \boldsymbol{QQ}^T, |\boldsymbol{Q}| > 0$, and vice versa. As such the $\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}})$ can be characterized by the affine transformation $(\tilde{\boldsymbol{m}}, \boldsymbol{Q})$. Let τ_1 be the mapping from the affine group $Aff^+_d = \{(\tilde{\boldsymbol{m}}, \boldsymbol{Q}) | \tilde{\boldsymbol{m}} \in \mathbb{R}^d, \boldsymbol{Q} \in \mathbb{R}^{d \times d}, |\boldsymbol{Q}| > 0\}$ to the simple Lie group $Sl_{d+1} = \{\boldsymbol{V} | \boldsymbol{V} \in \mathbb{R}^{(d+1)\times(d+1)}, |\boldsymbol{V}| > 0\}$ as:

$$\tau_1 : Aff^+_d \mapsto Sl_{d+1}, \quad (\tilde{\boldsymbol{m}}, \boldsymbol{Q}) \mapsto |\boldsymbol{Q}|^{-\frac{1}{d+1}} \begin{bmatrix} \boldsymbol{Q} & \tilde{\boldsymbol{m}} \\ \tilde{\boldsymbol{m}}^T & 1 \end{bmatrix} \tag{8}$$

Then we denote τ_2 as the mapping from Sl_{d+1} to the space of SPD matrices $Sym^+_{d+1} = \{\boldsymbol{P} | \boldsymbol{P} \in \mathbb{R}^{(d+1)\times(d+1)}, |\boldsymbol{P}| > 0\}$, i.e.,

$$\tau_2 : Sl_{d+1} \mapsto Sym^+_{d+1}, \quad \boldsymbol{V} \mapsto \boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} \tag{9}$$

Through the two mappings, a d-dimensional Gaussian $\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}})$ can be embedded into Sym^+_{d+1} and thus is uniquely represented by a $(d + 1) \times (d + 1)$ SPD matrix $\boldsymbol{P}$ as:

$$\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}}) \sim \boldsymbol{P} = |\boldsymbol{Q}|^{-\frac{2}{d+1}} \begin{bmatrix} \boldsymbol{QQ}^T + \tilde{\boldsymbol{m}}\tilde{\boldsymbol{m}}^T & \tilde{\boldsymbol{m}} \\ \tilde{\boldsymbol{m}}^T & 1 \end{bmatrix} \tag{10}$$

For detailed theory on the embedding process, please kindly refer to [26].

3.4 Hybrid Euclidean-and-Riemannian Metric Learning

Denote $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N]$ as the training set formed by N image sets, where $\boldsymbol{X}_i = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_i}] \in \mathbb{R}^{n_i \times d}$ indicates the i-th image set, $1 \le i \le N$, and n_i is the number of samples in this image set. It is known that the kernel function is always defined by first mapping the original features to a high dimension Hilbert space, that is $\phi : \mathbb{R}^d \to \mathcal{F}$ (or $Sym^+ \to \mathcal{F}$), and then calculating the dot product of high dimensional features $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Phi}_j$ in the new space. Though the mapping ϕ is usually implicit, we first consider it as an explicit mapping for simplicity. Hence, we first use $\boldsymbol{\Phi}^r_i$ as the high dimensional feature of r-th statistic feature extracted from the image set $\boldsymbol{X}_i$. Here, $1 \le r \le R$ and R is the number of statistics being used, which is 3 in the setting of our multiple statistics modeling. Now, given a pair of training sets $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ with the r-th statistic features $\boldsymbol{\Phi}^r_i, \boldsymbol{\Phi}^r_j$, we define the distance metric as:

$$d_{A_r}(\boldsymbol{\Phi}^r_i, \boldsymbol{\Phi}^r_j) = tr(\boldsymbol{A}_r(\boldsymbol{\Phi}^r_i - \boldsymbol{\Phi}^r_j)(\boldsymbol{\Phi}^r_i - \boldsymbol{\Phi}^r_j)^T) \tag{11}$$

where $\boldsymbol{A}_r$ is the learned Mahalanobis matrix for the r-th statistic in the high dimensional Hilbert space.

By assuming the high dimensional features of multiple statistics can be mapped to a common space, we can jointly optimize the unknown $\boldsymbol{A}_r$ $(r = 1, \ldots, R)$ for the multiple statistics lying in multiple Hilbert spaces. To learn these distance metrics, we attempt to maximize inter-class variations and minimize the intra-class variations with the regularizer of the LogDet divergence, which usually prevents overfitting due to the small training set and high model complexity. In addition, as stated in [25], the LogDet divergence forces the learned Mahalanobis matrices to be close to an initial Mahalanobis matrix and keep symmetric positive definite during the optimization. The objective function for our multiple metric learning problem is formulated as:

$$\begin{aligned} \min_{\boldsymbol{A_1}\succeq 0,\ldots,\boldsymbol{A_R}\succeq 0,\boldsymbol{\xi}} \quad & \frac{1}{R}\sum_{r=1}^{R} D_{\ell d}(\boldsymbol{A}_r,\boldsymbol{A}_0)+\gamma D_{\ell d}(diag(\boldsymbol{\xi}),diag(\boldsymbol{\xi}_0)),\\ s.t. \quad & \frac{\boldsymbol{\delta}_{ij}}{R}\sum_{r=1}^{R} d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)\le \boldsymbol{\xi}_{ij},\forall(i,j). \end{aligned} \tag{12}$$

where $d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)$ is obtained in Eq. 11 and $\boldsymbol{\xi}$ is initialized as $\boldsymbol{\xi}_0$, which is a vector with each elements equal to $\boldsymbol{\delta}_{ij}\rho-\zeta\tau$, ρ is the threshold for distance comparison, τ is the margin, ζ is the tuning scale of the margin. Another variable $\boldsymbol{\delta}_{ij}=1$ if the pair of samples come from the same class, otherwise $\boldsymbol{\delta}_{ij}=-1$. Since each Mahalanobis matrix $\boldsymbol{A}_r$ is symmetric and positive semi-definite, we can seek a non-square matrix $\boldsymbol{W}_r=[\boldsymbol{w}_1^r,\ldots,\boldsymbol{w}_{d_r}^r]$ by calculating the matrix square root $\boldsymbol{A}_r=\boldsymbol{W}_r\boldsymbol{W}_r^T$.

In general, because the form of ϕ^r is usually implicit, it is hard or even impossible to compute the distance $d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)$ in Eq. 11 directly in the Hilbert space. Hence, we use the kernel trick method [27] by expressing the basis $\boldsymbol{w}_k^r$ as a linear combination of all the training samples in the mapped space as:

$$\boldsymbol{w}_k^r=\sum_{j=1}^{N}\boldsymbol{u}_j^k\boldsymbol{\Phi}_j^r \tag{13}$$

where $\boldsymbol{u}_j^k$ are the expansion coefficients. Hence,

$$\sum_{r=1}^{R}(\boldsymbol{w}_k^r)^T\boldsymbol{\Phi}_i^r=\sum_{r=1}^{R}\sum_{j=1}^{N}\boldsymbol{u}_j^k(\boldsymbol{\Phi}_j^r)^T\boldsymbol{\Phi}_i^r=\sum_{r=1}^{R}(\boldsymbol{u}^k)^T\boldsymbol{K}_{.i}^r \tag{14}$$

where $\boldsymbol{u}^k$ is an $N\times 1$ column vector and its j-th entry is $\boldsymbol{u}_j^k$, and $\boldsymbol{K}_{.i}^r$ is the i-th column of the r-th kernel matrix $\boldsymbol{K}^r$. Here $\boldsymbol{K}^r$ is an $N\times N$ kernel matrix, calculated from the r-th statistic feature using the Euclidean kernel functions $\boldsymbol{\kappa}_m(\boldsymbol{m}_i,\boldsymbol{m}_j)=\boldsymbol{m}_i^T\boldsymbol{m}_j$ or Riemannian kernel functions in Eq. 2 for different set statistic features. If we denote Mahalanobis matrices as $\boldsymbol{B}_r=\boldsymbol{U}_r\boldsymbol{U}_r^T$ for $1\le r\le R$, then Eq. 12 can be rewritten as:

$$\begin{aligned} \min_{\boldsymbol{B_1}\succeq 0,\ldots,\boldsymbol{B_R}\succeq 0,\boldsymbol{\xi}} \quad & \frac{1}{R}\sum_{r=1}^{R} D_{\ell d}(\boldsymbol{B}_r,\boldsymbol{B}_0)+\gamma D_{\ell d}(diag(\boldsymbol{\xi}),diag(\boldsymbol{\xi}_0)),\\ s.t. \quad & \frac{\boldsymbol{\delta}_{ij}}{R}\sum_{r=1}^{R} d_{B_r}(\boldsymbol{K}_{.i}^r,\boldsymbol{K}_{.j}^r)\le \boldsymbol{\xi}_{ij},\forall(i,j). \end{aligned} \tag{15}$$

where $d_{B_r}(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j})$ indicates the distance between the i-th and j-th samples under the learned metric $\boldsymbol{B_r}$ for the r-th statistic mapping in the Hilbert space:

$$d_{B_r}(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j}) = tr(\boldsymbol{B}_r(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})^T) \tag{16}$$

3.5 Optimization

To solve the problem in Eq. 15, we adopt the cyclic Bregman projection method [28,29], which is to choose one constraint per iteration, and perform a projection so that the current solution satisfies the chosen constraint. In the case of inequality constraints, appropriate corrections of $\boldsymbol{B}_r$ and $\boldsymbol{\xi}_{ij}$ are also enforced. This process is then repeated by cycling through the constraints. The method of cyclic Bregman projections is able to converge to the globally optimal solution. Please kindly refer to [28,29] for more details. The updating rules for our proposed method are shown in the following proposition:

Proposition 1. *Given the solution $\boldsymbol{B}^t_r$ for $r = 1, \ldots, R$ at the t-th iteration, we update $\boldsymbol{B}_r$ and the corresponding $\boldsymbol{\xi}_{ij}$ as follows:*

$$\begin{cases} \boldsymbol{B^{t+1}_r} = \boldsymbol{B}^t_r + \beta_r \boldsymbol{B}_r(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})^T \boldsymbol{B}_r, & (17) \\ \boldsymbol{\xi}^{t+1}_{ij} = \dfrac{\gamma \boldsymbol{\xi}^t_{ij}}{\gamma + \boldsymbol{\delta}_{ij} \alpha \boldsymbol{\xi}^t_{ij}}, & (18) \end{cases}$$

where $\beta_r = \boldsymbol{\delta}_{ij}\alpha / (1 - \boldsymbol{\delta}_{ij}\alpha d_{B^t_r}(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j}))$ and α can be solved by:

$$\frac{\boldsymbol{\delta}_{ij}}{R} \sum_{r=1}^{R} \frac{d_{B^t_r}(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j})}{1 - \boldsymbol{\delta}_{ij}\alpha d_{B^t_r}(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j})} - \frac{\gamma \boldsymbol{\xi}^t_{ij}}{\gamma + \boldsymbol{\delta}_{ij}\alpha \boldsymbol{\xi}^t_{ij}} = 0. \tag{19}$$

Proof. Based on the cyclic projection method [28,29], we formulate the Lagrangian form of Eq. 15 and set the gradients to zero w.r.t $\boldsymbol{B}^{t+1}_r$, $\boldsymbol{\xi}^{t+1}_{ij}$ and α to get the following update equations:

$$\begin{cases} \nabla D(\boldsymbol{B}^{t+1}_r) = \nabla D(\boldsymbol{B}^t_r) + \boldsymbol{\delta_{ij}}\alpha(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})^T, & (20) \\ \nabla D(\boldsymbol{\xi}^{t+1}_{ij}) = \nabla D(\boldsymbol{\xi}^t_{ij}) - \dfrac{\boldsymbol{\delta}_{ij}\alpha}{\gamma}, & (21) \\ \dfrac{\boldsymbol{\delta}_{ij}}{R} \displaystyle\sum_{r=1}^{R} tr(\boldsymbol{B}^{t+1}_r(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})(\boldsymbol{K}^r_{.i} - \boldsymbol{K}^r_{.j})^T) = \boldsymbol{\xi}^{t+1}_{ij}. & (22) \end{cases}$$

Then, we can derive Eq. 17 and Eq. 18 from Eq. 20 and Eq. 21, respectively. Substituting Eqs. 17 and 18 into Eq. 22, we obtain the Eq. 19 related to α.

The resulting algorithm is given as Algorithm 1. The inputs to the algorithm are the starting Mahalanobis matrices $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R$, the constraint data, the slack parameter γ, distance threshold ρ, margin parameter τ and tuning scale ζ. If necessary, the projections can be computed efficiently over a factorization $\boldsymbol{U}$ of each Mahalanobis matrix, such that $\boldsymbol{B}_r = \boldsymbol{U}^T_r \boldsymbol{U}_r$. The main time cost is to update $\boldsymbol{B}^{t+1}_r$ in Step 5, which is $O(RN^2)$ (N is the number of samples) for each constraint projection. Therefore, the total time cost is $O(LRN^2)$ where L is the total number of the updating in Step 5 executed by the algorithm.

Algorithm 1. Hybrid Euclidean-and-Riemannian Metric Learning

```
Input: Training pairs $\{(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j}), \boldsymbol{\delta}_{ij}\}$, and slack parameter $\gamma$, input Mahalanobis matrix $\boldsymbol{B}_0$, distance thresholds $\rho$, margin parameter $\tau$ and tuning scale $\zeta$
$t \leftarrow 1$, $\boldsymbol{B}^1_r \leftarrow \boldsymbol{B}_0$ for $r = 1, \ldots, R$, $\boldsymbol{\lambda}_{ij} \leftarrow 0$, $\boldsymbol{\xi}_{ij} \leftarrow \boldsymbol{\delta_{ij}}\rho - \zeta\tau, \forall(i,j)$
Repeat
Pick a constraint $(i,j)$ and compute the distances $d_{B^t_r}(\boldsymbol{K}^r_{.i}, \boldsymbol{K}^r_{.j}))$ for $r = 1, \ldots, R$.
Solve $\alpha$ in Eq.19 and set $\alpha \leftarrow min(\alpha, \boldsymbol{\eta}_{ij})$ and $\boldsymbol{\eta}_{ij} \leftarrow \boldsymbol{\eta}_{ij} - \alpha$
Update $\boldsymbol{B}^{t+1}_r$ by using Eq. 17 for $r = 1, \ldots, R$.
Update $\boldsymbol{\xi}^{t+1}_{ij}$ by using Eq. 18.
Until convergence
Output: Mahalanobis matrices $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R$.
```

3.6 Discussion About Related Work

The original kernelized version of ITML [25] method implicitly solves the metric learning problem in a single high dimensional Hilbert space by learning the optimal kernel matrix $\boldsymbol{K}^*$. In contrast, our proposed method explicitly learns multiple metrics $\{\boldsymbol{B}^*_1, \ldots, \boldsymbol{B}^*_R\}$ on multiple Hilbert spaces for fusing hybrid Euclidean-and-Riemannian features. To some extent, our proposed metric learning framework is a generalized version of ITML. When the type of kernel function is linear and meanwhile the data lie in a single space, the proposed framework can be reduced to the original ITML.

In addition, there are a couple of previous works [13,30–35] for multiple kernel/metric learning in the literature. Nevertheless, most of these works mainly focus on fusing multiple homogeneous Euclidean (or Riemannian) features, while our method attempts to study the new problem of learning hybrid metrics for fusing heterogeneous Euclidean and Riemannian features. Thus, their problem domains are different from ours.

4 Experiments

In this section, we evaluate our proposed approach on two image set classification applications: set-based object categorization and video-based face recognition. The following describes the experiments and results.

4.1 Databases and Settings

For the set-based object categorization task, we use the database ETH-80 [36]. It consists of 8 categories of objects with each category including 10 object instances. Each object instance has 41 images of different views from one set. The task is to classify an image set of an object into a known category. The images were resized to 20 × 20 as [12,13] and the intensities were used for features.

For the video-based face recognition task, we consider two public datasets: YouTube Celebrities [37] and COX [38]. The YouTube is a quite challenging and widely used video face dataset. It has 1,910 video clips of 47 subjects collected from YouTube. Most clips contains hundreds of frames, which are often low resolution and highly compressed with noise and low quality. The COX is a large scale video dataset involving 1,000 different subjects, each of which has 3 videos captured by different camcorders. In each video, there is around $25 \sim 175$ frames of low resolution and low quality, with blur, and captured under poor lighting. Each face in YouTube was resized to a 20×20 image as [12,13] while the faces in COX were resized to 32×40. For all faces in the two datasets, histogram equalization was implemented to eliminate lighting effects.

On the three datasets, we followed the same protocol as the prior work [3,12,13], which conducted ten-fold cross validation experiments, i.e., 10 randomly selected gallery/probe combinations. Finally, the average recognition rates of different methods were reported. Specifically, for ETH-80, each category had 5 objects for gallery and the other 5 objects for probes. For YouTube, in each fold, one person had 3 randomly chosen image sets for the gallery and 6 for probes. Different from ETH and YouTube, COX dataset does also contain an additional independent training set [38], where each subject has 3 videos. Since there are 3 independent testing sets of videos in COX, each person had one video as the gallery and the remaining two videos for two different probes, thus in total 6 groups of testing need to be conducted.

4.2 Comparative Methods and Settings

We compared our approach with three categories of the state-of-the-art image set classification methods as following. Note that, we add ITML to sample-based methods as it performs metric learning on single samples/images, which can be considered as a kind of sample-based statistics here. Since ITML also has a kernel version, we feed our proposed kernel function of distribution-based statistics (DIS) to it for additional comparison.

1. Sample-based method:
   Maximum Mean Discrepancy (MMD)[2], Affine (Convex) Hull based Image Set Distance (AHISD, CHISD)[3], Set-to-Set Distance Metric Learning (SSDML) [7] and Information Theoretic Metric Learning (ITML)[25].
2. Subspace-based method:
   Mutual Subspace Method (MSM) [8], Discriminant Canonical Correlations (DCC)[1], Manifold Discriminant Analysis (MDA)[11], Grassmann Discriminant Analysis (GDA) [10], Covariance Discriminative Learning (CDL)[12] and Localized Multi-Kernel Metric Learning (LMKML)[13].
3. Distribution-based method:
   Single Gaussian Models (SGM) [14], Gaussian Mixture Models (GMM) [15] and kernel version of ITML [25] with our DIS-based set model (DIS-ITML).

Except SGM and GMM, the source codes of above methods are provided by the original authors. Since the codes of SGM and GMM are not publicly available, we carefully implemented them using the code[1] to generate Gaussian

[1] https://engineering.purdue.edu/~bouman/software/cluster/.

model(s). For fair comparison, the important parameters of each method were empirically tuned according to the recommendations in the original references: For MMD, we used the edition of Bootstrap and set the parameters $\alpha = 0.1, \sigma = -1$, the number of iteration to 5. For ITML, we used the default parameters as the standard implementation. For AHISD, CHISD and DCC, PCA was performed by preserving 95 % energy to learn the linear subspace and corresponding 10 maximum canonical correlations were used. For MDA, the parameters were configured according to [11]. For GDA, the dimension of Grassmannian manifold is set to 10. For CDL, since KPLS works only when the gallery data is used for training, the setting of COX prevent it from working. So, we use KDA for discriminative learning and adopt the same setting as [12]. For SSDML, we set $\lambda_1 = 0.001, \lambda_2 = 0.5$, numbers of positive and negative pairs per set is set to 10 and 20. For LMKML, we used median distance heuristic to tune the widths of Gaussian kernels. For our method HERML[2], we set the parameters $\gamma = 1$, ρ as the mean distances, τ as the standard variations and the tuning range of ζ is $[0.1, 1]$.

4.3 Results and Analysis

We present the rank-1 recognition results of comparative methods on the three datasets in Table 2. Each reported rate is an average over the ten-fold trials. Note that, since the LMKML method is too time-consuming to run in the setting of COX dataset, which has a large scale dataset, we alternately use 100 of 300 subject's data for training and 100 of 700 remaining subject's sets for testing.

Firstly, we are interested in the classification results of methods with different degree of match. Here, we focus on the comparison between those unsupervised methods MMD, AHISD, CHISD, MSM, SGM, GMM. On the ETH-80, the subspace-based method MSM and the distribution-based methods SGM, GMM outperform the sample-based methods MMD, AHISD, CHISD. This is mainly because the ETH-80 contains many sets of large variations. In this setting, MSM, SGM and GMM can capture the pattern variations, which are more robust to outlier and noise than MMD, AHISD and CHISD. In other two datasets, YouTube and COX, it is also reasonable that the three kinds of methods achieve comparable results for their used statistics are all effective for set modeling.

Secondly, we also care about which way to learn a discriminant function is more effective. So, we compare the results of the supervised methods SSDML, ITML, DCC, MDA, GDA, CDL. On the three datasets, GDA and CDL methods have clear advantage over SSDML, ITML, DCC and MDA. This is because ITML performs the metric learning and classification on single samples, which neglects the specific data structure of sets. SSDML, DCC and MDA methods learn the discriminant metrics in Euclidean space, whereas most of them classify the sets in non-Euclidean spaces. In contrast, GDA and CDL extract the subspace-based statistics in Riemannian space and match them in the same space, which is more favorable for the set classification task [10].

[2] The source code is released on the website: http://vipl.ict.ac.cn/resources/codes.

Table 2. Average recognition rate (%) of different image set classification methods on ETH-80, YouTube and COX-S2V datasets. Here, COX-ij represent the test using the i-th set of videos as gallery and the j-th set of videos as probe.

Method	ETH-80	YouTube	COX-12	COX-13	COX-23	COX-21	COX-31	COX-32
MMD [2]	77.5	52.6	36.4	19.6	8.90	27.6	19.1	9.60
AHISD [3]	77.3	63.7	53.0	36.1	17.5	43.5	35.0	18.8
CHISD [3]	73.5	66.3	56.9	30.1	15.0	44.4	26.4	13.7
SSDML [7]	80.0	68.8	60.1	53.1	28.7	47.9	44.4	27.3
ITML [25]	77.2	65.3	50.9	46.0	35.6	39.6	37.1	34.8
MSM [8]	87.8	61.1	45.5	21.5	11.0	39.8	19.4	9.50
DCC [1]	90.5	64.8	62.5	66.1	50.6	56.1	64.8	45.2
MDA [11]	89.0	65.3	65.8	63.0	36.2	55.5	43.2	30.0
GDA [10]	92.3	65.9	68.6	77.7	71.6	66.0	76.1	74.8
CDL [12]	**94.5**	69.7	**78.4**	**85.3**	**79.7**	**75.6**	**85.8**	**81.9**
LMKML [13]	90.0	**70.3**	66.0	71.0	56.0	74.0	68.0	60.0
SGM [14]	81.3	52.0	26.7	14.3	12.4	26.0	19.0	10.3
GMM [15]	89.8	61.0	30.1	24.6	13.0	28.9	31.7	18.9
DIS-ITML [25]	87.8	68.4	47.9	48.9	36.1	43.1	35.6	33.6
HERML	**94.5**	**74.6**	**94.9**	**96.9**	**94.0**	**92.0**	**96.4**	**95.3**

Thirdly, we compare the state-of-the-art methods with our approach and find they are impressively outperformed by ours on the three datasets. Several reasons are figured out as following: In terms of set modeling, as stated in Sect. 1, our combining of multiple complementary statistics can more robustly model those sets of arbitrary distribution, large variation and small size in the three datasets. In terms of discriminant function learning, by encoding the heterogeneous structure of the space of such statistics, our method jointly learns hybrid metrics to fuse them for more discriminant classification. In comparison, LMKML neglects the non-Euclidean data structure of two higher order statistics, i.e., the covariance matrix and the tensor. Thus, our proposed method is more desirable to learn metrics for non-Euclidean data and has a clear advantage over LMKML. In addition, the results also shows that our novel hybrid metric learning method has an impressive superiority over the original ITML.

In addition, we also compare the discriminative power of our proposed sample-based, subspace-based and distribute-based statistics (SAS, SUS, DIS) for image set classification. For each statistic, we performed our proposed method to train and classify sets with NN classifier. Table 3 tabulates the classification rates of multiple statistics. We can observe that the DIS achieves the best recognition performance than other two statistics because it jointly model the mean and the covariance matrix in a Gaussian distribution. Additionally, the results of combining of SAS and SUS sometimes are better than those of DIS on COX-S2V. This is because the dataset may contain some sets not in Gaussian distribution. Since the multiple statistics complement each other, the performance can be improved by our proposed metric learning with all of statistic models.

Table 3. Average recognition rate (%) of different statistics (SAS, SUS, DIS), combining SAS and SUS (SAS+SUS), fusing all multiple statistics (ALL) with our metric learning method on ETH-80, YouTube and COX-S2V. Here, COX-ij indicates the test using the i-th set of videos as gallery and the j-th set of videos as probe.

Statistics	ETH-80	YouTube	COX-12	COX-13	COX-23	COX-21	COX-31	COX-32
SAS	83.5	64.1	86.2	92.0	82.8	83.2	86.9	84.9
SUS	93.5	70.2	88.8	93.6	90.3	86.4	94.0	93.1
DIS	**94.3**	**73.5**	92.8	94.7	92.2	89.0	94.7	94.4
SAS+SUS	92.0	71.6	**93.1**	**95.2**	**93.1**	**91.2**	**95.2**	**95.0**
ALL	**94.5**	**74.6**	**94.9**	**96.9**	**94.0**	**92.0**	**96.4**	**95.3**

Table 4. Computation time (seconds) of different methods on the YouTube dataset for training and testing (classification of one video).

Method	MMD	SSDML	ITML	DCC	CDL	LMKML	SGM	GMM	**HERML**
Train	N/A	433.3	2459.7	11.9	4.3	17511.2	N/A	N/A	27.3
Test	0.1	2.6	0.5	0.1	0.1	247.1	0.4	1.9	0.1

Lastly, on the YouTube dataset, we compared the computational complexity of different methods on an Intel(R) Core(TM) i7-3770 (3.40 GHz) PC. Table 4 lists the time cost for each method. The presentation of training time is only required by discriminant methods. For testing, we report the classification time of one video. Since ITML has to train and test on large number of samples from sets and classify pairs of samples, it has high time complexities. Except DCC and CDL, our method is much faster than other methods especially the LMKML method. This is because it transformed the covariance matrices and third-order tensors to vectors, which lies in very high dimension Euclidean spaces. As a result, it is very time-consuming to perform metric learning and classification.

5 Conclusions

In this paper, we proposed a novel hybrid Euclidean-and-Riemannian metrics method to fuse multiple complementary statistics for robust image set classification. The extensive experiments have shown that our proposed method outperforms the state-of-the-art methods in both terms of accuracy and efficiency. To our best knowledge, the problem of hybrid metric learning across Euclidean and Riemannian spaces has not been investigated before and we made the first attempt to address this issue in this paper. In the future, it would be interesting to explore other possible metric learning methods to fuse multiple complement statistics or pursue more robust statistics to model image sets with different structures in real-world scenario.

Acknowledgement. The work is partially supported by Natural Science Foundation of China under contracts nos.61390511, 61379083, and 61222211.

References

1. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. PAMI **29**, 1005–1018 (2007)
2. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. JMLR **13**, 723–773 (2012)
3. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: CVPR (2010)
4. Hu, Y., Mian, A., Owens, R.: Sparse approximated nearest points for image set classification. In: CVPR (2011)
5. Yang, M., Zhu, P., Gool, L., Zhang, L.: Face recognition based on regularized nearest points between image sets. In: FG (2013)
6. Huang, Z., Zhao, X., Shan, S., Wang, R., Chen, X.: Coupling alignments with recognition for still-to-video face recognition. In: ICCV (2013)
7. Zhu, P., Zhang, L., Zuo, W., Zhang, D.: From point to set: extend the learning of distance metrics. In: ICCV (2013)
8. Yamaguchi, O., Fukui, K., Maeda., K.: Face recognition using temporal image sequence. In: FG (1998)
9. Wang, R., Shan, S., Chen, X., Dai, Q., Gao, W.: Manifold-Manifold distance and its application to face recognition with image sets. IEEE Trans. Image Proces. **21**, 4466–4479 (2012)
10. Hamm, J., Lee, D.D.: Grassmann discriminant analysis: a unifying view on subspace-based learning. In: ICML, pp. 376–383 (2008)
11. Wang, R., Chen, X.: Manifold discriminant analysis. In: CVPR (2009)
12. Wang, R., Guo, H., Davis, L., Dai, Q.: Covariance discriminative learning: a natural and efficient approach to image set classification. In: CVPR (2012)
13. Lu, J., Wang, G., Moulin, P.: Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In: ICCV (2013)
14. Shakhnarovich, G., Fisher III, J.W., Darrell, T.: Face recognition from long-term observations. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 851–865. Springer, Heidelberg (2002)
15. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: CVPR (2005)
16. Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 312–377 (1936)
17. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
18. Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C.: Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In: CVPR (2011)
19. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. IJCV **66**, 41–66 (2006)
20. Tuzel, O., Porikli, F., Meer, P.: Region covariance: a fast descriptor for detection and classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part II. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006)
21. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. **29**, 328–347 (2007)

22. Amari, S.I., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Oxford (2000)
23. Huang, Z., Wang, R., Shan, S., Chen, X.: Learning Euclidean-to-Riemannian metric for point-to-set classification. In: CVPR (2014)
24. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In: CVPR (2013)
25. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML (2007)
26. Lovrić, M., Min-Oo, M., Ruh, E.A.: Multivariate normal distributions parametrized as a Riemannian symmetric space. J. Multivar. Anal. **74**, 36–48 (2000)
27. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Comput. **12**, 2385–2404 (2000)
28. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**, 200–217 (1967)
29. Censor, Y., Zenios, S.: Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press, Oxford (1997)
30. Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: SimpleMKL. J. Mach. Learn. Res. (JMLR) **9**, 2491–2521 (2008)
31. McFee, B., Lanckriet, G.: Learning multi-modal similarity. JMLR **12**, 491–523 (2011)
32. Xie, P., Xing, E.P.: Multi-modal distance metric learning. In: IJCAI (2013)
33. Vemulapalli, R., Pillai, J.K., Chellappa, R.: Kernel learning for extrinsic classification of manifold features. In: CVPR (2013)
34. Cui, Z., Li, W., Xu, D., Shan, S., Chen, X.: Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In: CVPR (2013)
35. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Combining multiple manifold-valued descriptors for improved object recognition. In: DICTA (2013)
36. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: CVPR (2003)
37. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos. In: CVPR (2008)
38. Huang, Z., Shan, S., Zhang, H., Lao, S., Kuerban, A., Chen, X.: Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 589–600. Springer, Heidelberg (2013)