Discriminant Analysis on Riemannian Manifold of Gaussian Distributions for Face Recognition With Image Sets

Wen Wang, *Student Member, IEEE*, Ruiping Wang, *Member, IEEE*, Zhiwu Huang, *Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*— To address the problem of face recognition with image sets, we aim to capture the underlying data distribution in each set and thus facilitate more robust classification. To this end, we represent image set as the Gaussian mixture model (GMM) comprising a number of Gaussian components with prior probabilities and seek to discriminate Gaussian components from different classes. Since in the light of information geometry, the Gaussians lie on a specific Riemannian manifold, this paper presents a method named discriminant analysis on Riemannian manifold of Gaussian distributions (DARG). We investigate several distance metrics between Gaussians and accordingly two discriminative learning frameworks are presented to meet the geometric and statistical characteristics of the specific manifold. The first framework derives a series of provably positive definite probabilistic kernels to embed the manifold to a high-dimensional Hilbert space, where conventional discriminant analysis methods developed in Euclidean space can be applied, and a weighted Kernel discriminant analysis is devised which learns discriminative representation of the Gaussian components in GMMs with their prior probabilities as sample weights. Alternatively, the other framework extends the classical graph embedding method to the manifold by utilizing the distance metrics between Gaussians to construct the adjacency graph, and hence the original manifold is embedded to a lower-dimensional and discriminative target manifold with the geometric structure preserved and the interclass separability maximized. The proposed method is evaluated by face identification and verification tasks on four most challenging and largest databases, YouTube Celebrities, COX, YouTube Face DB, and Point-and-Shoot Challenge, to demonstrate its superiority over the state-of-the-art.**

***Index Terms*— Statistical manifold, kernel discriminative learning, graph embedding, gaussian distribution.**

I. Introduction

WITH the rapid develop of multimedia technologies, image-set based face recognition problem attracts more and more attention. This problem naturally arises to suffice for a wide range of real-world applications such as video surveillance, classification with images from multi-view cameras or photo albums, and classification based on long term observations [1]–[6]. For the task of image-set based face recognition, both the gallery and the probe samples are image sets, each of which contains many facial images or video frames belonging to one single person. Compared with the single-shot image, the numerous images in each set naturally cover more variations in the subject's face appearance due to changes of pose, expression and/or lighting. Some useful data variability information is incorporated implicitly in the image set, thus more appealing performance is expected. However, it also poses new challenges on the extraction and utilization of such information.

To represent the data variability in an image set, the probabilistic model seems a natural choice. Among many others, Gaussian Mixture Model (GMM) can precisely capture the data variations with a multi-modal density by using a varying number of Gaussian components. Theoretically, after modeling image set by GMM, the dissimilarity between any two sets can be computed as the distribution divergence between their GMMs. However, it is not adequate for classification tasks that need more discriminability. Especially, when the gallery and probe sets have weak statistical correlations, larger fluctuations in performance were observed [3], [5], [7]–[9].

To address the above problem, in this paper we propose to learn a discriminative representation for Gaussian distributions and then measure the dissimilarity of two sets with the distance between the learned representations of pair-wise Gaussian components respectively from either GMM. However, according to information geometry [10], Gaussian distributions lie on a specific statistical manifold which follows Riemannian geometry. Therefore, discriminant analysis methods developed

Manuscript received February 6, 2017; revised August 4, 2017; accepted August 15, 2017. Date of publication August 30, 2017; date of current version October 17, 2017. This work was supported in part by the Natural Science Foundation of China under Contract 61390511, Contract 61379083, Contract 61650202, and Contract 61672496, in part by the 973 Program under Contract 2015CB351802, and in part by the Youth Innovation Promotion Association, CAS, under Grant 2015085. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Junsong Yuan. *(Corresponding author: Ruiping Wang.)*

W. Wang, R. Wang, and X. Chen are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wen.wang@vipl.ict.ac.cn; wangruiping@ict.ac.cn; xlchen@ict.ac.cn).

Z. Huang is with the Computer Vision Laboratory, ETH Zürich, 8092 Zürich, Switzerland (e-mail: zhiwu.huang@vision.ee.ethz.ch).

S. Shan is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: sgshan@ict.ac.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes a PDF file which proves the positive definiteness of kernels for the proposed kernel-based DARG and also gives the formula derivation for graph-based DARG. Contact wangruiping@ict.ac.cn for further questions about this work.

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

Digital Object Identifier 10.1109/TIP.2017.2746993

in Euclidean space cannot be applied directly. We thus propose a novel method of Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG). We give a comprehensive investigation of the distance metrics between Gaussians. Accordingly, two discriminative learning frameworks are devised specifically for such manifold. Among them, one framework is based on deriving the provably positive definite kernels to embed the manifold into a high-dimensional Hilbert space, which follows Euclidean geometry. Moreover, through these kernels, a deliberately devised weighted Kernel Discriminant Analysis is utilized to discriminate the Gaussians from different subjects with their prior probabilities incorporated. The other framework extends the classical graph embedding method to the manifold, where the distance metric between Gaussians is used to construct the adjacency graph. This drives the graph embedding to satisfy that after projected, the Gaussians can be better classified while they also follow consistent geometric properties with the original Gaussians.

A preliminary version of the method has been published in [11]. Compared with the conference version, this paper has the following major extensions: 1) We present an alternative discriminative learning method on Riemannian manifold of Gaussians, namely the graph-based DARG, together with the kernel-based solution in the conference version, drives the proposed DARG method more comprehensive with enhanced scalability. 2) We provide a more detailed description of the proposed method and a wider analysis about the differences with related works. 3) More extensive experiments are carried out to compare with other state-of-the-art methods and to evaluate each stage in the whole method, followed with a more detailed discussion.

A. Previous Work

For face recognition with image sets, a lot of relevant approaches have been proposed recently. According to how to model the image sets, these approaches can be roughly classified into four categories: linear/affine subspace based methods, nonlinear manifold based methods, reconstruction model based methods and statistical model based methods. They are briefly reviewed as follows.

Linear/affine subspace based methods assume that each image set spans a linear or affine subspace. Among them, Mutual Subspace Method (MSM) [1] and Discriminant-analysis of Canonical Correlations (DCC) [12] represent each image set as a single linear subspace and compute the principal angles of two linear subspaces for classification. While in [3], [8], and [13]–[17],, each image set is approximated with one or multiple convex geometric region (the affine or convex hull) and a convex optimization is used to match the closest "virtual" points. Grassmann Discriminant Analysis (GDA) [18] and Grassmann Embedding Discriminant Analysis (GEDA) [19] both formulate the image sets as points (i.e. linear subspace) on the Grassmann manifold, and define Grassmann kernel based on principal angles to conduct discriminative learning on the manifold. Huang *et al.* [20] share similar image set model but propose to learn the Projection Metric directly on Grassmann manifold rather than in Hilbert space. Since the image sets usually have a relatively large number of images and cover complicated variations of view-point, lighting and expression, linear/affine subspace based methods are hard to satisfactorily model the nonlinear variations in facial appearance.

To address the limitation of subspace modeling, image set is modeled by more sophisticated nonlinear manifold which is usually approximated by a couple of linear subspaces in the literature. In Manifold-Manifold Distance (MMD) [7], [21], each image set is assumed to span a nonlinear manifold that can be partitioned into several local linear subspaces and the similarity between manifolds is converted into integrating the distances between pair-wise subspaces. Manifold Discriminant Analysis (MDA) [22] further extends MMD to work in a discriminative feature space rather than the original image space. Cui *et al.* [5] adopt the similar set modeling strategy but align the image sets with a generic reference set for more precise local model matching. Chen *et al.* [23] propose to utilize joint sparse approximation to search the nearest local linear subspaces and consequently measure the image set distance using distance between the nearest pair of subspaces. The main shortcoming of nonlinear manifold based methods is that they require a large data set with dense sampling to satisfy the manifold assumption and that they mainly use the relatively weak information (subspace angles) to measure the distance [4], [24].

Different from the two trends of image set modeling above, reconstruction models are proposed to learn the image set representation implicitly and the dissimilarity between image sets can then be computed by the reconstruction error. For instance, video-based dictionary [25] and joint sparse representation [26] generalize the works of sparse representation and dictionary learning from still image based to video based face recognition. More recently, Lu *et al.* [27] propose to learn discriminative features and dictionaries simultaneously. In addition, an Adaptive Deep Network Template (ADNT) [59] uses deep model to represent image sets. Chen [28] propose a Dual Linear Regression Classification (DLRC) method to find a "virtual" face image located in the intersection of the subspaces spanning from different image sets. As a further extension of DLRC, a pairwise linear regression classification (PLRC) method [29] is proposed to further increase the unrelated subspace for classification. For the reconstruction model based methods, classification is usually conducted based on the minimum residual from the learned class-specific models.

In the literature, statistical models have also been employed for image set modeling due to their capacity in characterizing the set data distribution more flexibly and faithfully. Two pioneering works [2], [30] in earlier years represent the image set with some well-studied probability density functions, such as single Gaussian in [2] and Gaussian Mixture Model (GMM) in [30]. The dissimilarity between two distributions is then measured by the classical Kullback-Leibler divergence (KLD). Since both approaches are unsupervised, it was observed that their performance may have large fluctuations when the gallery and probe data sets have weak statistical correlations [12]. In some image representation extraction works for single image classification, such as [31] and [32], GMM is also

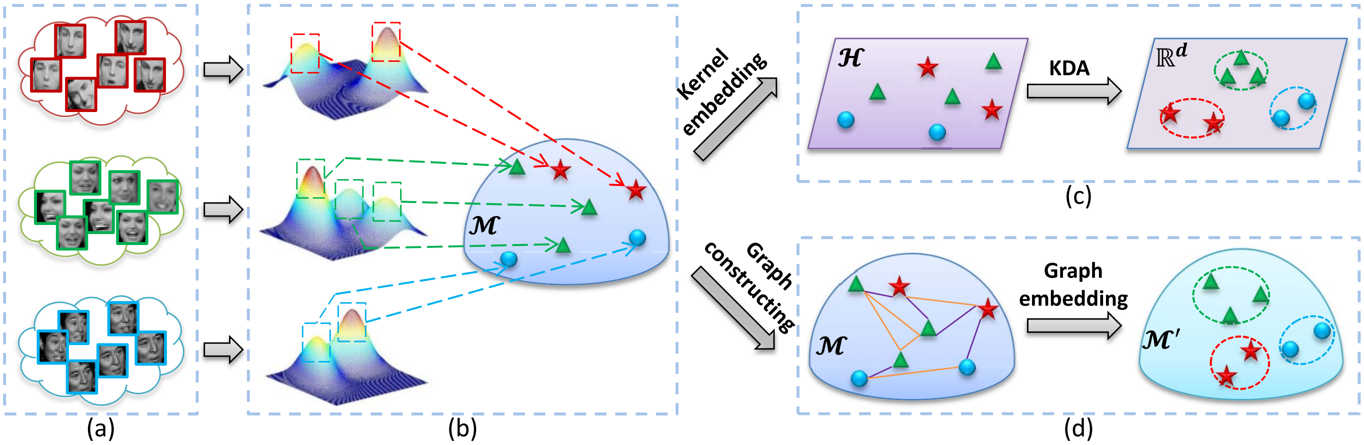


Fig. 1. Conceptual illustration of the proposed approach. (a) Training image sets in the gallery. Without loss of generality, we only demonstrate the image sets of three subjects here, with different colors denoting different subjects. (b) Modeling each image set with a GMM. The PDF of each component Gaussian is parameterized by its sample mean and covariance matrix and lies on a specific Riemannian manifold $\mathcal{M}$. Different legends (i.e. star, circle and triangle) denote the component Gaussians of different subjects. (c) Kernel-based DARG. By using kernels defined on $\mathcal{M}$, the Gaussian components are mapped to a high-dimensional Hilbert space $\mathcal{H}$, which is further discriminatively reduced to a lower-dimensional subspace $\mathbb{R}^d$. (d) Graph-based DARG. The adjacency graph is constructed which incorporates penalty and intrinsic information between the original Gaussian components. Through such graph, we perform graph embedding and enable the joint embedding of multiple Gaussian components into a more discriminative and low-dimensional statistical manifold $\mathcal{M}'$.

employed to represent the image appearance, which focuses on the distribution of the local descriptors extracted from one single image rather than the relationship between different images.

More recently, other statistical models, such as some natural statistics, are explored for image set modeling. Covariance Discriminative Learning (CDL) [4] is proposed to model the image set by its natural second-order statistic, i.e. covariance matrix, and further conduct discriminative learning on the manifold spanned by non-singular covariance matrices. While only covariance information is modeled in CDL, Lu *et al.* [24] propose to combine multiple order statistics. Symmetric Positive Definite Manifold Learning (SPDML) [33] is proposed to learn an orthonormal projection from the high-dimensional SPD manifold to a low-dimensional, more discriminative one. Log-Euclidean Metric Learning (LEML) [34] learns a tangent map from the original tangent space to a new discriminative one. An image set matching Beyond Gaussian (BG) method [35] is presented which exploits kernel density estimators to estimate the probability distribution function (PDF) of the image set and studies the kernel and dimensional reduction of PDFs on the statistical manifold.

Besides, some works, such as [36] and [37], propose to extract adapted features by deep networks. Specifically, the multi-manifold deep metric learning (MMDML) approach [36] learns feature learning networks, one for each class, by maximizing the manifold margin of different classes. Zhang *et al.* [37] address the video face clustering problem by performing the adaptation of deep representation through iteratively discovered weak pairwise identity constraints.

B. Overview of Our Approach

For our proposed approach, the overall schematic flowchart is illustrated in Fig. 1. As mentioned above, we aim at not only modeling the rich variations in each image set but also discovering discriminative invariant information hidden in the variations. Therefore, our method includes two main stages: statistical modeling of each image set with GMM and discriminative learning of the component Gaussians in the GMMs. The first stage shown in Fig. 1(b) is quite standard, which can be achieved by EM-like techniques. Each component Gaussian is then represented by its sample mean and covariance matrix, as well as an associated prior probability.

The second stage is however non-trivial as Gaussian distributions lie on a Riemannian manifold [10] while most existing discriminant analysis techniques only work in Euclidean space. This motivates us the idea of Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG). To bridge such gap between data on a statistical manifold (i.e. Gaussians) and the conventional discriminative methods in the Euclidean space, our DARG develops two discriminative learning frameworks, among which, one projects data on the manifold to some Euclidean space (i.e. the Hilbert space defined by the derived kernels) while the other reforms some conventional discriminative method in Euclidean space to precisely match the manifold.

As shown in Fig. 1(c), the kernel-based framework proposes to derive specific kernels to embed the Riemannian manifold of Gaussians $\mathcal{M}$ into a high-dimensional Hilbert space $\mathcal{H}$, which is then further discriminatively reduced to a lower-dimensional subspace $\mathbb{R}^d$. It respects the Riemannian geometry of the manifold and simultaneously enables seamless combination with conventional discriminative algorithms in Euclidean space. In our implementation, by treating the Gaussians in GMMs as samples and their prior probabilities as sample weights, we devise a weighted Kernel Discriminant Analysis to maximize the margin between Gaussians from different classes.

For the graph-based framework, Fig. 1(d) gives a conceptual illustration. To meet the specific statistical and geometric properties on the manifold $\mathcal{M}$, we extend the classical graph embedding method in Euclidean space [38] by constructing the adjacency graph with the distance metrics between Gaussians, which finally embeds the manifold into a lower-dimensional and discriminative manifold $\mathcal{M}'$. Our graph-based framework exactly fits the discriminative learning of Gaussians and simultaneously inherits the property of classical graph embedding method in preserving the geometric structure.

After learning discriminative representations of Gaussians via either of the two frameworks above, classification can be easily performed by exploiting the minimal distance between discriminative representations of component Gaussians from either GMM of the gallery and probe image sets.

II. GMM Modeling

In face recognition with image sets, it is often insufficient to model the face image set with one single model, because the image sets are usually highly nonlinear and cover large data variations. Therefore, a multi-modal density mixture model, i.e. Gaussian Mixture Models (GMM), is utilized to represent these variations efficiently in this study.

Formally, given an image set containing K images, denoted by $X = \{x_1, x_2, \ldots, x_K\}$, where x_j is the D-dimensional feature vector of the j-th image, we start with estimating GMM by the Expectation-Maximization (EM) algorithm. The estimated GMM can be written as:

$$\begin{aligned} G(x) &= \sum_{i=1}^{m} w_i g_i(x), \\ g_i(x) &= \mathcal{N}(x|\mu_i, \Sigma_i), \end{aligned} \tag{1}$$

where x denotes the feature vector of an image in this set, $g_i(x)$ is a Gaussian component with prior probability w_i, mean vector μ_i, and covariance matrix Σ_i. To facilitate subsequent processing, a small positive perturbation is added to the diagonal elements of this covariance matrix, which can make the matrix non-singular.

As an optimization method, the EM algorithm often gets stuck to local optima, and hence is sensitive to the initialization of the model. The simplest way to initialize GMM is to set a few clusters of data points randomly or by k-means clustering. However, different image sets usually have varying numbers of samples and thus the initial number of Gaussian components for each set should also be varying which is determined according to the set size. Considering the nonlinear data distribution in image set, we resort to the local linear model construction algorithm in [22] and [39], i.e. Hierarchical Divisive Clustering (HDC), which is able to generate the initialization adaptively and efficiently.

III. Kernel-Based DARG

In this section, we give a detailed description of our proposed kernel-based DARG. As mentioned above, the proposed kernel-based framework has two key ingredients: (a) Kernels derived from various Gaussian distances, (b) Kernel discriminative learning. These key ingredients are respectively detailed in the following two subsections.

A. Kernels Derived From Various Gaussian Distances

By GMM modeling, each image set that typically contains tens to hundreds of image samples is reduced to a number of Gaussian components with prior probabilities, which lie on a specific Riemannian manifold. Since Gaussian distribution functions have jointly encoded both the first order (mean) and second order (covariance) statistics, it is nontrivial to manipulate them with traditional algorithms in Euclidean space. Inspired by recent progress of learning on manifold [4], [18], [19], [40], [41], we derive corresponding positive definite kernels to encode the geometry of the manifold of Gaussians. Unlike existing methods, the kernel here is a measure of similarity between probability distributions rather than similarity between points in a feature space [42].

For constructing probabilistic kernels for Gaussians, we investigate the statistical distances quantifying the difference between two statistical distributions. The important and well established statistical distances include the following: f-divergence and its specific examples such as Kullback-Leibler Divergence and Hellinger distance, Bhattacharyya distance, Mahalanobis distance, Bregman divergence, Jensen-Shannon divergence, etc. Besides, there are also some distances specifically for Gaussians, such as the distance based on Lie Group [43], [44], the distance based on Siegel Group [45], etc. Because positive definite kernels can define valid Reproducing Kernel Hilbert Space (RKHS) and further allow the kernel methods in Euclidean space to be generalized to nonlinear manifolds, it should be guaranteed that the defined kernels are positive definite. Therefore, we study several representative distances that can be computed in closed-form and further derive provably positive definite probabilistic kernels.

1) Kullback-Leibler Kernel: A common distance between Gaussian distributions is Kullback-Leibler Divergence (KLD). Formally, given two Gaussian distributions $g_i = (\mu_i, \Sigma_i)$ and $g_j = (\mu_j, \Sigma_j)$, their KLD is computed by

$$\begin{aligned} KLD(g_i \| g_j) = \frac{1}{2}\bigg(& tr(\Sigma_j^{-1}\Sigma_i) + (\mu_j - \mu_i)^T \Sigma_j^{-1}(\mu_j - \mu_i) \\ & - \ln\Big(\frac{det\,\Sigma_i}{det\,\Sigma_j}\Big) - D\bigg), \end{aligned} \tag{2}$$

where D is the feature dimension and thus the dimension of Gaussian distributions.

By exponentiating the symmetric KLD (a.k.a. Jeffreys divergence), we define Kullback-Leibler kernel for Gaussian distributions as follows,

$$K_{KLD}(g_i, g_j) = \exp\Big(-\frac{KLD(g_i\|g_j) + KLD(g_j\|g_i)}{2t^2}\Big), \tag{3}$$

where t is the kernel width parameter. Hereinafter, it is similarly used in the following kernel functions.

2) Bhattacharyya Kernel: Bhattacharyya Distance (BD) is also a widely-used distance measure in statistics. For Gaussian distributions g_i and g_j, BD can be computed as follows,

$$\begin{aligned} BD(g_i, g_j) = & \frac{1}{8}(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j) \\ & + \frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_i \det\Sigma_j}}\right), \end{aligned} \tag{4}$$

where $\Sigma = \frac{\Sigma_i + \Sigma_j}{2}$.

Then, by exponentiating BD, we define Bhattacharyya kernel for Gaussian distributions as

$$K_{BD}(g_i, g_j) = \exp\left(-\frac{BD(g_i, g_j)}{2t^2}\right). \tag{5}$$

3) Hellinger Kernel: Hellinger Distance (HD) is closely related to BD, and can be formulated as

$$HD(g_i, g_j) = \left(1 - \frac{\det(\Sigma_i)^{1/4}\det(\Sigma_j)^{1/4}}{\det(\Sigma)^{1/2}} \cdot \exp\left\{-\frac{1}{8}(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)\right\}\right)^{1/2}, \tag{6}$$

where $\Sigma = \frac{\Sigma_i + \Sigma_j}{2}$.

Thus the corresponding Hellinger kernel is

$$K_{HD}(g_i, g_j) = \exp\left(-\frac{HD^2(g_i, g_j)}{2t^2}\right). \tag{7}$$

4) Kernel Based on Lie Group: Under the framework of information geometry, it is developed in [43] that the space of D-dimensional Gaussian distributions can be embedded into a space of $(D + 1) \times (D + 1)$ symmetric positive definite (SPD) matrices. The embedding is accomplished via mapping from affine transformation $(\mu, \Sigma^{1/2})$ into a simple Lie Group and then mapping from the Lie Group into the space of $(D+1) \times (D+1)$ SPD matrices. Thus Log-Euclidean distance [46] can be readily used to measure the distance in this space of $(D + 1) \times (D + 1)$ SPD matrices. Let P_i and P_j denote the SPD matrices corresponding to two Gaussian distributions g_i and g_j respectively. Then, the distance based on Lie Group (LGD) is defined as follows:

$$LGD(g_i, g_j) = \|\log(P_i) - \log(P_j)\|_F, \tag{8}$$

where $P = |\Sigma|^{-\frac{1}{D+1}} \begin{pmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{pmatrix}$.

Then by exponentiating the square of LGD, we define a kernel based on Lie Group to measure the similarity between $(D+1) \times (D+1)$ SPD matrices, which further measures the similarity between Gaussians as follows.

$$K_{LGD}(g_i, g_j) = \exp\left(-\frac{LGD^2(P_i, P_j)}{2t^2}\right). \tag{9}$$

5) Kernel Based on Mahalanobis Distance and Log-Euclidean Distance: Besides the above statistical distances, we can also measure the similarity respectively for the two main statistics in Gaussian distribution, i.e. mean and covariance matrix. While the former lies in the Euclidean space, the latter, after regularized to symmetric positive definite (SPD) matrix, lies on the SPD manifold. We choose Mahalanobis distance (MD) to measure the dissimilarity between means

$$MD(\mu_i, \mu_j) = \sqrt{(\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)}, \tag{10}$$

and Log-Euclidean distance (LED) for covariance matrices

$$LED(\Sigma_i, \Sigma_j) = \|\log(\Sigma_i) - \log(\Sigma_j)\|_F. \tag{11}$$

Then we tend to fuse the two distances and construct an integrated kernel for Gaussians. However, simply exponentiating their sum cannot yield a positive definite kernel and will suffer from numerical instability. Instead, we derive kernels from the two distances respectively and subsequently linearly combine them to form a valid kernel for Gaussians. Specifically, the kernel based on MD is defined as

$$K_{MD}(\mu_i, \mu_j) = \exp\left(-\frac{MD^2(\mu_i, \mu_j)}{2t^2}\right), \tag{12}$$

while the kernel based on LED is formulated by

$$K_{LED}(\Sigma_i, \Sigma_j) = \exp\left(-\frac{LED^2(\Sigma_i, \Sigma_j)}{2t^2}\right). \tag{13}$$

Finally we fuse the two kernels in a linear combination form to measure the similarity between Gaussians as follows,

$$K_{MD+LED}(g_i, g_j) = \gamma_1 K_{MD}(\mu_i, \mu_j) + \gamma_2 K_{LED}(\Sigma_i, \Sigma_j), \tag{14}$$

where γ_1 and γ_2 are the combination coefficients.

a) Positive definiteness of the kernels: Following the definition, we can easily prove that such kernels (except Kullback-Leibler kernel) derived above are positive definite. For space limitation, please refer to our supplementary materials for detailed proof analysis of the validity and positive definiteness of these kernel. While currently it is hard to theoretically justify the positive definiteness of Kullback-Leibler kernel, it can still be used as a valid kernel and the numerical stability is guaranteed by shifting the kernel width [47].

B. Kernel Discriminative Learning

Exploiting the kernels for Gaussian distributions introduced in the above section, we can naturally extend the kernel algorithms in Euclidean space to Riemannian manifold of Gaussian distributions. Here we develop a weighted Kernel Discriminant Analysis to discriminate component Gaussians of different classes with their prior probabilities incorporated as sample weights.

Formally, suppose we have n image sets belonging to c classes for training. From their GMM models, we collect all the N Gaussian components denoted by $g_1, g_2, \ldots, g_N$, which lie on a Riemannian manifold $\mathcal{M}$. Among them, the Gaussians from the i-th class are denoted as $g_1^i, g_2^i, \ldots, g_{N_i}^i$, $(\sum_{i=1}^{c} N_i = N)$, with each g_j^i accompanied a prior probability w_j^i. Let $k(g_i, g_j) = \langle\phi(g_i), \phi(g_j)\rangle$ denote a kernel function (which can be any one of the kernels in Section III-A) measuring the similarity of two Gaussians, where $\phi(\cdot)$ maps points on $\mathcal{M}$ into a high-dimensional Hilbert space $\mathcal{H}$. For a local Gaussian g_j^i, we denote $k_j^i = [k(g_j^i, g_1), \ldots, k(g_j^i, g_N)]^T \in \mathbb{R}^N$.

To perform discriminative learning with the samples g_j^i as well as their corresponding weights w_j^i, in this study we develop a weighted extension of KDA, which can be formulated as maximizing the following optimization objective $J(\alpha)$ using kernel trick similar to [48].

$$J(\alpha) = \frac{|\alpha^T B\alpha|}{|\alpha^T W\alpha|}, \tag{15}$$

Algorithm 1 Kernel-Based DARG

```
Input:
    GMMs and labels of $n$ image sets for training: $\{G_1, l_1\}, ..., \{G_n, l_n\}$. Denote the number of Gaussians in the $k$-th image set by $N_k$, and the Gaussians from all the training GMMs by $g_1, ..., g_N$, where $N = \Sigma_{k=1}^{n} N_k$;
    GMM of an image set $G^{te}$ for test, and its component Gaussians are denoted by $g_1^{te}, ..., g_M^{te}$;
Output:
    Label of the test image set $l^{te}$.
Compute $k_i^{tr} = [k(g_i, g_1), ..., k(g_i, g_N)]^T$ and $k_j^{te} = [k(g_j^{te}, g_1), ..., k(g_j^{te}, g_N)]^T$ by Equation (3) (or any of Equation (5), (7), (9), (14)), $i \in [1, N]$, $j \in [1, M]$;
Compute transformation matrix A by maximizing Equation (15);
Compute projections $z_1^k, ..., z_{N_k}^k$ of the $N_k$ Gaussians belonging to the $k$-th image set, $k \in [1, n]$;
Compute projections $z_1^{te}, ..., z_M^{te}$ of the $M$ Gaussians belonging to the test set;
Compute cosine similarity $cos(z_i^{te}, z_j^k)$ between $z_i^{te}$ and $z_j^k$;
Compute $\hat{k} = \arg\max_k cos(z_i^{te}, z_j^k)$, for all $i \in [1, M]$, $j \in [1, N_k]$;
return $l^{te} = l_{\hat{k}}$;
```

where

$$B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T,$$

$$W = \sum_{i=1}^{c} \frac{1}{\omega_i} \sum_{j=1}^{N_i} (k_j^i - m_i)(k_j^i - m_i)^T, \tag{16}$$

and

$$m_i = \frac{1}{N_i \omega_i} \sum_{j=1}^{N_i} w_j^i k_j^i, \quad m = \frac{1}{N} \sum_{i=1}^{c} \frac{1}{\omega_i} \sum_{j=1}^{N_i} w_j^i k_j^i, \tag{17}$$

Note that $\omega_i = \sum_{j=1}^{N_i} w_j^i$ is used to normalize the weights of samples belonging to a single class to guarantee the sum of them is equal to 1. Then the optimization problem can be reduced to solving a generalized eigenvalue problem: $B\alpha = \lambda W \alpha$. Supposing its $(c-1)$ leading eigenvectors are $\alpha_1, \alpha_2, \ldots, \alpha_{c-1}$, we obtain $A = [\alpha_1, \alpha_2, \ldots, \alpha_{c-1}] \in \mathbb{R}^{N \times\ (c-1)}$. Furthermore, the discriminative projection of a new Gaussian $g_t \in \mathcal{M}$ is given by $z_t = A^T k_t$, where $k_t = [k(g_t, g_1), \ldots, k(g_t, g_N)]^T \in \mathbb{R}^N$.

In the testing stage, given a test image set modeled by a GMM, we first compute the discriminative representations of its component Gaussians. Then face recognition can be simply achieved by finding the maximal one among all possible cosine similarities between the discriminative representations of the test set and those of all the training sets. The Algorithm 1 summarizes the training and testing process of our proposed kernel-based DARG.

IV. Graph-Based DARG

As an alternative solution of the kernel-based framework which often scales poorly with large data size and high data dimension, we propose to exploit a graph-based discriminative learning framework to conduct Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG).

While the proposed graph-based framework is inspired from the graph embedding methods working in Euclidean space [38], we tailor the framework to data on the manifold of Gaussian distributions. Illuminated by graph-based dimensional reduction on Riemannian manifold in [33], [35], and [49], we similarly aim to learn a r $(r < D)$ dimensional latent feature space where data distributions from different classes can be better discriminated with the geometric structure of the original component Gaussians preserved properly. Note that when we map the original features associated with a component Gaussian with a linear transformation $F \in \mathbb{R}^{r \times D}$, the obtained projections still follow Gaussian distribution according to the specific property of Gaussain distribution. It guarantees that the projected points still fall into the manifold of Gaussians, i.e., for $\forall x \sim g(x) = \mathcal{N}(x|\mu, \Sigma)$, we have

$$y = F^T x \sim \hat{g}(y) = \mathcal{N}(y|F^T\mu, F^T \Sigma F), \tag{18}$$

which supports the theoretical rationality of our proposed graph-based DARG.

As described in Section III-B, we collect all the N Gaussian components $g_1, \ldots, g_N$ from the n GMMs which respectively estimated on each image set.

Formally, the objective function is of the following form.

$$\mathcal{J}(F) = \sum_{i,j=1}^{N} S(g_i, g_j) Dist(\hat{g}_i, \hat{g}_j), \tag{19}$$

where $\hat{g}_i = \mathcal{N}(F^T \mu_i, F^T \Sigma_i F)$ is the data distribution after mapping g_i by a transformation F. S denotes the affinity matrix defined based on the corresponding distances between Gaussians $Dist(\cdot, \cdot)$. Here, we employ KLD, BD or HD as $Dist(\cdot, \cdot)$ to measure the distance between Gaussians, as LGD is invariant under any F.

For a more stable optimization solution, we constrain F to be orthonormal, i.e., $F^T F = I_r$, where I_r denotes a r-dimensional matrix. Thus the optimization problem can be formatted as follows.

$$F^* = \mathop{\arg\min}_{F^T F = I_r} \mathcal{J}(F) \tag{20}$$

Different from the graph studied in [38], here we construct the adjacency graph with respect of the geometric properties of statistical manifold. Let $\{\{g_i\}_{i=1}^N, S\}$ be an undirected weighted graph, where each vertex is a Gaussian component g_i and the affinity matrix $S \in \mathbb{R}^{N \times N}$ measures the compactness of intraclass Gaussians and the separability of interclass Gaussians. Formally, it can be defined in the following.

$$S(g_i, g_j) = \begin{cases} s_{ij}, & \text{if } g_i \in N_w(g_j) \text{ or } g_j \in N_w(g_i) \\ -s_{ij}, & \text{if } g_i \in N_b(g_j) \text{ or } g_j \in N_b(g_i) \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where $s_{ij} = exp(-Dist(g_i, g_j)^2/\sigma)$ gives a larger weight for a closer pair of Gaussian components and σ is a constant. $N_w(g_i)$ contains K_w nearest neighbors of g_i that share the same label y_i, while $N_b(g_i)$ consists of K_b nearest neighbors of g_i belonging to different classes from g_i. Here we define the nearest neighbor as the Gaussian component with the smallest distance. Note that we can alternatively define the nearest neighbor through other strategies, such as some ε-ball.

Since Equation (20) is an optimization problem with an orthonormality constraint, we rewrite it as an unconstrained optimization problem on a Grassmann manifold, which can be solved using the gradient descent method on the Grassmann manifold. Its reasonability can be verified easily as the objective function $\mathcal{J}(F)$ obviously satisfies the fact that $\mathcal{J}(F) = \mathcal{J}(FT)$ for any orthogonal matrix $T \in \mathbb{R}^{r\times r}$. For further details on optimization by gradient descent on Grassmann manifold, please refer to [49].

Corresponding to different distances between Gaussians $Dist(\cdot,\cdot)$, we calculate the partial derivative of the objective function $\mathcal{J}$ with respect to the transformation matrix F, i.e.,

$$\frac{\partial}{\partial F}\mathcal{J}(F) = \sum_{i,j=1}^{N} S(g_i, g_j)\frac{\partial}{\partial F}Dist(\hat{g}_i, \hat{g}_j) \tag{22}$$

The detailed formulation and the corresponding derivation of the above formula can be found in our supplementary materials. Though without a theoretical proof for the convergence of the proposed optimization algorithm, we will conduct experiment to illustrate its convergence in Section VI.

For the test stage, we take a test image set as an example to give an introduction. Through Equation (18), we first calculate the projections of its component Gaussians as more discriminative and lower-dimensional representations. To perform matching between this test set and some training image set, we compute the distances $Dist(\cdot,\cdot)$ between all the projected Gaussians of the test set and those of the training set. Then we utilize the minimal one among these distances as the dissimilarity between the test image set and the training image set and classify the test image set with a simple nearest neighbor classifier. The proposed graph-based DARG method is summarized in Algorithm 2.

Algorithm 2 Graph-Based DARG

```
Input:
    GMMs and labels of $n$ image sets for training: $\{G_1, l_1\}, ..., \{G_n, l_n\}$. Denote the number of Gaussians in the $k$-th image set by $N_k$, and the Gaussians from all the training GMMs by $g_1, ..., g_N$, where $N = \Sigma_{k=1}^{n} N_k$;
    GMM of an image set $G^{te}$ for test, and its component Gaussians are denoted by $g_1^{te}, ..., g_M^{te}$;
Output:
    Label of the test image set $l^{te}$.
1: Compute the undirected weighted graph $\{\{g_i\}_{i=1}^{N}, S\}$ by Equation (21);
2: Initialize the transformation matrix $F$ with appropriate values;
3: Compute the value of $\mathcal{J}(F)$ by Equation (19);
4: Update $F$ by Equation (22);
5: Return to Step 3, until it converges and an optimal value of $F$, denoted by $F^*$, is finally obtained;
6: Compute projections $\hat{g}_i{}^k, ..., \hat{g}_{N_k}^k$ of the $N_k$ Gaussians belonging to the $k$-th image set by Equation (18), for $k \in [1, n]$;
7: Compute projections $\hat{g}_1^{te}, ..., \hat{g}_M^{te}$ of the $M$ Gaussians belonging to the test set by Equation (18);
8: Compute $Dist(\hat{g}_i^{te}, \hat{g}_j^k)$ between $\hat{g}_i^{te}$ and $\hat{g}_j^k$;
9: Compute $\hat{k} = \arg\min_k Dist(\hat{g}_i^{te}, \hat{g}_j^k)$, for all $i \in [1, M]$, $j \in [1, N_k]$;
10: return $l^{te} = l_{\hat{k}}$;
```

V. Discussion

A. Differences From Related Works

While our method reveals the data structure in an image set with a statistical model (i.e. GMM) comprising of multiple local models (i.e. Gaussian components) and performs discriminant analysis on a statistical manifold, it bears certain relationship and also has its unique merits compared with related works in the literature. We highlight them as follows.

1) Differences From Other Statistical Models: Compared with [2] using single Gaussian and [30] using GMM, the main difference is that discriminative information is used in our method such that it can achieve significantly improved resistance to the weak statistical correlation between training and test data. CDL [4] models the image set with its covariance matrix, which also inherited in SPDML [33] and LEML [34] which further perform discriminative learning on the SPD manifold and in the corresponding tangent space respectively. However, these methods all ignore the mean information, which may lead to missing of some useful underlying data variability information. LMKML [24] combines multiple order statistics as the feature of image set, but simply treats both the 2nd order covariance matrix and the 3rd order tensor as vectors, which ignores the inherent manifold geometric structure. Different from the methods above, our method creatively proposes to explore discriminative learning on a specific Riemannian manifold of statistical distributions.

After our conference version [11], a more recent method BG [35] adopts a similar strategy of modeling the image set with some probability distribution function (PDF). The main differences are listed in the following. 1) Density estimation. To estimate the PDF for each image set, BG utilizes the non-parametric and data-driven Kernel density estimation (KDE) which gives a flexibility estimation theoretically, whereas in the implementation a good KDE is usually difficult to calculate with sensitivity to the kernel width. On the contrary, as a semi-parametric model, GMM can leverage the efficiency of parametric methods and the flexibility of non-parametric methods. 2) Distance metric. In BG, for each set, the estimated PDF is used as a single model, and the divergence between PDFs needs to be approximated, which leads to a high complexity and consuming time. On the contrary, our method alleviates it by learning representations for multiple local models, rather than for each GMM, which ensures that the corresponding PDF distance can be easily computed in closed-form.

2) Differences From Other Set Structure Models: DCC [12], GDA [18], GEDA [19] and PML [20] model the data structure of an image set under the assumption of linear subspace. For these methods, the linear subspace is formulated by an orthonormal basis matrix, and the distance between them is measured with the principal angles or the projection metric. That is to say, for the image set modeling and measuring, the mean information is ignored, which however usually incorporates some useful information. In contrast, our approach endeavors to measure the distribution distinction of local Gaussian models with the distance incorporating both mean and covariance information.

MMD [7], MDA [22] and SANS [23] all approximate the data structure in an image set with multiple linear subspaces, but the linear subspaces are computed by a hard partition that neglects the probabilistic distribution of the set data, which is mainly encoded with the Gaussian distribution in this work. Moreover, for measuring the distance between linear subspaces, MMD considers the mean and variance of data, but makes no use of discriminative information. MDA is a discriminative extension of MMD, but only involves mean information during discriminative learning. SANS measures image set distance with average distance of the nearest subspace pairs extracted by sparse approximation, but the distance is based on the relatively weaker principal angles [4], [24]. Again, SANS is non-discriminative.

B. Gaussian Component Weights

In this subsection, we give discussions about the effect of Gaussian component weights in our proposed method. 1) For the kernel-based DARG, the Gaussian component weights are used in the training stage which leads to accuracy gain. These weights are not used in the testing stage. In our experiments, we have tested several linear combination schemes, including the classical EMD, to impose the weights on the similarity computation, which however leads to trivial gain. The possible reason behind can be adduced that in the training stage we have incorporated the component weights in KDA learning, therefore the resulting discriminative model has already emphasized the components with larger weights and weakened those with smaller weights (possibly formed by noisy samples). Therefore in the testing stage it is unnecessary to weight them again. 2) The component weights are not incorporated into the graph-based DARG as in our experiments it barely improves the performance to simply employ them to weight the affinity matrix. The possible cause may be that it exerts too much influence on the encoded relationship between Gaussians while these weights mainly work when used as estimated auxiliary information in discriminative learning.

C. Contribution

In summary, our contributions mainly lie in four folds: 1) We propose a new method for discriminative learning with Gaussians on Riemannian manifold to encourage more robust image set classification. 2) Two discriminative learning frameworks are devised by respectively using the kernel function and the adjacency graph to bridge the data on the statistical manifold and the conventional discriminative learning methods in Euclidean space. 3) The kernel-based framework derives a series of kernels for Gaussians with proved positive definiteness to embed the manifold into a high-dimensional Hilbert space, which is then discriminatively reduced to a lower-dimensional subspace by designing a weighted extension of KDA. 4) The graph-based framework constructs an adjacency graph with the corresponding distance metrics between Gaussians to embed the original manifold to a lower-dimensional and discriminative target manifold with the geometric structure preserved and the interclass separability maximized.

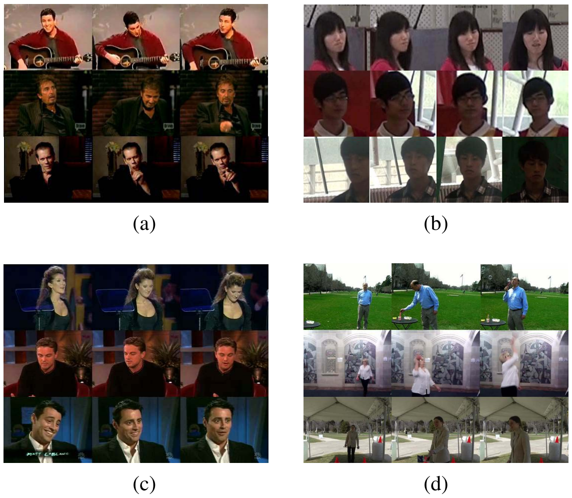


Fig. 2. Some examples of the datasets. (a) YTC. (b) COX. (c) YTF. (d) PaSC.

VI. Experiments

In this section, we firstly introduce the experimental settings for the datasets. Then we investigate different strategies for the two main stages in our proposed DARG, i.e., GMM modeling and discriminative learning. Finally, we compare our DARG with the state-of-the-art in accuracy performance and time complexity following with a detailed analysis.

A. Databases Description and Settings

We used four most challenging and largest datasets: YouTube Celebrities (YTC) [50], COX [51], YouTube Face DB (YTF) [52] and Point-and-Shoot Challenge (PaSC) [53]. Their protocol and performance metric all follow the original literature. Examples in the four datasets are shown in Fig. 2.

We performed face identification experiments on YTC and COX. YTC contains 1,910 videos of 47 subjects. We conducted ten-fold cross validation experiments and randomly selected 3 clips for training and 6 for testing in each of the ten folds. This enables the whole testing sets to cover all of the 1,910 clips in the database, which is similar with the protocol in [4], [5], [22], and [27]. COX contains 3,000 video sequences from 1,000 different subjects and has a training set containing 3 video sequences for each subject. Since the dataset contains three settings of videos captured by different cameras, we conducted ten-fold cross validation respectively with one setting of video clips as gallery and another one as probe.

To evaluate the experimental performance on face verification, we used another two datasets, YTF and PaSC. YTF contains 3,425 videos of 1,595 subjects. We followed the same settings with benchmark tests in [52]. 5,000 video pairs are collected randomly and half of them are from the same subject, half from different subjects. These pairs are then divided into 10 splits and each contains 250 'same' pairs and 250 'not-same' pairs. PaSC contains 2,802 videos of 265 people. Half of these videos are captured by controlled video camera, and the rest are captured by hand held video camera. It has a total of 280 sets for training and experiments were conducted using control or handheld videos as target and query respectively.

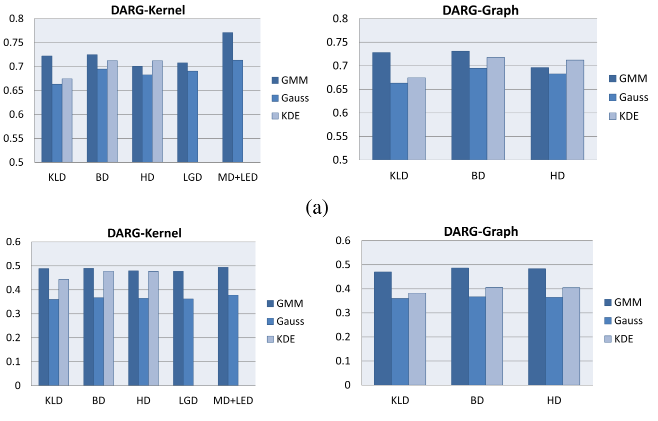
In our experiments, the cropped faces were resized to 20×20 on YTC, 32 × 40 on COX, 24 × 40 on YTF and 256 × 256 on PaSC as previous works [20], [24], [35], [54]. Then histogram equalization was implemented for the gray features of faces in YTC, COX and YTF. For PaSC is relatively difficult, we further followed [54] to extract the state-of-the-art Deep Convolutional Neural Network (DCNN) features by using Caffe [55]. We used the CFW dataset [56] for pre-training and the training data of PaSC and COX for fine-tuning.

B. Evaluations of the Main Stages

For the two stages discussed in Section I-B, we respectively investigate their performance on YTC for identification task and on control videos of PaSC for verification task.

1) GMM Modeling: Firstly we experimentally show the efficiency of GMM as a density estimation strategy for image set. Here GMM is compared with other density estimation methods which have been used to model the image set in the literature, i.e., the single Gaussian model and KDE. In our method, following GMM modeling, its Gaussian components are treated as multiple local models, while for the single Gaussian model and KDE, the estimated PDF for each image set is used to model the image set directly. To measure the dissimilarity between Gaussians, the chosen statistical distances are in close-form, while for PDFs estimated by KDE, KLD and HD are approximated as [49]. The kernel and graph-based method following with KDE respectively refer to the kernel and dimensionality reduction method in [35], where the source code is provided by the original author.

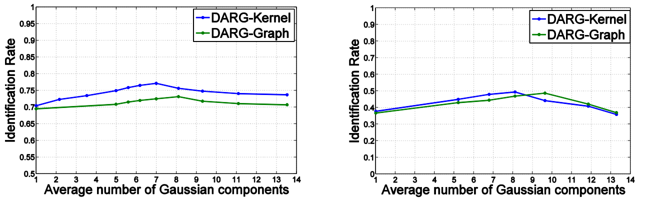
The comparison results of the three density estimation strategies (GMM, single Gaussian and KDE) are reported in Fig. 3. Note that we refer to the kernel-based DARG as "DARG-Kernel", while the graph-based framework is denoted by "DARG-Graph". In our kernel-based DARG, since kernels based on LGD and MD+LED are specific for Gaussian, we only reported their performance for GMM and single Gaussian. We can see that for different distances, it generally performs better to exploit GMM to conduct discriminative learning on multiple models than single Gaussian or KDE model, which implies that one single model is inadequate to represent the complicated variations inside the image set. Furthermore, this also gives an experimental support for the theoretical superiority of GMM over KDE as discussed in Section V-A.1.



(a)

(b)

Fig. 3. Comparison of different density estimation strategies on YTC for face identification task and control videos of PaSC for face verification task.



(a)

(b)

Fig. 4. Comparison of different average Gaussian component numbers on YTC and control videos of PaSC.

Having shown the superiority of GMM, we compared the performance of different Gaussian component numbers. Fig. 4a and Fig. 4b respectively show how the accuracy changes for DARG-Kernel with kernel based on MD+LED and DARG-Graph with BD when using different average numbers of Gaussian components on YTC and control videos of PaSC. The number of Gaussian components is different for each image set such that the average number of Gaussian components is not necessarily integer value. The results show favorable stability within a proper range of Gaussian numbers. For instance, on YTC, we get the best result with an average of about 7 or 8 Gaussian components in GMM. From the curve, we can analyze that GMM with too small number of Gaussian components may be insufficient to represent the complex variation of face images, and using too many Gaussian components in GMM may make the statistics of each Gaussian component difficult to estimate.

2) Discriminative Learning: For the stage of discriminative learning, we compared our DARG and unsupervised nearest neighbor (NN) classifier with different distances between Gaussians. The results are reported in Table I. Besides, we also compared with classifying each image set with the largest average of image probabilities estimated by GMM [57], which achieves an accuracy of 53.22% on YTC.

From the comparison results tabulated in Table I, we can see that by discriminative learning, our DARG performs

TABLE I

COMPARISON OF DIFFERENT DISTANCE METRICS ON YTC

Setting \ Method	KLD	BD	HD	LGD	MD+LED
DARG-Kernel	72.21	72.49	70.03	68.72	77.09
DARG-Graph	72.81	73.09	69.61	N/A	N/A
NN	68.86	71.42	66.58	58.61	69.22/22.06

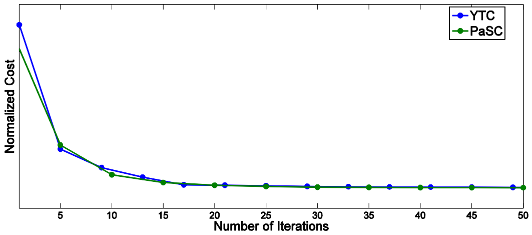


Fig. 5. Convergence of the optimization algorithm in graph-based DARG on YTC and control videos of PaSC.

better than the unsupervised classification, which indicates that discriminative information can facilitate more robust classification than directly classifying Gaussians without supervision. Thus the experiments supports our motivation of learning discriminative representation for component Gaussians. As shown in Table I, for kernel-based DARG, the kernel based on MD+LED works best among the derived kernels for Gaussians. The reason can be attributed to the fusing scheme of two statistics (i.e. mean and covariance) in the kernel combination level. This scheme is less dependent on Gaussian hypothesis and thus alleviates the measurement error in case of distribution deviating from Gaussian in real-world data.

Besides, we also give an experimental proof for the convergence of the optimization algorithm in the graph-based DARG. Fig. 5 gives the cost changing with the iteration number on YTC and on control videos of PaSC. Note that the cost is respectively normalized to $[0.1, 0.9]$. It shows that the objective function can finally achieve convergence to a stable value after a few iterations.

C. Comparison With the State-of-the-Art

We compared our performance to several groups of state-of-the-art methods for face recognition with image sets:

(1) Single/multiple linear/affine subspaces based methods:

MMD [7], MDA [22], AHISD [8], CHISD [8], GDA [18] and GEDA [19].

(2) Statistical model based methods:

SGM [2], MDM [30], CDL [4], LMKML [24] and BG [35].

Except SGM and MDM, the source codes of above methods are provided by the original authors. Since the codes of SGM and MDM have not been publicly available, we implemented them using the same GMM estimation code in our approach to generate Gaussian models. For fair comparison, the important parameters of each method were empirically tuned according to the recommendations in the original references. For all methods, we first used PCA to reduce the data dimension by preserving 95% of data energy on YTC, COX and YTF, and 80% of data energy on PaSC. In MMD and MDA, we used the default parameters as the standard implementation in [7] and [22]. For AHISD and CHISD, we searched the PCA energy when learning the linear subspace through $\{80\%, 85\%, 90\%, 95\%\}$, and reported the best result for each method. For both GDA and GEDA, the dimension of Grassmannian manifold was searched to find the best result. In CDL, we used KDA for discriminative learning and the same setting as [4] on YTC, COX and PaSC. Note that on YTF we used a kernel version of SILD [58] rather than KDA in CDL, BG and our approach because the restricted protocol of YTF limits the information available for training to the same/not-same labels. For LMKML, we utilized the same setting as [24]. In BG, we reported its performance based on Hellinger distance and tuned the parameters in the empirical range given in [35], where the KFDA and dimensionality reduction method are respected denoted by “BG-K” and “BG-DR”.

TABLE II

IDENTIFICATION RATES (%) ON YTC AND COX. HERE, “ COX-ij ” REPRESENTS THE EXPERIMENT USING THE i-TH SET OF VIDEOS AS GALLERY AND THE j-TH SET OF VIDEOS AS PROBE

Method \ Dataset	YTC	COX					
		COX-12	COX-13	COX-23	COX-21	COX-31	COX-32
MMD [7]	65.30	38.29	30.34	15.24	34.86	22.21	11.44
MDA [22]	66.98	65.82	63.01	36.17	55.46	43.23	29.70
AHISD [8]	63.69	57.54	37.99	18.57	47.91	34.91	18.79
CHISD [8]	66.46	56.87	30.10	14.80	44.37	26.44	13.68
GDA [18]	65.91	72.26	80.70	74.36	71.44	81.99	77.57
GEDA [19]	66.83	76.73	83.80	76.59	72.56	82.84	79.99
SGM [2]	52.00	26.74	14.32	12.39	26.03	19.21	10.50
MDM [30]	62.12	30.70	24.98	14.30	28.90	31.72	19.30
CDL [4]	69.70	78.37	85.25	79.74	75.59	85.83	81.87
LMKML [24]	70.31	66.00	71.00	56.00	74.00	68.00	60.00
BG-K [35]	72.71	81.62	87.70	82.65	81.91	87.17	85.45
BG-DR [35]	70.29	67.81	72.93	68.26	70.10	73.07	71.21
DARG-Kernel	77.09	83.71	90.13	85.08	81.96	89.99	88.35
DARG-Graph	73.09	76.17	83.54	77.86	74.65	83.88	79.00

For the kernel-based DARG, we took the kernel based on MD+LED as an example due to its good performance in Section VI-B. For kernel based on MD+LED, we fixed the fusing coefficient γ_1 as 1, and γ_2 was searched in the range of [0.5,1,2]. In our graph-based DARG, for constructing the graph, K_w was fixed as 3, and K_b was tuned from 10 to 30. Among the distances, we chosen BD as an example.

For face identification task, Table II reports the average recognition accuracy over multiple-fold trials on YTC and COX. For face verification task, Table III shows the area under ROC curve (AUC) on YTF. The comparisons on PaSC are shown in Table IV and performance is evaluated by the verification rate (%) at a false accept rate (FAR) of 0.01.

From these tables, it is shown that our proposed approach achieves superior performances in most tests.

(1) Among the non-discriminative methods, compared with the single modeling methods AHISD, CHISD, SGM, most of

TABLE III

COMPARISONS ON YTF. THE PERFORMANCE IS EVALUATED BY THE AREA UNDER ROC CURVE (AUC) IN THIS TABLE

Method	MMD [7]	AHISD [8]	CHISD [8]	CDL [4]	BG-K [35]	BG-DR [35]	**DARG -Kernel**	**DARG -Graph**
Result	64.96	66.50	66.24	69.74	67.15	69.32	73.01	70.84

TABLE IV

COMPARISONS ON PASC. NOTE THAT THE VERIFICATION RATES (%) AT A FALSE ACCEPT RATE (FAR) OF 0.01 ON PASC IS REPORTED IN THIS TABLE

Method Setting	GDA [18]	GEDA [19]	CDL [4]	BG-K [35]	BG-DR [35]	**DARG -Kernel**	**DARG -Graph**
Control	39.73	40.57	46.07	47.62	40.44	49.37	48.65
Handheld	37.52	38.96	44.53	44.21	39.16	48.54	47.32

TABLE V

COMPUTATION TIME (SECONDS) ON YTC FOR TRAINING AND TESTING (CLASSIFICATION OF ONE IMAGE SET)

Method Process	MDA [22]	AHISD [8]	GDA [18]	CDL [4]	BG-K [35]	BG-DR [35]	**DARG -kernel**	**DARG -Graph**
Training	11.34	N/A	3.86	4.15	167.50	235.00	114.70	107.30
Testing	0.31	0.28	0.42	0.32	0.96	1.45	0.80	0.83

the multi-model methods such as MMD, MDM achieve better performance on both datasets. This supports our motivation to apply multiple Gaussian components to model each image set.

(2) Among the discriminative methods, GDA, GEDA, CDL, BG and our proposed DARG conduct discriminative learning on the manifold, which yield better results than MDA. This is because MDA learns the discriminative metrics in Euclidean space, whereas most of them classify the sets in non-Euclidean spaces. In contrast, these methods extract the subspace-based statistics in Riemannian space and match them in the same space, which is more favorable for the set classification task.

(3) Compared with GDA and GEDA, the statistical model based methods, i.e., CDL, LMKML, BG and the proposed DARG, have shown their superiority in most of the experiments. The reason can be analyzed to be that GDA and GEDA both depend on the linearity assumption, which is hard to satisfy, while the statistical models are free from such assumption and can better represent the set variations.

(4) Among the statistical model based methods, our method achieves better performance than CDL and LMKML. This is because they only utilize the relatively weak information of set variations while our method attempts to model the data distribution and jointly fuse both mean and covariance information. In contrast with BG, our method performs better in all the experiments, which experimentally supports the discussions in Section V..

Besides the performance, another important factor is the time complexity. In Table V, we compared time costs of our method and some closely related methods on YTC using an Intel i7-3770, 3.40 GHz PC. For our method, we take DARG-Kernel with kernel based on MD+LED and DARG-Graph with BD as examples and the average number of Gaussian components is about 7. Clearly, our testing speed is comparable to those of the state-of-the-art methods. Though our training time is relatively long, it is not a big problem as the training stage can be conducted offline.

VII. CONCLUSION

In this paper, we propose a Discriminant Analysis on the Riemannian manifold of Gaussian distributions (DARG) to solve the problem of face recognition with image sets. Our method differs from tradition methods in conducting kernel discriminative learning and graph embedding for Gaussian distributions on a statistical manifold rather than for vectors in Euclidean space. We utilized GMM to represent each image set by a number of Gaussian components with prior probabilities and then gave a comprehensive investigation of the distances between Gaussians to measure the geometric properties on the manifold. Based on these distances, a series of simple but valid probabilistic kernels were derived and accordingly, a weighted Kernel Discriminant Analysis technique was devised to maximize the margin between Gaussians from different classes. Alternatively, a graph-based discriminative learning framework was established by constructing the adjacency graphs according to the distances between Gaussians to encode the geometric structure and discriminative information on the manifold. The experiments have demonstrated the superiority of our proposed approach over the state-of-the-art methods.

In the future, we intend to further advance the framework of discriminative learning on the statistical manifold with more divergences or other conventional learning methods. Moreover, the proposed method will be extended to support more general application scenarios, rather than limited to faces.

REFERENCES

[1] O. Yamaguchi, K. Fukui, and K.-I. Maeda, “Face recognition using temporal image sequence,” in *Proc. IEEE Conf. Autom. Face Gesture Recognit. (FG)*, Apr. 1998, pp. 318–323.

[2] G. Shakhnarovich, J. W. Fisher, and T. Darrell, “Face recognition from long-term observations,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2002, pp. 851–865.

[3] Y. Hu, A. S. Mian, and R. Owens, “Sparse approximated nearest points for image set classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 121–128.

[4] R. Wang, H. Guo, L. S. Davis, and Q. Dai, “Covariance discriminative learning: A natural and efficient approach to image set classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2496–2503.

[5] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, “Image sets alignment for video-based face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2626–2633.

[6] M. Du, A. C. Sankaranarayanan, and R. Chellappa, “Robust face recognition from multi-view videos,” *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1105–1117, Mar. 2014.

[7] R. Wang, S. Shan, X. Chen, and W. Gao, “Manifold-manifold distance with application to face recognition based on image set,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[8] H. Cevikalp and B. Triggs, “Face recognition based on image sets,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2567–2573.

[9] T.-K. Kim, J. Kittler, and R. Cipolla, “On-line learning of mutually orthogonal subspaces for face recognition by image sets,” *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1067–1074, Apr. 2010.

[10] S. Amari and H. Nagaoka, *Methods of Information Geometry* (Translations of Mathematical Monographs). New York, NY, USA: AMS, 2000.

[11] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, “Discriminant analysis on riemannian manifold of Gaussian distributions for face recognition with image sets,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2048–2057.

[12] T.-K. Kim, J. Kittler, and R. Cipolla, “Discriminative learning and recognition of image set classes using canonical correlations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[13] M. Yang, P. Zhu, L. V. Gool, and L. Zhang, “Face recognition based on regularized nearest points between image sets,” in *Proc. IEEE Conf. Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

[14] A. Mian, Y. Hu, R. Hartley, and R. Owens, “Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5252–5262, Dec. 2013.

[15] S. Chen, A. Wiliem, C. Sanderson, and B. C. Lovell. (Mar. 2014). “Matching image sets via adaptive multi convex hull.” [Online]. Available: https://arxiv.org/abs/1403.0320

[16] W. Wang, R. Wang, S. Shan, and X. Chen, “Probabilistic nearest neighbor search for robust classification of face image sets,” in *Proc. IEEE Conf. Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–7.

[17] W. Wang, R. Wang, S. Shan, and X. Chen, “Prototype discriminative learning for image set classification,” *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1318–1322, Sep. 2017.

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

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

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

[21] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao, “Manifold–manifold distance and its application to face recognition with image sets,” *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4466–4479, Oct. 2012.

[22] R. Wang and X. Chen, “Manifold discriminant analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 429–436.

[23] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell, “Improved image set classification via joint sparse approximated nearest subspaces,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 452–459.

[24] J. Lu, G. Wang, and P. Moulin, “Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 329–336.

[25] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, “Dictionary-based face recognition from video,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 766–779.

[26] Y.-C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips, “Video-based face recognition via joint sparse representation,” in *Proc. IEEE Conf. Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[27] J. Lu, G. Wang, W. Deng, and P. Moulin, “Simultaneous feature and dictionary learning for image set based face recognition,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 265–280.

[28] L. Chen, “Dual linear regression based classification for face cluster recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2673–2680.

[29] Q. Feng, Y. Zhou, and R. Lan, “Pairwise linear regression classification for image set retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4865–4872.

[30] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, “Face recognition with image sets using manifold density divergence,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 581–588.

[31] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, “Hierarchical Gaussianization for image classification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 2009, pp. 1971–1977.

[32] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image classification with the fisher vector: Theory and practice,” *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[33] M. T. Harandi, M. Salzmann, and R. Hartley, “From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 17–32.

[34] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, “Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 720–729.

[35] M. Harandi, M. Salzmann, and M. Baktashmotlagh, “Beyond Gauss: Image-set matching on the Riemannian manifold of PDFs,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4112–4120.

[36] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, “Multi-manifold deep metric learning for image set classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1137–1145.

[37] Z. Zhang, P. Luo, C. L. Chen, and X. Tang, “Joint face representation adaptation and clustering in videos,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 236–251.

[38] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: A general framework for dimensionality reduction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[39] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao, “Maximal linear embedding for dimensionality reduction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1776–1792, Sep. 2011.

[40] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. T. Harandi, “Kernel methods on the riemannian manifold of symmetric positive definite matrices,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 73–80.

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

[42] A. B. Chan, N. Vasconcelos, and P. J. Moreno, “A family of probabilistic kernels based on information divergence,” Univ. California, Dept. Elect. Comput. Eng., San Diego, CA, USA, Tech. Rep. SVCL-TR-2004-1, 2004.

[43] M. Lovrić, M. Min-Oo, and E. A. Ruh, “Multivariate normal distributions parametrized as a Riemannian symmetric space,” *J. Multivariate Anal.*, vol. 74, no. 1, pp. 36–48, 2000.

[44] P. Li, Q. Wang, and L. Zhang, “A novel earth mover’s distance methodology for image matching with Gaussian mixture models,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1689–1696.

[45] M. Calvo and J. M. Oller, “A distance between multivariate normal distributions based in an embedding into the Siegel group,” *J. Multivariate Anal.*, vol. 35, no. 2, pp. 223–242, 1990.

[46] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, “Log-Euclidean metrics for fast and simple calculus on diffusion tensors,” *Magn. Reson. Med.*, vol. 56, no. 2, pp. 411–421, 2006.

[47] P. J. Moreno, P. Ho, and N. Vasconcelos, “A Kullback–Leibler divergence based kernel for svm classification in multimedia applications,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2003, pp. 1385–1392.

[48] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, 2000.

[49] K. M. Carter, “Dimensionality reduction on statistical manifolds,” Ph.D. dissertation, Univ. of Michigan, Ann Arbor, MI, USA, 2009.

[50] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, “Face tracking and recognition with visual constraints in real-world videos,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[51] Z. Huang *et al.*, “A benchmark and comparative study of video-based face recognition on COX face database,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5967–5981, Dec. 2015.

[52] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 529–534.

[53] B. Ross *et al.*, “The challenge of face recognition from digital point-and-shoot cameras,” in *Proc. IEEE Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2013, pp. 1–8.

[54] J. R. Beveridge *et al.*, “Report on the FG 2015 video person recognition evaluation,” in *Proc. IEEE Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–8.

[55] Y. Jia *et al.*, “Caffe: Convolutional architecture for fast feature embedding,” in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[56] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, “Finding celebrities in billions of Web images,” *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 995–1007, Aug. 2012.

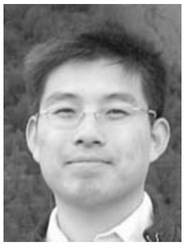
[57] C. Sanderson, S. Bengio, and Y. Gao, “On transforming statistical models for non-frontal face verification,” *Pattern Recognit.*, vol. 39, no. 2, pp. 288–302, 2006.

[58] M. Kan, S. Shan, D. Xu, and X. Chen, “Side-information based linear discriminant analysis for face recognition,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 125.0–125.1.

[59] M. Hayat, and M. Bennamoun, and S. An, “Deep reconstruction models for image set classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 713–727, Apr. 2015.

Wen Wang (S'15) received the B.S. degree in information and computing Science from Jilin University, Changchun, China, in 2011, and the Ph.D. degree in computer science and technology from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2017.

She is currently a Post-Doctoral Researcher with ICT, CAS. Her research interests mainly include computer vision, pattern recognition, machine learning and, in particular, manifold learning, metric learning, and deep learning and their application in video-based face recognition.



Ruiping Wang (S'08–M'11) 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 Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, from 2010 to 2012. He also spent one year as a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies, University of Maryland at College Park, College Park, from 2010 to 2011. He has been with the Faculty of the Institute of Computing Technology, Chinese Academy of Sciences, since 2012, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.

Zhiwu Huang (S'13–M'16) received the B.S. degree in computer science and technology from Huaqiao University, Quanzhou, China, in 2007, the M.S. degree in computer software and theory from Xiamen University, Xiamen, China, in 2010, and the Ph.D. degree in computer science and technology from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2015. Since 2015, he has been with the Computer Vision Laboratory, ETH Zürich, Zürich, Switzerland, where he is currently a Post-Doctoral Researcher. His research interests include computer vision, Riemannian computing, metric learning, and deep learning.

Shiguang Shan (M'04–SM'15) 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 (CAS), Beijing, China, in 2004. In 2002, he joined ICT, CAS, where he has been a Professor since 2010. He is currently the Deputy Director of the Key Laboratory of Intelligent Information Processing, CAS. He has authored over 200 papers in refereed journals and proceedings in computer vision and pattern recognition. His research interests include computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics. He was a recipient of the Chinas State Natural Science Award in 2015 and the Chinas State S&T Progress Award in 2005 for his research work. He is an Associate Editor of several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *Computer Vision and Image Understanding*, the *Neurocomputing*, and the *Pattern Recognition Letters*. He has served as the Area Chair for many international conferences, including ICCV11, ICPR12, ACCV12, FG13, ICPR14, ICASSP14, and ACCV16.

Xilin Chen (M'00–SM'09–F'16) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively. He was a Professor with the Harbin Institute of Technology from 1999 to 2005. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), since 2004. He has authored one book and over 200 papers in refereed journals and proceedings in computer vision, pattern recognition, image processing, and multimodal interfaces. He served as an Organizing Committee/Program Committee Member for over 50 conferences. He is a fellow of China Computer Federation. He was a recipient of several awards, including the Chinas State Natural Science Award in 2015, the Chinas State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work. He is an AE of the IEEE TRANSACTIONS ON MULTIMEDIA, a Leading Editor of the *Journal of Computer Science and Technology*, and an Associate Editor-in-Chief of the *Chinese Journal of Computers*.