Pattern Recognition 48 (2015) 3113–3124

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning

Zhiwu Huang [a,b], Ruiping Wang [a], Shiguang Shan [a,*], Xilin Chen [a]

[a] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form
12 February 2015
Accepted 13 March 2015
Available online 20 March 2015

Keywords:
Face recognition
Large-scale video
Multiple heterogeneous statistics
Hybrid Euclidean-and-Riemannian metric learning

ABSTRACT

Face recognition on large-scale video in the wild is becoming increasingly important due to the ubiquity of video data captured by surveillance cameras, handheld devices, Internet uploads, and other sources. By treating each video as one image set, set-based methods recently have made great success in the field of video-based face recognition. In the wild world, videos often contain extremely complex data variations and thus pose a big challenge of set modeling for set-based methods. In this paper, we propose a novel Hybrid Euclidean-and-Riemannian Metric Learning (HERML) method to fuse multiple statistics of image set. Specifically, we represent each image set simultaneously by mean, covariance matrix and Gaussian distribution, which generally complement each other in the aspect of set modeling. However, it is not trivial to fuse them since mean, covariance matrix and Gaussian model typically lie in multiple heterogeneous spaces equipped with Euclidean or Riemannian metric. Therefore, we first implicitly map the original statistics into high dimensional Hilbert spaces by exploiting Euclidean and Riemannian kernels. With a LogDet divergence based objective function, the hybrid kernels are then fused by our hybrid metric learning framework, which can efficiently perform the fusing procedure on large-scale videos. The proposed method is evaluated on four public and challenging large-scale video face datasets. Extensive experimental results demonstrate that our method has a clear superiority over the state-of-the-art set-based methods for large-scale video-based face recognition.

1. Introduction

With the increased usage of surveillance and personal video capturing devices, enormous amount of video data is being captured everyday. For instance, a large number of videos are being uploaded every day on YouTube alone. Significant amount of data are also captured by the ubiquitous surveillance and handheld cameras. One of the most popular applications on the large-scale video data is video-based face recognition, which commonly identifies a person by matching his/her video sequence taken somewhere against the surveillance videos recorded elsewhere.

In recent years, a large variety of methods have been suggested for the problem of video-based face recognition. Broadly speaking, these methods can be grouped into sequence-based ones and set-based ones [1]. The former methods (e.g. [2–6]) exploit the temporal or dynamic information of the faces in the video, while the latter ones (e.g. [7–11]) represent videos as image sets of the separated video frames, without using the temporal information. In this paper, we focus on the set-based methods due to their less assumption on face video sequence, high-efficiency on large-scale video data as well as their state-of-the-art achievements on the problem of recognizing people in videos. Generally, existing set-based methods mainly focus on the key issues of how to quantify the degree of match between two image sets and how to learn discriminant function from training image sets [8].

In the aspect of how to quantify the degree of match, set-based methods can be broadly partitioned into sample-based methods [10,12–16], subspace-based methods [7–9,17–20] and distribution-based methods [21,22]. Sample-based methods compare sets based on matching their sample-based statistics such as sample mean and affine (convex) combination of samples. This kind of methods includes Maximum Mean Discrepancy (MaxMD) [12], Affine (Convex) Hull based Image Set Distance (AHISD, CHISD) [10] and Sparse Approximated Nearest Point (SANP) [13], etc. In contrast, subspace-based methods typically apply subspace-based statistics to model sets and classify them with given similarity function. For example, Mutual Subspace Method (MSM) [7] represents sets as linear subspaces and match them using canonical correlations [23]. The distribution-based methods, e.g. Single Gaussian Model (SGM) [21] and Gaussian mixture models (GMM) [22], model each set with distribution-based statistics (i.e., Gaussian distribution), and then measure the similarity between two distributions in terms of the Kullback–Leibler Divergence (KLD) [24].

[*] Corresponding author. Tel./fax: +86 10 6260 0558.
E-mail address: sgshan@ict.ac.cn (S. Shan).

http://dx.doi.org/10.1016/j.patcog.2015.03.011
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

In the real-world scenario, video sequences are very likely to cover large variations in a subject's appearance due to camera pose changes, non-rigid deformations, and different illumination conditions. Therefore, data in videos are often of arbitrary distribution, which may cause single type of set model fail to faithfully characterize the set data structure. As shown in Fig. 1(a) and (b), the sample mean encodes the position information of the observed samples by averaging them, while the sample covariance matrix captures the *tight* variation modes of the observed samples by computing the variance between the involved samples and the population mean. Obviously, using either the sample mean (i.e., the position) or the covariance matrix (i.e., the variation) of set data separately can only characterize the set data from one side of the coin. To handle this problem, Gaussian model is commonly employed to simultaneously represent the position and the variation of data samples by estimating their mean and covariance matrix. As studied in [21,22,25], the observation covariance matrix in Gaussian model is typically a maximum-likelihood estimation from the sample covariance matrix, and thus can more *loosely* encode the variation of the data (see Fig. 1(c)). However, Gaussian model assumes the data in each set follow Gaussian distribution, which however cannot always be satisfied in real-world applications. Consequently, if the data follow normal distribution strictly, Gaussian model individually can characterize the set data structure. When the data is of non-normal distribution, the fusion of sample mean and sample covariance would be a better choice to represent each set. To cover all possible cases, in this paper, we combine three such statistics to model the sets from video sequences for more robustness.

Another important problem in set-based methods is how to learn discriminant function from training image sets. The first kind of methods [8,16,18,20] is to learn the discriminant function in Euclidean space. For instance, Set-to-Set Distance Metric Learning (SSDML) [16] 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) [20] treats three order statistics of each set as single vectors in Euclidean spaces and attempts to learn one metric for them by embedding Euclidean spaces into Hilbert spaces. 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 [17,19,26] treat each subspace-based statistics as a point in a specific non-Euclidean space, and perform discriminant function learning in the same space. For example, Grassmann Discriminant Analysis (GDA) [17] and Covariance Discriminative Learning (CDL) [19] represent each linear subspace or covariance matrix as a point on a specific type of Riemannian manifolds and learn discriminant functions on those manifolds.

In this paper, we propose a novel hybrid metric learning approach to combine multiple statistics on set modeling for more robust video face recognition in the wild. Specifically, we model each set by simultaneously fusing sample mean, covariance matrix and Gaussian distribution due to their complementary properties especially in the real-world settings as discussed above. 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 [27–29], the nonsingular 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 [30,31]. In such heterogeneous spaces, inspired by our previous work [32], we first exploit classical Euclidean and Riemmannian metrics to define Euclidean and Riemannian kernels, employing which the heterogeneous statistics are implicitly mapped into high dimension Hilbert spaces. Then, our method jointly learns multiple Mahalanobis matrices to fuse the hybrid Euclidean and Riemannian kernels in a unified framework for more robust video-based face recognition.

In fact, this work is an extension of our previous work [33]. The differences between this work and the conference paper are as follows: (1) this paper extends the Single Gaussian Model (SGM) in the conference version to Gaussian Mixture Model (GMM), which is essentially a general version of SGM, for modeling the Gaussian distribution. (2) Furthermore, we exploit a novel kernel functions for GMM in the Hilbert space embedding to facilitate fusing of our employed statistics in our hybrid metric learning framework. (3) Besides face identification task in the conference version, we evaluate the proposed method in video face verification task by conducting extensive experiments on two public and very challenging large-scale video face datasets: YouTube Face [34] and Point-and-Shoot Face Recognition Challenge [35]. In addition, we also report results of the extended method and present the performances of its each component on all datasets. *In summary, there are three main contributions in this work:*

(1) We represent the image set (one video) simultaneously by three statistics, i.e., sample mean, sample covariance matrix and Gaussian model, and investigate their complementary properties for more robust video-based face recognition.

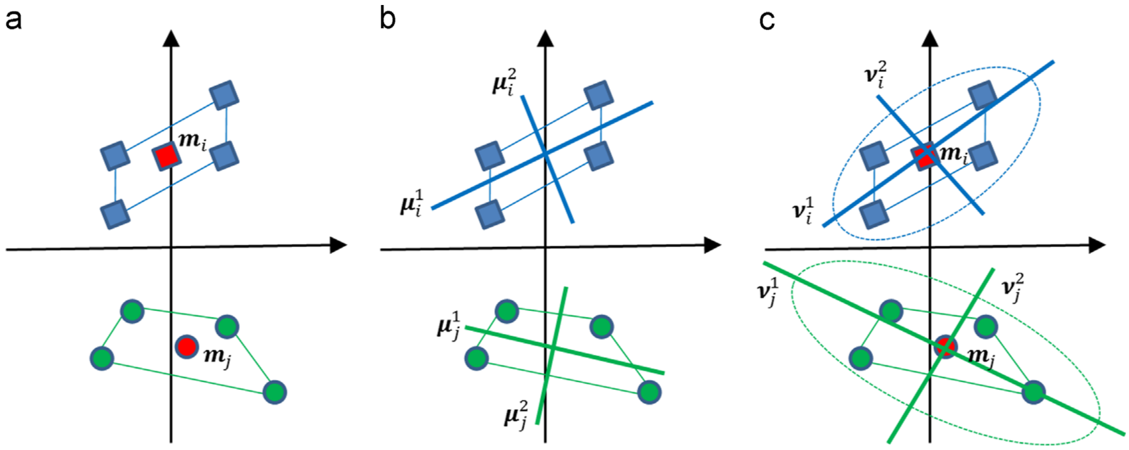

Fig. 1. Illustration of different roles of sample mean, sample covariance matrix and Gaussian model when representing one set. Here, the squares and the circles denote data samples respectively from two different sets. In (a), sample mean m_i/m_j characterizes position of one set. In (b), different directions $\mu_i^1, \mu_i^2/\mu_j^1, \mu_j^2$ of sample covariance matrix encode tight variations of set data. In (c), Gaussian model takes into account the mean and its leading directions $v_i^1, v_i^2/v_j^1, v_j^2$ capture loose variations in one set.

(2) To alleviate the heterogeneity with the other two spaces of mean (i.e., Euclidean space) and covariance matrix (i.e., SPD manifold), we embed the space of Gaussian distribution, which has been commonly studied within the context of statistical manifold in previous literatures, into another specific SPD manifold by exploiting the information theory [31].

(3) To fuse the complementary but heterogeneous statistics, we develop a hybrid metric learning framework to jointly learn three different Mahalanobis matrices respectively in three kernel Hilbert spaces for these statistics. While commonly used kernel functions are employed for mean and covariance matrix, we propose a novel kernel function for Gaussian model by leveraging KL divergence-based kernel approximation theory [36] and the Riemannian geometry of SPD manifold.

The rest of the paper is organized as follows. Section 2 reviews the related work including Information-Theoretic Metric Learning and Multiple Kernel Learning. Section 3 details the proposed Hybrid Euclidean-and-Riemannian Metric Learning (HERML) method to fuse multiple statistics of image sets for more robust video-based face recognition. Section 4 evaluates our proposed method and the state-of-the-art set-based methods on both video-based face identification and video-based face verification, followed by conclusions in Section 5.

2. Background

To learn the hybrid metrics, our new proposed method exploits the Burg divergence based objective function which are employed in Information-Theoretic Metric Learning method [37]. Therefore, in this section, we first introduce the Information-Theoretic Metric Learning method and its kernelized version. Then, since our formulation is to fuse multiple kernels to combine different statistics of image sets, we review the conventional Multiple Kernel Learning methods [20,38–43], which often fuse multiple kernels derived from homogeneous Euclidean spaces with different dimensionalities.

2.1. Information-theoretic metric learning

Information-Theoretic Metric Learning (ITML) [37] method formulates the problem of metric learning as a particular Bregman optimization, which aims to minimize the LogDet divergence subject to linear constraints:

$$\begin{aligned}
&\min_{\boldsymbol{A}\succcurlyeq 0,\boldsymbol{\xi}} \quad D_{\ell d}(\boldsymbol{A},\boldsymbol{A}_0)+\gamma D_{\ell d}(\mathrm{diag}(\boldsymbol{\xi}),\mathrm{diag}(\boldsymbol{\xi}_0))\\
&\text{s.t.} \quad \mathrm{tr}(\boldsymbol{A}(\boldsymbol{x}_i-x_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T)\le \boldsymbol{\xi}_{ij}, \quad (i,j)\in S\\
&\mathrm{tr}(\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T)\ge \boldsymbol{\xi}_{ij}, \quad (i,j)\in D
\end{aligned} \tag{1}$$

where $\boldsymbol{A},\boldsymbol{A}_0\in\mathbb{R}^{d\times d}$, $D_{\ell d}(\boldsymbol{A},\boldsymbol{A}_0)=\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}_0^{-1})-\mathrm{logdet}(\boldsymbol{A}\boldsymbol{A}_0^{-1})-d$, 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 are equal to an 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, i.e., $\boldsymbol{K}_0(i,j)=\phi(\boldsymbol{x}_i)^T\boldsymbol{A}_0\phi(\boldsymbol{x}_j)$, where ϕ is an implicit mapping from the original space to a high dimensional Hilbert space. Note that the Euclidean distance in kernel space may be written as $\boldsymbol{K}(i,i)+\boldsymbol{K}(j,j)-2\boldsymbol{K}(i,j)=\mathrm{tr}(\boldsymbol{K}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^T)$, 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 Hilbert space, 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

$$\begin{aligned}
&\min_{\boldsymbol{K}\succcurlyeq 0,\boldsymbol{\xi}} \quad D_{\ell d}(\boldsymbol{K},\boldsymbol{K}_0)+\gamma D_{\ell d}(\mathrm{diag}(\boldsymbol{\xi}),\mathrm{diag}(\boldsymbol{\xi}_0))\\
&\quad\text{s.t.} \quad \mathrm{tr}(\boldsymbol{K}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^T)\le \boldsymbol{\xi}_{ij}, \quad (i,j)\in S\\
&\qquad \mathrm{tr}(\boldsymbol{K}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^T)\ge \boldsymbol{\xi}_{ij}, \quad (i,j)\in D
\end{aligned} \tag{2}$$

2.2. Multiple kernel learning

The Multiple Kernel Learning (MKL) refers to the process of learning a kernel machine with multiple kernel functions or kernel matrices. In other word, the existing MKL algorithms use different learning methods for determining the kernel combination function. Suppose we have a set of base kernel functions $\{\kappa_r\}_{r=1}^R$, where R is the number of base kernels. An ensemble kernel function κ is then defined by

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sum_{r=1}^{R}\boldsymbol{\beta}_r\kappa_r(\boldsymbol{x}_i,\boldsymbol{x}_j), \quad \boldsymbol{\beta}_r\ge 0 \tag{3}$$

Consequently, one often-used MKL model from binary class data $\{(\boldsymbol{x}_i,\boldsymbol{y}_i\in\pm 1)\}_{i=1}^N$ is formulated as

$$\begin{aligned}
f(\boldsymbol{x}) &= \sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{y}_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+\boldsymbol{b}\\
&= \sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{y}_i\sum_{r=1}^{R}\boldsymbol{\beta}_r\kappa_r(\boldsymbol{x}_i,\boldsymbol{x})+\boldsymbol{b}
\end{aligned} \tag{4}$$

Optimizing over both the coefficients $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ and $\{\boldsymbol{\beta}_r\}_{r=1}^R$ is one particular form of the MKL problems. Recent research efforts on MKL, e.g. [20,38–43] have shown that learning the combination of multiple kernels not only increases the accuracy but also enhances the interpretability. As far as we know, most of the conventional MKL methods learn the combinations of the kernel function derived from the metric of multiple homogeneous Euclidean spaces or Riemannian manifolds with different dimensionalities.

3. Proposed method

In this section, we first describe an overview of our proposed approach for video face recognition. Then, we study multiple statistics for set modeling, which lie in multiple heterogeneous spaces, i.e., one Euclidean space and two different Riemannian manifolds. Subsequently, we exploit the kernel functions for the three statistics in the Hilbert space embedding and then present the Hybrid Euclidean-and-Riemannian Metric Learning (HERML) to fuse such statistics by learning multiple Mahalanobis matrices respectively transforming the hybrid elements from different spaces to a common Euclidean space. 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 to fuse multiple statistics of image sets for more robust video face recognition. As discussed in the prior sections, simultaneously exploiting multiple statistics can be expected to 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 features in three such heterogeneous spaces spanned by our employed multiple statistics. To this end, we exploit kernel functions for such statistics in their Hilbert space embeddings. Since classical MKL algorithms fail to take a small number of kernels as their direct inputs to learn their combination coefficients, we then present an efficient hybrid metric learning framework to fuse the hybrid Euclidean-and-Riemannian features alternatively, which aims to learn multiple distance metrics in the corresponding Hilbert spaces. A conceptual illustration of our approach is shown in Fig. 2.

3.2. *Multiple statistics of image set*

This part will describe the detailed modeling of our employed multiple statistics, which are in one Euclidean space and two different dimensional Riemannian manifolds respectively. Then, we exploit Euclidean and Riemannian kernel functions to embed their original spaces into high dimensional Hilbert space, which facilitates the subsequent hybrid metric learning.

3.2.1. *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 view of probability theory and statistics, we model each set as the following three statistics with different properties.

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

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

As is well known, the mean is a form of vector lying in Euclidean space $\mathbb{R}^d$, where d is the dimension of the samples.

3.2.1.2. Sample covariance matrix. With no assumption about the data distribution, the sample covariance matrix models the variation modes of the set data by computing the variation between the involved samples and the population mean. Given one set with n samples, the covariance matrix is calculated as

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

As studied in [27,29], the covariance matrix resides on Riemannian manifold Sym_+^d.

3.2.1.3. Gaussian model. In probability theory, the Gaussian distribution is a very commonly occurring probability distribution, which simultaneously captures the mean and the variations of one set. Specifically, we exploit the well-known Gaussian Mixture Model (GMM) to represent the probability of one image set by employing the classical Expectation-Maximization (EM) algorithm to estimate the GMMs and Minimum Description Length (MDL) [44] criterion to calculate the number of component Gaussians that best fit the data. The estimated GMM on each image set can be written as

$$G = \sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{x} | \tilde{\boldsymbol{m}}_i, \tilde{\boldsymbol{C}}_i) \tag{7}$$

where $\boldsymbol{x}$ is the feature vector of data samples, $\tilde{\boldsymbol{m}}_i, \tilde{\boldsymbol{C}}_i$ are the estimated first-order and second-order statistics, $\mathcal{N}(\boldsymbol{x} | \tilde{\boldsymbol{m}}_i, \tilde{\boldsymbol{C}}_i)$ denotes a k-dimensional Gaussian component with prior probability w_i, mean vector $\tilde{\boldsymbol{m}}_i$ and covariance matrix $\tilde{\boldsymbol{C}}_i$. Note that, when $M=1$, GMM is essentially a Single Gaussian Model (SGM), which is employed in the conference version of this work.

Based on the information geometry [30,31] theory, we can embed the space of Gaussian components in GMMs into a Riemannian manifold Sym_+^{d+1}. In the field of information geometry, if the random vector $\boldsymbol{x}$ 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 the observation covariance matrix $\tilde{\boldsymbol{C}}$ has a decomposition $\tilde{\boldsymbol{C}} = \boldsymbol{QQ}^T, |\boldsymbol{Q}| > 0$ (here, $|\boldsymbol{Q}|$ means the determinant of $\boldsymbol{Q}$), and vice versa. Therefore, such a Gaussian model $\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}})$ can be characterized by the affine transformation $(\tilde{\boldsymbol{m}}, \boldsymbol{Q})$. According to the information geometry theory in [31], a d-dimensional Gaussian component $\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}})$ can be embedded into Sym_+^{d+1} and thus is uniquely represented by a $(d+1) \times (d+1)$-dimensional SPD matrix $\boldsymbol{P}$ as

$$\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}}) \sim \boldsymbol{P} = |\boldsymbol{Q}|^{-2/(d+1)} \begin{bmatrix} \boldsymbol{QQ}^T + \tilde{\boldsymbol{m}}\tilde{\boldsymbol{m}}^T & \tilde{\boldsymbol{m}} \\ \tilde{\boldsymbol{m}}^T & 1 \end{bmatrix} \tag{8}$$

For detailed theory on the embedding process refer to [31].

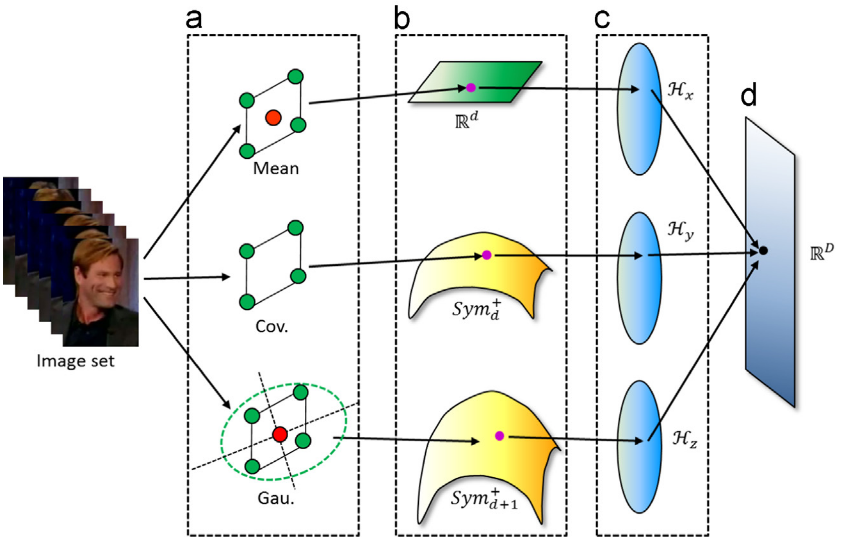

Fig. 2. Conceptual illustration of the proposed Hybrid Euclidean-and-Riemannian Metric Learning (HERML) framework for video face recognition. (a) We first model each image set by its sample mean, covariance matrix and Gaussian distribution, which lie in (b) one Euclidean space $\mathbb{R}^d$ and two Riemannian manifolds Sym_+^d, Sym_+^{d+1} respectively. Finally, by further embedding such heterogeneous spaces into Hilbert spaces (c), the hybrid Euclidean/Riemannian elements are unified in a common subspace (d) by learning multiple Mahalanobis matrices, which can be reduced to transformations from the Hilbert spaces.

3.2.2. *Hilbert space embedding for multiple statistics*

3.2.2.1. For mean vectors. As a positive definite kernel, the linear kernel has proven very effective in Euclidean space for kernel based algorithms. It maps the data points to a high dimensional Hilbert space for yielding a very rich representation. Specifically, for the points in the Euclidean space, the Gaussian kernel can be expressed as

$$\kappa_m(\boldsymbol{m}_i, \boldsymbol{m}_j) = \boldsymbol{m}_i^T \boldsymbol{m}_j \tag{9}$$

which makes use of the Euclidean distance between two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. In addition to the linear kernel, there are many other well-studied kernels such as Gaussian Radial Basis Function (RBF) kernel and polynomial kernel. Without loss of generality, in this paper, we only exploit the linear kernel for the Hilbert embedding of mean vectors in our proposed metric learning framework.

3.2.2.2. For covariance matrices. Each nonsingular covariance matrix $\boldsymbol{C}$ is actually a SPD matrix. The three most widely used distance measures of SPD matrices are the Affine-Invariant Distance (AID) [27,45,46], the Log-Euclidean Distance (LED) [19,29,47] and the Stain Divergence based Distance (SDD) [48–50]. As studied in [51], only AID and LED yield true geodesic distances and only LED and SDD can induce positive definite kernel. Therefore, in this work, we focus on the LED, which is not only a true geodesic distance on Sym_+ but also yields a positive definite kernel as studied in [19,51].

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

$$d(\boldsymbol{C}_i, \boldsymbol{C}_j) = \| \log(\boldsymbol{C}_i) - \log(\boldsymbol{C}_j) \|_F \tag{10}$$

where $\| \cdot \|_F$ denotes the matrix Frobenius form.

As studied in [19], a Riemannian kernel function on the Sym_+ manifold can be derived by computing the corresponding inner product in the space. So, we embed the space of covariance matrices into Hilbert space by using this kind of Riemannian kernel:

$$\kappa_C(\boldsymbol{C}_i, \boldsymbol{C}_j) = \mathrm{tr}(\log(\boldsymbol{C}_i) \cdot \log(\boldsymbol{C}_j)) \tag{11}$$

3.2.2.3. For Gaussian distributions. To embed the Gaussian distribution into Hilbert space, Campbell et al. [52] reported a GMM-supervector kernel based KL divergence, which is perhaps the most widely used dissimilarity measure between two probability distributions p_a, p_b:

$$\Psi_{KL}(p_i \| p_j) = \int_{R^n} p_i(x) \log\left(\frac{p_i(x)}{p_j(x)}\right) dx \tag{12}$$

However, since the KL divergence dose not satisfy Mercer's theorem [36], an approximation is considered, for the case of GMM, by bounding the divergence with log-sum inequality:

$$\begin{aligned} \Psi_{KL}(p_i \| p_j) &\le \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \Psi_{KL}(p_i^a \| p_j^b) \\ &= \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \Psi_{KL}(\omega_a f(\tilde{\boldsymbol{m}}_i^a, \tilde{\boldsymbol{C}}_i^a) \| \omega_b f(\tilde{\boldsymbol{m}}_j^b, \tilde{\boldsymbol{C}}_j^b)) \\ &= \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \omega_a \omega_b \Psi_{KL}(f(\tilde{\boldsymbol{m}}_i^a, \tilde{\boldsymbol{C}}_i^a) \| f(\tilde{\boldsymbol{m}}_j^b, \tilde{\boldsymbol{C}}_j^b)) \end{aligned} \tag{13}$$

where M_a, M_b are the numbers of Gaussian components in the two GMMs, p_i^a, p_j^b are the probability distributions of the corresponding Gaussian components, w_a, w_b are their prior probabilities, $\Psi_{KL}(\cdot)$ is calculated by the canonical KL divergence. In this paper, since we embed each Gaussian component into a Riemannian manifold Sym_+^{d+1} (i.e., each Gaussian component can be reformulated as a $d+1$ dimensional SPD matrix $\boldsymbol{P}$ as computed in Eq. (8)), we alternately exploit the Riemannian distance of pairs of Gaussian components to calculate $\Psi_{KL}(\cdot)$. So, by employing the LED metric of SPD matrices, we formulate the corresponding kernel function in the Hilbert Space Embedding of Gaussian Distributions as

$$\kappa_G(\boldsymbol{G}_i, \boldsymbol{G}_j) = \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \omega_a \omega_b \, \mathrm{tr}(\log(\boldsymbol{P}_i^a) \cdot \log(\boldsymbol{P}_j^a)) \tag{14}$$

where $\boldsymbol{P}_i^a, \boldsymbol{P}_j^a$ are the corresponding $d+1$ dimensional SPD matrices for the two involved Gaussian components.

3.3. *Hybrid Euclidean-and-Riemannian metric learning*

In this part, we will present the formulation of our proposed Hybrid Euclidean-and-Riemannian Metric Learning method in details. Then, we introduce the optimization of the proposed method.

3.3.1. *Formulation*

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, i.e., $\phi : \mathbb{R}^d \to \mathcal{F}$ (or $\mathrm{Sym}_+ \to \mathcal{F}$), 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 $\boldsymbol{\Phi}_i^r$ 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 $\boldsymbol{\Phi}_i^r, \boldsymbol{\Phi}_j^r$, we define the distance metric as

$$d_{A_r}(\boldsymbol{\Phi}_i^r, \boldsymbol{\Phi}_j^r) = \mathrm{tr}(\boldsymbol{A}_r(\boldsymbol{\Phi}_i^r - \boldsymbol{\Phi}_j^r)(\boldsymbol{\Phi}_i^r - \boldsymbol{\Phi}_j^r)^T) \tag{15}$$

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

As shown in Fig. 2, assuming the high dimensional features of multiple statistics can be mapped to a common space, we can jointly optimize multiple Mahalanobis matrices $\boldsymbol{A}_r$ $(r = 1, \ldots, R)$, which can be reduced to multiple transformations respectively mapping the multiple statistics from the corresponding Hilbert spaces to the common space. To learn these Mahalanobis matrices, 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 [37], 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

$$\begin{aligned} &\min_{\boldsymbol{A_1} \succcurlyeq 0, \ldots, \boldsymbol{A_R} \succcurlyeq 0, \boldsymbol{\xi}} \quad \frac{1}{R} \sum_{r=1}^{R} D_{\ell d}(\boldsymbol{A}_r, \boldsymbol{A}_0) + \gamma D_{\ell d}(\mathrm{diag}(\boldsymbol{\xi}), \mathrm{diag}(\boldsymbol{\xi}_0)), \\ &\text{s.t.} \quad \frac{\boldsymbol{\delta}_{ij}}{R} \sum_{r=1}^{R} d_{A_r}(\boldsymbol{\Phi}_i^r, \boldsymbol{\Phi}_j^r) \le \boldsymbol{\xi}_{ij}, \quad \forall (i,j) \end{aligned} \tag{16}$$

where $d_{A_r}(\boldsymbol{\Phi}_i^r, \boldsymbol{\Phi}_j^r)$ is obtained in Eq. (15) and $\boldsymbol{\xi}$ is initialized as $\boldsymbol{\xi}_0$, which is a vector with each element equal to $\boldsymbol{\delta}_{ij}\rho - \zeta\tau$, ρ is the threshold for distance comparison, τ is the margin, and 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. (15) directly in the Hilbert space. Hence, we use the kernel trick method [53] by expressing the basis $\boldsymbol{w}_k^r$ as a linear combination of all the training samples in the mapped space as

$$\boldsymbol{w}_k^r = \sum_{j=1}^{N} \boldsymbol{u}_j^k \boldsymbol{\Phi}_j^r \tag{17}$$

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

$$\sum_{r=1}^{R} (\boldsymbol{w}_k^r)^T \boldsymbol{\Phi}_i^r = \sum_{r=1}^{R} \sum_{j=1}^{N} \boldsymbol{u}_j^k (\boldsymbol{\Phi}_j^r)^T \boldsymbol{\Phi}_i^r = \sum_{r=1}^{R} (\boldsymbol{u}^k)^T \boldsymbol{K}_{\cdot i}^r \tag{18}$$

where $\boldsymbol{u}^k$ is an $N \times 1$ column vector and its j-th entry is $\boldsymbol{u}_j^k$, and $\boldsymbol{K}_{\cdot i}^r$ 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 and Riemannian kernel functions respectively in Eqs. (9), (11), and (14) 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. (16) can be rewritten as

$$\min_{\boldsymbol{B_1} \succcurlyeq 0, \ldots, \boldsymbol{B_R} \succcurlyeq 0, \boldsymbol{\xi}} \quad \frac{1}{R} \sum_{r=1}^{R} D_{\ell d}(\boldsymbol{B}_r, \boldsymbol{B}_0) + \gamma D_{\ell d}(\mathrm{diag}(\boldsymbol{\xi}), \mathrm{diag}(\boldsymbol{\xi}_0)),$$

$$\text{s.t.} \quad \frac{\boldsymbol{\delta}_{ij}}{R} \sum_{r=1}^{R} d_{B_r}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r) \le \boldsymbol{\xi}_{ij}, \quad \forall (i,j) \tag{19}$$

where $d_{B_r}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r)$ 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:

$$d_{B_r}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r) = \mathrm{tr}(\boldsymbol{B}_r (\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)(\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)^T) \tag{20}$$

The objective function of our proposed HERML framework in Eq. (19) simultaneously learns three Mahalanobis matrices $\boldsymbol{B}_r^*$ ($r = 1, \ldots, 3$), while the kernel version of ITML learns a single kernel matrix $\boldsymbol{K}^*$ (see Eq. (2)). Compared with traditional MKL methods that typically learn the combination coefficients for fusing a large number of basic homogeneous kernels, our proposed HERML algorithm alternately learns multiple Mahalanobis matrices for combining a small number of hybrid Euclidean/Riemannian kernels. For more discussions about the differences of the proposed HERML from ITML and MKL refer to Section 3.4.

3.3.2. *Optimization*

To solve the problem in Eq. (19), we adopt the cyclic Bregman projection method [54–56], 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. Refer to [54–56] for more details. The updating rules for our proposed method are shown in the following proposition:

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

$$\begin{cases} \boldsymbol{B}_r^{t+1} = \boldsymbol{B}_r^t + \beta_r \boldsymbol{B}_r (\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)(\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)^T \boldsymbol{B}_r, & \text{(a)} \\ \boldsymbol{\xi}_{ij}^{t+1} = \dfrac{\gamma \boldsymbol{\xi}_{ij}^t}{\gamma + \boldsymbol{\delta}_{ij} \alpha \boldsymbol{\xi}_{ij}^t}, & \text{(b)} \end{cases} \tag{21}$$

where $\beta_r = \boldsymbol{\delta}_{ij} \alpha / (1 - \boldsymbol{\delta}_{ij} \alpha d_{B_r^t}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r))$ *and* α *can be solved by*

$$\frac{\boldsymbol{\delta}_{ij}}{R} \sum_{r=1}^{R} \frac{d_{B_r^t}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r)}{1 - \boldsymbol{\delta}_{ij} \alpha d_{B_r^t}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r)} - \frac{\gamma \boldsymbol{\xi}_{ij}^t}{\gamma + \boldsymbol{\delta}_{ij} \alpha \boldsymbol{\xi}_{ij}^t} = 0. \tag{22}$$

Proof. Based on the cyclic projection method [54–56], we formulate the Lagrangian form of Eq. (19) and set the gradients to zero w.r.t $\boldsymbol{B}_r^{t+1}$, $\boldsymbol{\xi}_{ij}^{t+1}$ and α to get the following update equations:

$$\begin{cases} \nabla D(\boldsymbol{B}_r^{t+1}) = \nabla D(\boldsymbol{B}_r^t) + \boldsymbol{\delta_{ij}} \alpha (\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)(\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)^T, & \text{(a)} \\ \nabla D(\boldsymbol{\xi}_{ij}^{t+1}) = \nabla D(\boldsymbol{\xi}_{ij}^t) - \dfrac{\boldsymbol{\delta}_{ij} \alpha}{\gamma}, & \text{(b)} \\ \dfrac{\boldsymbol{\delta}_{ij}}{R} \displaystyle\sum_{r=1}^{R} \mathrm{tr}(\boldsymbol{B}_r^{t+1} (\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)(\boldsymbol{K}_{\cdot i}^r - \boldsymbol{K}_{\cdot j}^r)^T) = \boldsymbol{\xi}_{ij}^{t+1}. & \text{(c)} \end{cases} \tag{23}$$

Then, we can derive Eqs. (21a) and (21b) from Eqs. (23a) and (23b), respectively. Substituting Eqs. (21a) and (21b) into Eq. (23c), we obtain Eq. (22) related to α. For an inequality constraint in Eq. (19), we use $\boldsymbol{\eta}_{ij} \ge 0$ as the dual variable of α. To maintain nonnegativity of the dual variable (which is necessary for satisfying the KKT conditions), following the work [56], we solve Eq. (19) by updating α as

$$\alpha \leftarrow \min(\alpha, \boldsymbol{\eta}_{ij}), \quad \boldsymbol{\eta}_{ij} \leftarrow \boldsymbol{\eta}_{ij} - \alpha \tag{24}$$

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. The main time cost is to update $\boldsymbol{B}_r^{t+1}$ 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 following algorithm.

Algorithm 1. Hybrid Euclidean-and-Riemannian Metric Learning.

```
Input: Training pairs $\{(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r), \boldsymbol{\delta}_{ij}\}$, and slack parameter $\gamma$, input
    Mahalanobis matrix $\boldsymbol{B}_0$, distance thresholds $\rho$, margin
    parameter $\tau$ and tuning scale $\zeta$
1. $t \leftarrow 1$, $\boldsymbol{B}_r^1 \leftarrow \boldsymbol{B}_0$ for $r = 1, \ldots, R$, $\boldsymbol{\eta}_{ij} \leftarrow 0$, $\boldsymbol{\xi}_{ij} \leftarrow \boldsymbol{\delta_{ij}} \rho - \zeta \tau$, $\forall (i,j)$
2. Repeat
3. Pick a constraint $(i,j)$ and compute the distances $d_{B_r^t}(\boldsymbol{K}_{\cdot i}^r, \boldsymbol{K}_{\cdot j}^r)$
    for $r = 1, \ldots, R$ by using Eq. (20).
4. Solve $\alpha$ in Eq. (22) and update $\alpha$ by using Eq. (24).
5. Update $\boldsymbol{B}_r^{t+1}$ by using Eq. (21a) for $r = 1, \ldots, R$.
6. Update $\boldsymbol{\xi}_{ij}^{t+1}$ by using Eq. (21b).
7. Until convergence
Output: Mahalanobis matrices $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R$.
```

3.4. *Discussion about related work*

The idea of the proposed hybrid metric learning framework is inspired by our previous work in [32] which aims to learn the cross-space (i.e., Euclidean space to Riemannian manifold) distance metric for matching Euclidean points against Riemannian elements. Different from [32], this presented work attempts to simultaneously learn multiple distance metrics (each working in a single space) respectively for Euclidean and Riemannian elements under a hybrid metric fusion framework.

To solve our problem of hybrid metric learning, as ITML [37], we employ the LogDet divergence based constraint which has excellent ability to implicitly maintain the positive semidefiniteness (i.e., fix the rank) of the learned Mahalanobis matrices for fusing the hybrid kernels in our proposed framework.

Although using the same constraint of LogDet divergence as ITML [37], our proposed metric learning framework works in a different way from ITML. ITML aims to learn a *single* Mahalanobis matrix in a Euclidean space or to learn a kernel matrix in a Hilbert space. In contrast, our method attempts to jointly learn *multiple* Mahalanobis matrices in multiple Hilbert spaces for fusing hybrid Euclidean and Riemannian data. In some sense, our method tends to be 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 naturally reduces to the original ITML. In other words, ITML can be viewed as a special case of the proposed HERML algorithm in this paper.

In addition, there exist several works [20,38–43] for Multiple Kernel Learning (MKL) in the literature. While these works also exploit multiple kernels, they mainly focus on fusing multiple *homogeneous* Euclidean (or Riemannian) features. In contrast, our method addresses the new problem of fusing *heterogeneous* Euclidean and Riemannian features. In the sense of techniques, most MKL methods aim to learn the combination coefficients to fuse a lot of basic kernels, while our hybrid metric learning algorithm HERML alternatively learns multiple Mahalanobis matrices to fuse a small number of kernels (only 3 in our work). Therefore, both the problem and technique of our method differ from the existing MKL methods.

4. Experiments

In this section, we evaluate our proposed approach on four large-scale video face datasets for both video face identification and video face verification tasks. The following describes the experimental results and our analysis in details.

4.1. Comparative methods and settings

We compare our proposed approach with three categories of the state-of-the-art set-based methods as follows. 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 of image set here. Since ITML also has a kernel version, we feed our proposed kernel function of Gaussian distribution to it for additional comparison.

(1) *Sample-based method*: Maximum Mean Discrepancy (MaxMD) [12], Affine (Convex) Hull based Image Set Distance (AHISD, CHISD) [10], Set-to-Set Distance Metric Learning (SSDML) [16] and Information Theoretic Metric Learning (ITML) [37].
(2) *Subspace-based method*: Mutual Subspace Method (MSM) [7], Discriminant Canonical Correlations (DCC) [8], Manifold Discriminant Analysis (MDA) [18], Grassmann Discriminant Analysis (GDA) [17], Covariance Discriminative Learning (CDL) [19] and Localized Multi-Kernel Metric Learning (LMKML) [20].
(3) *Distribution-based method*: Single Gaussian Models (SGM) [21], Gaussian Mixture Models (GMM) [22] and kernelized ITML [37] with our DIS-based set model (DIS-ITML).

Except for 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 model(s). For fair comparison, the important parameters of each method are empirically tuned according to the recommendations in the original references: For MaxMD, we use the edition of Bootstrap and set the parameters $\alpha=0.1$, $\sigma=-1$, the number of iteration to 5. For ITML, we use the default parameters as the standard implementation. For AHISD, CHISD and DCC, PCA is performed by preserving 95% energy to learn the linear subspace and corresponding 10 maximum canonical correlations are used. For MDA, the parameters are configured according to [18]. 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, such setting prevents KPLS from working in many cases. So, we use KDA for discriminative learning and adopt the same setting as [19]. 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 use 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 $[0.1, 1]$. In HERML, we first calculate three distances for each pair of video samples under the learned Mahalanobis matrices $\boldsymbol{B}_r$ for the r-th statistics (see Eq. (20)) and then average the three distances as the final distance for the involved sample pair. For identification task, we identify the query videos using the NN classifier based on the above distance calculation. For verification task, with a given threshold, the calculated final distance is used for verifying the associating video pair as the same or different identity.

4.2. Evaluation on video face identification

4.2.1. Datasets

For video face recognition task, we use two public large-scale video face datasets: YouTube Celebrities (YTC) [5] and COX [57]. The YTC is a quite challenging and widely used video face dataset. It has 1910 video clips of 47 subjects collected from YouTube. Most clips contain hundreds of frames, which are often of low resolution and highly compressed with noise and low quality. The COX is a large-scale video dataset involving 1000 different subjects, each of which has 3 videos captured by different camcorders. In each video, there are around 25–175 frames of low resolution and low quality, with blur, and captured under poor lighting. As shown in Figs. 3 and 4, there are some examples on YTC and COX datasets.

In our experiments, For COX, we coarsely align all the faces and normalize them to the same size based on the face bounding-box and the eye positions provided by the original dataset providers. For YTC, following the works [19,20,16], we used directly the detected face regions without further alignment due to the low resolution of the faces. More specifically, each face in YTC is resized to a 20 × 20 image as [19,20] while the faces in COX are resized to 32 × 40. For all faces in the two datasets, histogram equalization is implemented to eliminate lighting effects. On the two video face datasets, we follow the same protocol as the prior work [10,19,20], which conducted ten-fold cross validation experiments, i.e., 10 randomly selected gallery/probe combinations. Finally, the average recognition rates of different methods are reported. For YTC, in each fold, one person has 3 randomly chosen image sets for the gallery and 6 for probes. Different from YTC, COX dataset does also contain an additional independent training set [57], where each subject has 3 videos. Since there are 3 independent testing sets of videos in COX, each person has 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.

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

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

Fig. 3. Examples on YTC dataset.

Fig. 4. Examples on COX dataset.

Table 1
Average recognition rate (%) of different set-based methods on YouTube Celebrities (YTC) and COX face 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. In each column of this table, the bold values indicate the first three highest performances on the corresponding database.

Method	YTC	COX-12	COX-13	COX-23	COX-21	COX-31	COX-32
MaxMD [12]	52.6	36.4	19.6	8.9	27.6	19.1	9.6
AHISD [10]	63.7	53.0	36.1	17.5	43.5	35.0	18.8
CHISD [10]	66.3	56.9	30.1	15.0	44.4	26.4	13.7
SSDML [16]	68.8	60.1	53.1	28.7	47.9	44.4	27.3
ITML [37]	65.3	50.9	46.0	35.6	39.6	37.1	34.8
MSM [7]	61.1	45.5	21.5	11.0	39.8	19.4	9.5
DCC [8]	64.8	62.5	66.1	50.6	56.1	64.8	45.2
MDA [18]	65.3	65.8	63.0	36.2	55.5	43.2	30.0
GDA [17]	65.9	68.6	77.7	71.6	66.0	76.1	74.8
CDL [19]	69.7	**78.4**	**85.3**	**79.7**	**75.6**	**85.8**	**81.9**
LMKML [20]	**70.3**	66.0	71.0	56.0	74.0	68.0	60.0
SGM [21]	52.0	26.7	14.3	12.4	26.0	19.0	10.3
GMM [22]	61.0	30.1	24.6	13.0	28.9	31.7	18.9
DIS-ITML [37]	68.4	47.9	48.9	36.1	43.1	35.6	33.6
HERML-SGM	**74.6**	**94.9**	**96.9**	**94.0**	**92.0**	**96.4**	**95.3**
HERML-GMM	**73.3**	**95.1**	**96.3**	**94.2**	**92.3**	**95.4**	**94.5**

4.2.2. *Results and analysis*

We present the rank-1 recognition results of comparative methods on the two datasets in Table 1. Each reported rate is an average recognition rate over the ten-fold trials. Note that, since the multiple kernel learning method LMKML [20] is too time-consuming to run in the setting of COX dataset, which is a large-scale dataset, we alternately use 100 of 300 subject's videos for training and 100 of 700 remaining subject's videos 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 MaxMD, AHISD, CHISD, MSM, SGM, and GMM. On YTC and COX, the sample-based methods (MaxMD, AHISD, and CHISD), the distribution-based methods (SGM and GMM) and the subspace-based method (MSM) are comparable in the term of recognition rate.

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. Of 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 the Riemannian space and match them in the same space, which is more favorable for the set classification task [17].

Thirdly, we compare the state-of-the-art methods with our approach and find that they are impressively outperformed by ours on the two datasets. Several reasons are figured out as following: In terms of set modeling, as stated in Section 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 especially the non-Euclidean data structure of covariance matrices and Gaussian models, our method jointly learns hybrid metrics to fuse them for more discriminant classification. In comparison, the MKL method LMKML [20] simply transforms both the covariance matrices and the third-order tensors to vectors, which lie in Euclidean spaces. As a result, it neglects the non-Euclidean geometrical structure of the covariance matrices and third-order tensors. We argue that the underlying data structure and distributions in the learning stage will probably lead to undesirable metrics. 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 show that our novel hybrid metric learning method has an impressive superiority over the original ITML.

Fourthly, we also compare the discriminative power of different basic set modelings (i.e., sample mean, sample covariance matrix and Gaussian model) for video face recognition. For each statistic, we performed our proposed method to train and classify sets with NN classifier. Table 2 tabulates the classification rates of multiple statistics. We can observe that the SGM/GMM 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 mean and covariance matrix sometimes are better than those of SGM/GMM on COX-S2V. This is because the dataset may contain some sets not strictly following Gaussian distribution. Since the multiple statistics complement each other, the performance can be improved by our proposed metric learning with all statistic models.

As shown in Table 1, compared with SGM method, the higher performances of GMM method demonstrate that it is more qualified to faithfully characterize the set structure. However, from Table 2, we observe that, in our proposed framework, the GMM statistics is outperformed by the SGM statistics in several cases. This is possibly because, to allow sufficient flexibility and avoid overfitting, GMM usually needs to tune an appropriate number of components. Though we have done this in a principled way according to the MDL [44] criterion, it is still difficult to find the best parameter to fit real data with arbitrary distribution. However, in practical applications, we still prefer

Table 2
Average recognition rates (%) of separatively using mean, covariance matrix (Cov.) and Gaussian model (SGM/GMM), combining mean and covariance matrix (Mean+Cov.), fusing all of them (ALL) with our metric learning method on YouTube Celebrities (YTC) and COX face datasets. Here, COX-*ij* indicates the test using the *i*-th set of videos as gallery and the *j*-th set of videos as probe. Note that, ALL-SGM and ALL-GMM mean fusing the first two statistics with SGM or GMM. In each column of this table, the bold values indicate the first three highest performances on the corresponding database.

Statistics	YTC	COX-12	COX-13	COX-23	COX-21	COX-31	COX-32
Mean	64.1	86.2	92.0	82.8	83.2	86.9	84.9
Cov.	70.2	88.8	93.6	90.3	86.4	94.0	93.1
SGM	**73.5**	92.8	94.7	92.2	89.0	94.7	94.4
GMM	**72.3**	93.1	94.3	92.5	90.0	93.3	93.7
Mean+Cov.	71.6	**93.1**	**95.2**	**93.1**	**91.2**	**95.2**	**95.0**
ALL-SGM	**74.6**	**94.9**	**96.9**	**94.0**	**92.0**	**96.4**	**95.3**
ALL-GMM	**73.3**	**95.1**	**96.3**	**94.2**	**92.3**	**95.4**	**94.5**

Table 3
Computation time (seconds) of different methods on the YTC dataset for training and testing (classification of one video).

Method	MaxMD	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

to the more general GMM modeling due to its stronger ability to capture the data variations especially when they are with a multi-model density.

Lastly, on the YouTube dataset, we compare the computational complexities of different methods on an Intel(R) Core(TM) i7-3770 (3.40 GHz) PC. Table 3 lists the time cost for each method. The presentation of training time is only required by discriminant methods. 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 LMKML need to learn on very high dimensional Euclidean vectors transformed from the covariance matrices and the third-order tensors. In contrast, our HERML method work directly on covariance matrices and the SPD model of Gaussian statistics, whose sizes are much smaller than their vector forms.

4.3. Evaluation on video face verification

4.3.1. Datasets

For video face verification task, we conduct experiments on two challenging large-scale datasets: YouTube Face (YTF) [34] and Point-and-Shoot Face Recognition Challenge (PaSC) [35]. The YTF [34] contains 3425 videos of 1595 different persons collected from the YouTube website. There are large variations in pose, illumination, and expression in each video, and the average length of each video clip is 181.3 frames. The PaSC [35] includes 2802 videos of 265 people carrying out simple actions. Every action was filmed by two cameras: a high quality, 1920 × 1080 pixel, Panasonic camera on a tripod and one of five alternative handheld video cameras. The tripod-based Panasonic data serves as a control. The handheld cameras have resolutions ranging from 640 × 480 up to 1280 × 720. As shown in Figs. 5 and 6, there are some examples on YTF and PaSC datasets.

On YTF, we follow the standard evaluation protocol [34] and test our method for unconstrained face verification with 5000 video pairs. These pairs are equally divided into 10 folds, and each fold has 250 intra-personal pairs and 250 inter-personal pairs. The experiment is performed in the restricted training setting. On PaSC, there are two video face verification experiments: control-to-control and handheld-to-handheld experiments. In both of the two experiments, the target and query sigsets contain the same set of videos. The task was to verify a claimed identity in the query video by comparing with the associated target video. Since the same 1401 videos served as both the target and query sets, 'same video' comparisons were excluded.

For PaSC and YTF, we directly used face detection and positions of eyes to rotate and crop each face image to a normal size. Specifically, in our experiments, we directly crop the face images according to the provided data and then resize them into 24 × 40 pixels for YTF as [42] while 64 × 80 pixels for PaSC. On YTF dataset, we extract the raw intensity feature of resized video frames. Compared with the YTF datasets, the PaSC dataset is so challenging that using pixels intensity as feature performs too badly. Therefore, we employed the state-of-the-art DCNN to learn features on this dataset. Specifically, we employ the Caffe [58] to extract the Deep Convolutional Neural Network (DCNN) feature of the video frames. The DCNN model is pretrained on CFW [59], and then fine-tuned on the data from the training sets of PaSC and COX datasets.

4.3.2. Results and analysis

In the video face verification evaluation, we compare six representative set-based methods, i.e., AHISD [10], CHISD [10], SSDML [16], DCC [8], GDA [17] and CDL [19], due to their high performances in the video face identification task. As LMKML [20] are very time-consuming on large-scale video, we do not evaluate it in this task. On YTF dataset, since DCC [8], GDA [17] and CDL [19] were not specifically designed for face verification, we modify them by constructing the within-class scatter matrix from intra-class pairs and the between-class scatter matrix from inter-class pairs.

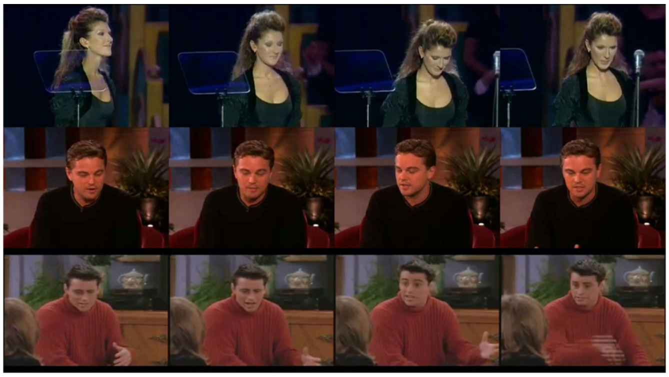

Fig. 5. Examples on YTF dataset.

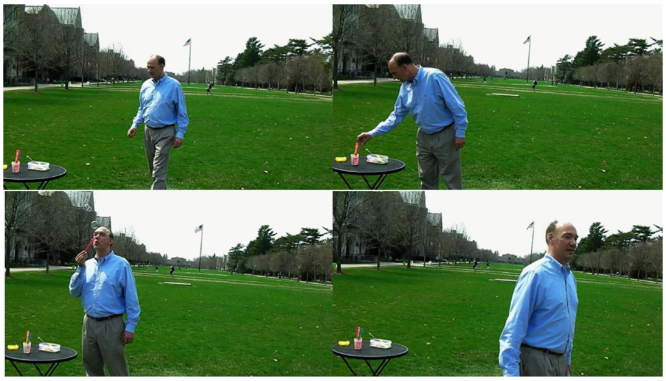

Fig. 6. Examples on PaSC dataset.

In Table 4, we report mean accuracies plus standard deviations of comparative methods on YTF dataset while listing their verification rates at false accept rate (FAR) = 0.01 on PaSC dataset. Figs. 7–9 show the corresponding ROC of these methods on YTF and PaSC. As can be seen in these results, our method coupled with either SGM or GMM outperforms the other set-based methods. Most performances of our implemented set-based methods on YTF are lower than those reported in [60], possibly because it used higher resolution of facial images and more robust feature were used in their methods. We also see that the result of GDA on YTF is much lower than other set-based methods. This may be possible for that the published results of its corresponding basic set-to-set distance (i.e., projection metric) are also much lower than others in the original work [34], where the YTF is released. On PaSC, we extract the state-of-the-art DCNN feature and find that most of the comparative set-based methods significantly outperform the published state-of-the-art method Eigen-PEP [61], whose performance is 26% on the handheld experiment. As can be seen, our method yields an impressive improvement of 20% above the state-of-the-art result.

In the end, we also compare performances of separatively/simultaneously using sample mean, sample covariance matrix and Gaussian model in the video face identification task. Table 5 lists the verification rates of our employed multiple statistics coupling with our metric learning framework. On both datasets, the SGM/GMM outperform other two statistics due to its jointly modeling of mean and the covariance matrix in a Gaussian distribution. As the multiple statistics complement each other, the performance can be improved by fusing all statistic models in our proposed metric learning framework.

Table 4
Verification rates (%) of different set-based methods on YouTube Face DB (YTF) and PaSC face datasets. Note that, the reported rates on YTF are mean accuracies with standard deviations, and those on PaSC are verification rates when false accept rate is 0.01. In each column of this table, the bold values indicate the first three highest performances on the corresponding database.

Method	YTF	PaSC—control	PaSC—handheld
AHISD [10]	64.80 ± 1.54	21.96	14.29
CHISD [10]	66.30 ± 1.21	26.12	20.97
SSDML [16]	65.38 ± 1.86	29.19	22.89
DCC [8]	**68.28 ± 2.21**	38.87	37.53
GDA [17]	59.14 ± 1.98	41.88	**43.25**
CDL [19]	64.94 ± 2.38	**42.62**	42.97
HERML-SGM	**75.16 ± 0.84**	**45.40**	**45.46**
HERML-GMM	**74.36 ± 1.53**	**46.61**	**46.23**

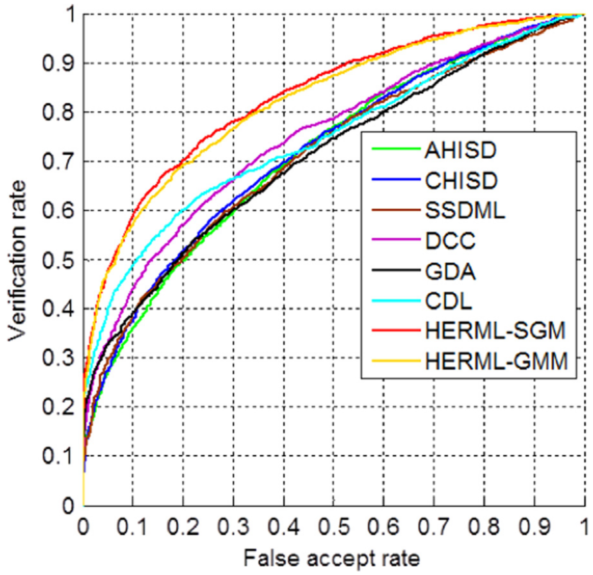

Fig. 7. ROC for the video-to-video face verification experiment on YTF.

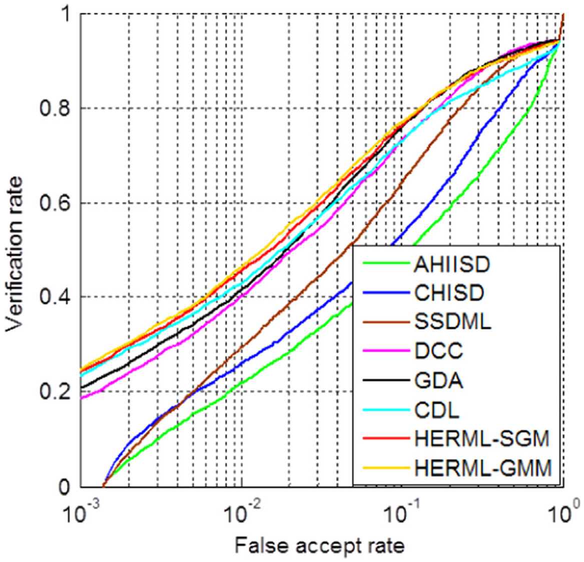

Fig. 8. ROC for the control video-to-video face verification experiment on PaSC.

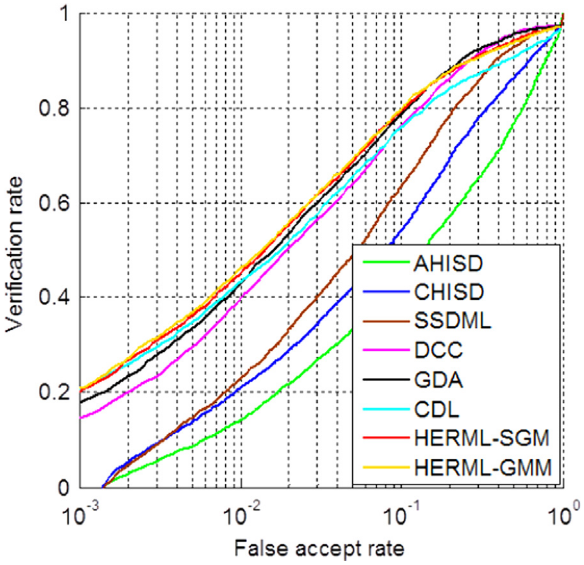

Fig. 9. ROC for the handheld video-to-video face verification experiment on PaSC.

5. Conclusions

In this paper, we proposed a novel set-based hybrid metric learning method to fuse multiple statistics of image sets for more robust large-scale video face recognition in the wild. Our contributions lie in modeling multiple complementary statistics in heterogeneous spaces and learning hybrid Euclidean-and-Riemannian metrics to combine them. 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.

The extensive experiments on four large-scale datasets have shown that our proposed method outperforms the state-of-the-art

Table 5
Average verification rates (%) of separatively using mean, covariance matrix (Cov.) and Gaussian model (SGM/GMM), combining mean and covariance matrix (Mean+Cov.), fusing all of them (ALL) with our metric learning method on YouTube Face (YTF) and PaSC datasets. Here, ALL-SGM and ALL-GMM mean fusing the first two statistics with DIS-SGM or DIS-GMM.

Statistics	YTF	PaSC—control	PaSC—handheld
SAS	72.10 ± 2.25	43.57	41.46
SUS	71.22 ± 2.04	44.88	45.33
DIS-SGM	72.60 ± 1.09	**45.46**	**44.91**
DIS-GMM	69.74 ± 2.21	**45.65**	**45.21**
SAS+SUS	**74.40 ± 1.11**	45.07	44.40
ALL-SGM	**75.16 ± 0.84**	**45.39**	**45.46**
ALL-GMM	**74.36 ± 1.53**	**46.61**	**46.23**

set-based methods in both tasks of video face identification and verification. The comparison of our employed statistics coupled separately or jointly with our hybrid metric learning framework demonstrates that they are complementary for each other to improve the performance of face recognition when they are fused together in the real-world setting. In terms of the efficiency, compared with the existing multiple kernel learning method, our proposed method is much more efficient to fuse multiple hybrid kernels on the large-scale video data.

In the future, several possible directions of our proposed method can be as follows. Firstly, in addition to the application on video face recognition, our proposed method can also be employed to other applications such as action recognition, person re-identification and so on. Secondly, the significant improvement by the state-of-the-art DCNN image feature indicates that jointly learning image feature and image set feature may be a very promising direction. Lastly, it would be also 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.

Conflict of interest

None declared.

Acknowledgments

This work is partially supported by National Basic Research Program of China (973 Program) under No. 2015CB351802 and by Natural Science Foundation of China under contract Nos. 61222211, 61390511, and 61379083.

References

[1] J.H. Barr, K. Boyer, P. Flynn, S. Biswas, Face recognition from video: a review, Int. J. Pattern Recognit. Artif. Intell. 26 (5) (2012).
[2] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Comput. Vis. Image Underst. 91 (1) (2003) 214–245.
[3] X. Liu, T. Cheng, Video-based face recognition using adaptive hidden Markov models, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2003, pp. 340–345.
[4] S. Zhou, R. Chellappa, B. Moghaddam, Visual tracking and recognition using appearance-adaptive models in particle filters, IEEE Trans. Image Process. 13 (11) (2004) 1491–1506.
[5] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2008, pp. 1–8.
[6] N. Ye, T. Sim, Towards general motion-based face recognition, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2010, pp. 2598–2605.
[7] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: Proceedings of the International Conference on Automatic Face Gesture Recognition, 1998, pp. 318–323.
[8] T. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1005–1018.
[9] R. Wang, S. Shan, X. Chen, Q. Dai, W. Gao, Manifold–manifold distance and its application to face recognition with image sets, IEEE Trans. Image Process. 21 (10) (2012) 4466–4479.
[10] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2010, pp. 2567–2573.
[11] S. Chen, C. Sanderson, M. Harandi, B. Lovell, Improved image set classification via joint sparse approximated nearest subspaces, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2013, pp. 452–459.
[12] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (2012) 723–773.
[13] Y. Hu, A. Mian, R. Owens, Sparse approximated nearest points for image set classification, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2011.
[14] M. Yang, P. Zhu, L. Gool, L. Zhang, Face recognition based on regularized nearest points between image sets, in: Proceedings of the International Conference on Automatic Face Gesture Recognition, 2013.
[15] Z. Huang, X. Zhao, S. Shan, R. Wang, X. Chen, Coupling alignments with recognition for still-to-video face recognition, in: Proceedings of the International Conference on Computer Vision, 2013, pp. 3296–3303.
[16] P. Zhu, L. Zhang, W. Zuo, D. Zhang, From point to set: extend the learning of distance metrics, in: Proceedings of the International Conference on Computer Vision, 2013.
[17] J. Hamm, D.D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: Proceedings of the International Conference on Machine Learning, 2008.
[18] R. Wang, X. Chen, Manifold discriminant analysis, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2009.
[19] R. Wang, H. Guo, L. Davis, Q. Dai, Covariance discriminative learning: a natural and efficient approach to image set classification, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2012.
[20] J. Lu, G. Wang, P. Moulin, Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning, in: Proceedings of the International Conference on Computer Vision, 2013.
[21] G. Shakhnarovich, J.W. Fisher, T. Darrell, Face recognition from long-term observations, in: Proceedings of the European Conference on Computer Vision, 2002.
[22] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2005.
[23] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 312–377.
[24] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[25] M. Tipping, C. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. 61 (3) (1999) 611–622.
[26] M.T. Harandi, C. Sanderson, S. Shirazi, B.C. Lovell, Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2011.
[27] X. Pennec, P. Fillard, N. Ayache, A Riemannian framework for tensor computing, Int. J. Comput. Vis. 66 (1) (2006) 41–66.
[28] O. Tuzel, F. Porikli, P. Meer, Region covariance: a fast descriptor for detection and classification, in: Proceedings of the European Conference on Computer Vision, 2006.
[29] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in a novel vector space structure on symmetric positive-definite matrices, SIAM J. Matrix Anal. Appl. 29 (1) (2007) 328–347.
[30] S.-I. Amari, H. Nagaoka, Methods of Information Geometry, Oxford University Press, New York, 2000.
[31] M. Lovrić, M. Min-Oo, E.A. Ruh, Multivariate normal distributions parametrized as a Riemannian symmetric space, J. Multivar. Anal. 74 (1) (2000) 36–48.
[32] Z. Huang, R. Wang, S. Shan, X. Chen, Learning Euclidean-to-Riemannian metric for point-to-set classification, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2014, pp. 1677–1684.
[33] Z. Huang, R. Wang, S. Shan, X. Chen, Hybrid Euclidean-and-Riemannian metric learning for image set classification, in: Proceedings of the Asian Conference on Computer Vision, 2014.
[34] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: Proceedings of the Computer Vision and Pattern Recognition Conference, IEEE, 2011, pp. 529–534.
[35] J.R. Beveridge, P.J. Phillips, D.S. Bolme, B.A. Draper, G.H. Given, Y.M. Lui, M.N. Teli, H. Zhang, W.T. Scruggs, K.W. Bowyer, et al., The challenge of face recognition from digital point-and-shoot cameras, in: Sixth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2013, IEEE, 2013, pp. 1–8.
[36] W.M. Campbell, D.E. Sturim, D.A. Reynolds, A. Solomonoff, SVM based speaker verification using a GMM supervector kernel and NAP variability

compensation, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2006, pp. 97–100.

[37] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the International Conference on Machine Learning, 2007.

[38] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, Simple MKL, J. Mach. Learn. Res. 9 (11) (2008).

[39] B. McFee, G. Lanckriet, Learning multi-modal similarity, J. Mach. Learn. Res. 12 (2011) 491–523.

[40] P. Xie, E.P. Xing, Multi-modal distance metric learning, in: International Joint Conference on Artificial Intelligence, 2013.

[41] R. Vemulapalli, J.K. Pillai, R. Chellappa, Kernel learning for extrinsic classification of manifold features, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2013.

[42] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2013.

[43] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, M. Harandi, Combining multiple manifold-valued descriptors for improved object recognition, in: International Conference on Digital Image Computing: Techniques and Applications, 2013.

[44] A.R. Barron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling, IEEE Trans. Inf. Theory 44 (6) (1998) 2743–2772.

[45] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on Riemannian manifolds, IEEE Trans. Pattern Anal. Mach. Intell. 30 (10) (2008) 1713–1727.

[46] M. Harandi, M. Salzmann, R. Hartley, From manifold to manifold: geometry-aware dimensionality reduction for SPD matrices, in: Proceedings of the European Conference on Computer Vision, 2014.

[47] H. Minh, M. Biagio, V. Murino, Log-Hilbert–Schmidt metric between positive definite operators on Hilbert spaces, in: Proceedings of the Neural Information Processing Systems, 2014, pp. 144–152.

[48] S. Sra, A new metric on the manifold of kernel matrices with application to matrix geometric means, in: Proceedings of the Neural Information Processing Systems, 2012, pp. 144–152.

[49] M.T. Harandi, C. Sanderson, R. Hartley, B.C. Lovell, Sparse coding and dictionary learning for symmetric positive definite matrices: a kernel approach, in: Proceedings of the European Conference on Computer Vision, Springer, 2012.

[50] M. Harandi, M. Salzmann, F. Porikli, Bregman divergences for infinite dimensional covariance matrices, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2014, pp. 1003–1010.

[51] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, M. Harandi, Kernel methods on the Riemannian manifold of symmetric positive definite matrices, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2013.

[52] W.M. Campbell, D.E. Sturim, D.A. Reynolds, Support vector machines using GMM supervectors for speaker verification, IEEE Signal Process. Lett. 13 (5) (2006) 308–311.

[53] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.

[54] L.M. Bregman, 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 (3) (1967) 200–217.

[55] Y. Censor, S. Zenios, Parallel Optimization: Theory, Algorithms, and Applications, Oxford University Press, New York, 1997.

[56] B. Kulis, M. Sustik, I.S. Dhillon, Low-rank kernel learning with Bregman matrix divergences, J. Mach. Learn. Res. 10 (2009) 341–376.

[57] Z. Huang, S. Shan, H. Zhang, H. Lao, A. Kuerban, X. Chen, Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset, in: Proceedings of Asian Conference on Computer Vision, 2012.

[58] J. Y, An Open Source Convolutional Architecture for Fast Feature Embedding, Available at ⟨http://caffe.berkeleyvision.org⟩.

[59] X. Zhang, L. Zhang, X.-J. Wang, H.-Y. Shum, Finding celebrities in billions of web images, IEEE Trans. Multimed. 14 (4) (2012) 107–995.

[60] J. Hu, J. Lu, Y.-P. Tan, Discriminative deep metric learning for face verification in the wild, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2014.

[61] H. Li, G. Hua, X. Shen, Z. Lin, J. Brandt, Eigen-PEP for video face recognition, in: Proceedings of Asian Conference on Computer Vision, 2014.

Zhiwu Huang received the B.S. degree from Xiamen University. Now he is pursuing the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His research interests include pattern recognition and computer vision, especially face recognition.

Ruiping Wang received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2010. He was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, Beijing, from July 2010 to June 2012. He also spent one year working as a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies (UMIACS), at the University of Maryland, College Park, from November 2010 to October 2011. He has been with the faculty of the Institute of Computing Technology, Chinese Academy of Sciences, since July 2012, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.

Shiguang Shan received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences Beijing, in 2004. He has been with ICT, CAS since 2002 and has been a Professor since 2010. He is also the Vice Director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover image analysis, pattern recognition, and computer vision. He is focusing especially on face recognition related research topics. He received the China's State Scientific and Technological Progress Awards in 2005 for his work on face recognition technologies.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1988, 1991, and 1994 respectively. He was a Professor with the HIT from 1999 to 2005 and was a Visiting Scholar with Carnegie Mellon University from 2001 to 2004. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, since August 2004. His research interests include image processing, pattern recognition, computer vision, and multimodal interface. He has received several awards, including the China's State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work.