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

Wen Wang[1,2], Ruiping Wang[1], Zhiwu Huang[1,2], Shiguang Shan[1], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
{wen.wang, zhiwu.huang}@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Abstract

This paper presents a method named Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG) to solve the problem of face recognition with image sets. Our goal is to capture the underlying data distribution in each set and thus facilitate more robust classification. To this end, we represent image set as Gaussian Mixture Model (GMM) comprising a number of Gaussian components with prior probabilities and seek to discriminate Gaussian components from different classes. In the light of information geometry, the Gaussians lie on a specific Riemannian manifold. To encode such Riemannian geometry properly, we investigate several distances between Gaussians and further derive a series of provably positive definite probabilistic kernels. Through these kernels, a weighted Kernel Discriminant Analysis is finally devised which treats the Gaussians in GMMs as samples and their prior probabilities as sample weights. 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.

1. Introduction

In contrast to traditional face recognition task based on single-shot images, image-set based face recognition problem attracts more and more attention recently. For the task of image-set based face recognition, both the gallery and probe samples are image sets, each of which contains many facial images or video frames belonging to one single person. Compared with recognition from single 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. Therefore, how to represent the variations and further discover invariance from them are the key issues of image-set based face recognition [23, 36, 34].

To represent the variations in an image set, a 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 image sets can be computed as the distribution divergence between their GMMs. However, divergence in distribution 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 [23, 36, 6, 18, 13].

To address the above problem, in this paper we propose to learn a discriminative and compact representation for Gaussian distributions and thus measure the dissimilarity of two sets with the distance between the learned representations of pair-wise Gaussian components respectively from either GMM. However, Gaussian distributions lie on a specific Riemannian manifold according to information geometry [1]. Therefore, discriminant analysis methods developed in the Euclidean space cannot be applied directly. We thus propose a novel method of Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG). In this method, by exploring various distances between Gaussians, we derive corresponding provably positive definite probabilistic kernels, which encode the Riemannian geometry of such manifold properly. Then through these kernels, a deliberately devised weighted Kernel Discriminant Analysis is utilized to discriminate the Gaussians from different subjects with their prior probabilities incorporated.

1.1. 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 three categories: linear/affine subspace based methods, nonlinear manifold based methods and statistical

model based methods. They are briefly reviewed as follows.

Linear/affine subspace based methods [38, 23, 6, 18, 39, 14, 16, 8, 10] assume that each image set spans a linear or affine subspace. Among them, the Mutual Subspace Method (MSM) [38] and Discriminant-analysis of Canonical Correlations(DCC) [23] represent each image set as a single linear subspace and compute the principal angles of two linear subspaces for classification. While in [6, 18, 39, 10], 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. Further Chen *et al*. [8] model image sets similarly, but propose a Dual Linear Regression Classification (DLRC) method to perform classification by a regression technique. Grassmann Discriminant Analysis (GDA) [14] and Grassmann Embedding Discriminant Analysis (GEDA) [16] 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.

Since the image sets usually have a relatively large number of images and cover complicated variations of viewpoint, lighting and expressions, 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 in the literature [36, 33, 13, 9]. In Manifold-Manifold Distance (MMD) [36], each image set is assumed to span a nonlinear manifold that can be partitioned into several local linear models and the similarity between manifolds is converted into integrating the distances between pair-wise subspaces. Manifold Discriminant Analysis (MDA) [33] further extends MMD to work in a discriminative feature space rather than the original image space. Cui *et al*. [13] adopt the similar set modeling strategy with MMD, but align the image sets with a generic reference set for more precise local model matching. Chen *et al*. [9] 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.

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 [31, 2] in earlier years represent the image set with some well-studied probability density functions, such as single Gaussian in [31] and Gaussian Mixture model (GMM) in Manifold Density Method (MDM) [2]. 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 [23]. More recently, Covariance Discriminative Learning(CDL) [34] is proposed to model the image set by its natural second-order statistic, *i.e*. covariance matrix, and further derive a Riemannian kernel function to conduct discriminative learning on the manifold spanned by non-singular covariance matrices. While only covariance information is modeled in CDL, Lu *et al*. [28] propose to combine multiple order statistics as features of image sets, and develop a localized multi-kernel metric learning (LMKML) algorithm for classification.

Besides the above three main trends, there also exist some other methods with different set models. For instance, video-based dictionary [11] and joint sparse representation [12] 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) [17] uses deep model to represent image sets. For these works, classification is conducted based on the minimum reconstruction error from the learned class-specific models.

1.2. Overview of our approach

In this paper we propose a new method named Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG) for face recognition with image sets. Fig. 1 shows the overall schematic flowchart of our approach.

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: modeling each image set statistically with GMM and discriminant analysis of the component Gaussians of the GMMs. The first stage 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. In the light of information geometry, Gaussian distributions lie on a Riemannian manifold [1]. But unfortunately, most existing discriminant analysis techniques only work in Euclidean space. This motivates us the idea of Discriminant Analysis on the Riemannian manifold of Gaussian distributions (DARG). Specifically, by exploring various distances between Gaussian distributions, we first derive several kernels to embed the Riemannian manifold of Gaussians into a high-dimensional Hilbert space, which is then further discriminatively reduced to a lower-dimensional subspace.

Our kernel-based solution 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 prob-

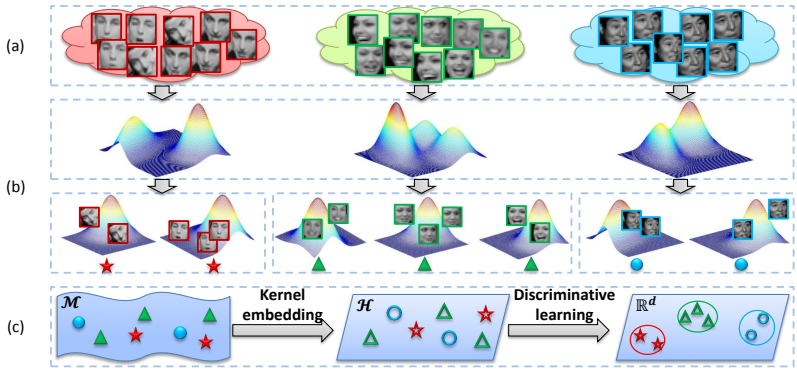

Figure 1: Conceptual illustration of the proposed approach. (a) Training image sets in the gallery. Without loss of generality, we only demonstrate the image sets of three subjects here, with different colors denoting different subjects. (b) Modeling each image set with GMM. Different legends (*i.e.* star, circle and triangle) denote the component Gaussians of different subjects. Each component Gaussian is parameterized by its sample mean and covariance matrix. (c) Discriminant analysis for the Gaussians. By using kernels defined on Riemannian manifold of Gaussian distributions $\mathcal{M}$, the Gaussian components are mapped to a high-dimensional Hilbert space $\mathcal{H}$, which is further discriminatively reduced to a lower-dimensional subspace $\mathbb{R}^d$. As in space $\mathcal{H}$ and $\mathbb{R}^d$ each component Gaussian is transformed to a vector, the legends are changed from solid shapes to hollow ones to reflect this changes.

abilities as sample weights, we devise a weighted Kernel Discriminant Analysis, which aims to maximize the margin between Gaussians from different classes.

After each component Gaussian is transformed to a vector representation in the low-dimensional Euclidean space $\mathbb{R}^d$, classification can be easily performed by exploiting the minimal distance between component Gaussians from either GMM of the gallery and probe image sets.

From above description, we can summarize that the proposed DARG method has three key ingredients: (a) GMM modeling of image sets, (b) Kernels derived from various Gaussian distances, (c) Kernel discriminative learning. These key ingredients are respectively detailed in the following three sections.

2. 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 $X = \{x_1, x_2, ..., x_K\}$, 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:

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

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 particularly 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 determined according to the set size. Considering the nonlinear data distribution in image set, we resort to the local linear model construction algorithm in [33, 35], *i.e.* Hierarchical Divisive Clustering (HDC), which is able to generate the initialization adaptively and efficiently.

3. 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 [14, 16, 34, 20, 15], we derive corresponding positive definite kernels to encode the geometry of mani-

fold 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 [7].

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 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 [26, 24], the distance based on Siegel Group [5], *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.

3.1. 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

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

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,

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

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

3.2. 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

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

3.3. Hellinger Kernel

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

$$HD(g_i, g_j) = \sqrt{1 - BC(g_i, g_j)}, \tag{6}$$

where Bhattacharyya Coefficient (BC) is

$$BC(g_i, g_j) = \exp(-BD(g_i, g_j)). \tag{7}$$

Thus the corresponding Hellinger kernel is

$$K_{HD}(g_i, g_j) = \exp\left(-\frac{HD^2(g_i, g_j)}{2t^2}\right). \tag{8}$$

3.4. Kernel based on Lie Group

Under the framework of information geometry, it is developed in [26] 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 [3] 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:

$$LGD(P_i, P_j) = \| \log(P_i) - \log(P_j) \|_F, \tag{9}$$

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.

$$K_{LGD}(g_i, g_j) = \exp\left(-\frac{LGD^2(P_i, P_j)}{2t^2}\right). \tag{10}$$

3.5. 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

$$MD(\mu_i, \mu_j) = \sqrt{(\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)}, \quad (11)$$

and Log-Euclidean distance (LED) [3] for covariance matrices

$$LED(\Sigma_i, \Sigma_j) = \| \log(\Sigma_i) - \log(\Sigma_j) \|_F. \quad (12)$$

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 a problem in numerical stability. Instead, we derive kernels from the two distances respectively and then linearly combine them to form a valid kernel for Gaussian distributions. Specifically, the kernel based on MD is defined as

$$K_{MD}(\mu_i, \mu_j) = \exp\left(-\frac{MD^2(\mu_i, \mu_j)}{2t^2}\right), \quad (13)$$

while the kernel based on LED is formulated by

$$K_{LED}(\Sigma_i, \Sigma_j) = \exp\left(-\frac{LED^2(\Sigma_i, \Sigma_j)}{2t^2}\right). \quad (14)$$

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

$$\begin{aligned} &K_{MD+LED}(g_i, g_j) \\ &= \gamma_1 K_{MD}(\mu_i, \mu_j) + \gamma_2 K_{LED}(\Sigma_i, \Sigma_j), \end{aligned} \quad (15)$$

where γ_1 and γ_2 are the combination coefficients.

3.6. 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 and analysis of the validity and positive definiteness of these kernels. 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 [29].

4. 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 $g_1, g_2, ..., g_N$, which lie on a Riemannian manifold $\mathcal{M}$. Among them, the Gaussians from the i-th class are denoted as $g_1^i, g_2^i, ..., g_{N_i}^i$, ($\sum_{i=1}^{c} N_i = N$), 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 Sec. 3) measuring 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 $k_j^i = [k(g_j^i, g_1), ..., k(g_j^i, g_N)]^T \in \mathbb{R}^N$.

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 [4].

$$J(\alpha) = \frac{|\alpha^T B \alpha|}{|\alpha^T W \alpha|}, \quad (16)$$

where

$$\begin{aligned} B &= \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T, \\ W &= \sum_{i=1}^{c} \frac{1}{w_i} \sum_{j=1}^{N_i} (k_j^i - m_i)(k_j^i - m_i)^T, \end{aligned} \quad (17)$$

and

$$m_i = \frac{1}{N_i w_i} \sum_{j=1}^{N_i} w_j^i k_j^i, \quad m = \frac{1}{N} \sum_{i=1}^{c} \frac{1}{w_i} \sum_{j=1}^{N_i} w_j^i k_j^i, \quad (18)$$

Note that $w_i = \sum_{j=1}^{N_i} w_j^i$ 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 $\alpha_1, \alpha_2, ..., \alpha_{c-1}$, we obtain $A = [\alpha_1, \alpha_2, ..., \alpha_{c-1}] \in \mathbb{R}^{N \times (c-1)}$. Furthermore, the discriminative projection of a new Gaussian $g_t \in \mathcal{M}$ is given by $z_t = A^T k_t$, where $k_t = [k(g_t, g_1), ..., k(g_t, g_N)]^T \in \mathbb{R}^N$.

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 component Gaussian representations of the test set and those of all the training sets. The Algorithm 1 summarizes the training and testing process of our proposed DARG method.

5. Discussion

While our method represents the image set with a statistical model (*i.e.* GMM) comprising of multiple local models (*i.e.* Gaussian components) and performs set classification in a discriminative way, it bears certain relationship and also has its unique merits comparing with related works in the literature. We highlight them in the following.

Algorithm 1 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 $k_i^{tr} = [k(g_i, g_1), ..., k(g_i, g_N)]^T$ and $k_j^{te} = [k(g_j^{te}, g_1), ..., k(g_j^{te}, g_N)]^T$ by (3) (or any of (5), (8), (10), (15)), $i \in [1, N]$, $j \in [1, M]$;

2: Compute transformation matrix A by maximizing (16);

3: Compute projections $z_1^k, ..., z_{N_k}^k$ of the N_k Gaussians belonging to the k-th image set, $k \in [1, n]$;

4: Compute projections $z_1^{te}, ..., z_M^{te}$ of the M Gaussians belonging to the test set;

5: Compute cosine similarity $cos(z_i^{te}, z_j^k)$ between z_i^{te} and z_j^k;

6: Compute $\hat{k} = \arg\max_k cos(z_i^{te}, z_j^k)$, for all $i \in [1, M]$, $j \in [1, N_k]$;

7: **return** $l^{te} = l_{\hat{k}}$;

Differences from other statistical model based methods [31, 2, 34, 28] . Compared with [31] using single Gaussian and [2] 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 [34] models the image set with its covariance matrix, but ignores the mean information. LMKML [28] combines multiple order statistics as features of image sets, but simply treats both the 2nd order covariance matrix and the 3rd order tensor as vectors which ignores the inherent manifold geometric structure.

Differences from other multi-model based methods [36, 33, 9]. MMD [36], MDA [33] and SANS [9] all employ multi-model set representation, but they obtain the local models by a hard partition that neglects the probabilistic distribution of the set data, which is mainly encoded with the Gaussian distribution in this work. 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 [34, 28]. Again, SANS is non-discriminative.

Differences from other discriminative methods [33, 14, 16]. MDA [33] uses Euclidean distance between mean vectors to measure the distance between image sets, and GDA [14]/GEDA [16] utilize the subspace modeling without considering mean information. In contrast, our approach endeavors to measure the distribution distinction of local Gaussian models with distance incorporating both mean and covariance information.

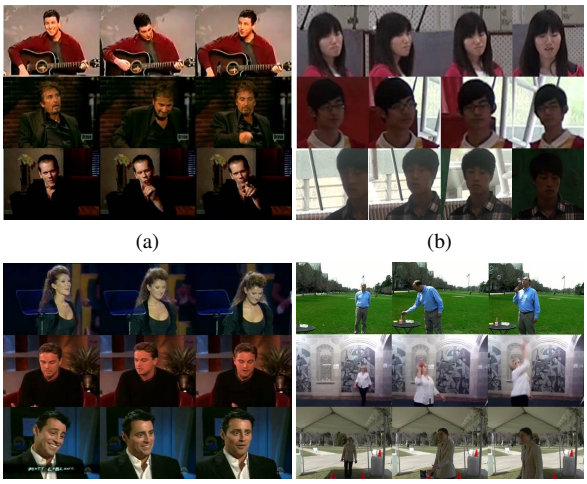

Figure 2: Some examples of the datasets. (a) YTC (b) COX (c) YTF (d) PaSC

In summary, our contributions mainly lie in three folds: 1) we propose a new discriminative framework for learning with Gaussians on Riemannian manifold; 2) we derive a series of kernels for Gaussians with proved positive definiteness; 3) a weighted extension of KDA is devised to classify Gaussians with component weights incorporated.

6. Experiments

6.1. Databases description and settings

We used four most challenging and largest datasets: YouTube Celebrities (YTC) [22], COX [19], YouTube Face DB (YTF) [37] and Point-and-Shoot Challenge (PaSC) [30]. In our experiments, 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 three clips for training and six for testing in each of the ten folds. This enables the whole testing sets in our experiment to cover all of the 1,910 clips in the database, which is similar with protocol in [33, 34, 13, 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 [37]. 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 split contains 250 'same' pairs and 250 'not-same' pairs. PaSC consists of 2,802 videos of 265 people, and half of these videos are captured by controlled video camera, the rest are captured by hand held video camera. It has a total of 280 sets for training and verification experiments were conducted using control or handheld videos as target and query respectively.

For all the four datasets, a cascaded face detector [32] was used to detect faces in the frames, and then the faces were resized to 20×20 on YTC, 32×40 on COX, 24×40 on YTF and 64×80 on PaSC as previous works. In order to eliminate lighting effects, histogram equalization was implemented for the gray features of faces in the four datasets above. Note that for PaSC, we further extracted the dense SIFT feature [25] from the pre-processed faces.

6.2. Comparison results and analysis

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

(1) Linear/affine subspace based methods:

MSM [38], DCC [23], AHISD [6], CHISD [6], SANP [18], GDA [14] and GEDA [16].

(2) Nonlinear manifold based methods:

MMD [36] and MDA [33].

(3) Statistical model based methods:

SGM [31], MDM [2] and CDL [34].

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 model(s). 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 [36, 33]. For MSM, AHISD, CHISD and SANP, 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 [34] on YTC, COX and PaSC. Note that on YTF we used a kernel version of SILD [21] rather than KDA in CDL and our approach because we cannot get the exact label , but only know whether an image pair belong to the same subject on YTF. In our method, for kernel based on MD and LED, we fixed the fusing coefficient γ_1 as 1, and γ_2 was searched in the range of [0.5,1,2].

For face identification task, Tab. 1 shows the average recognition accuracy over multiple-fold trials on YTC and COX. For our proposed framework, we tested the performance of kernels induced from different distances between Gaussians, which is denoted by "DARG-" in the table.

Method \ Dataset	YTC	COX					
		COX-11	COX-12	COX-23	COX-21	COX-31	COX-32
MSM [38]	61.14	45.53	21.47	10.96	39.83	19.36	9.50
AHISD [6]	63.69	57.54	37.99	18.57	47.91	34.91	18.79
CHISD [6]	66.46	56.87	30.10	14.80	44.37	26.44	13.68
DCC [23]	66.81	62.51	66.11	50.59	56.12	63.84	45.21
GDA [14]	65.91	72.26	80.70	74.36	71.44	81.99	77.57
GEDA [16]	66.83	76.73	83.80	76.59	72.56	82.84	79.99
MMD [36]	65.30	38.29	30.34	15.24	34.86	22.21	11.44
MDA [33]	66.98	65.82	63.01	36.17	55.46	43.23	29.70
SGM [31]	52.00	26.74	14.32	12.39	26.03	19.21	10.50
MDM [2]	62.12	30.70	24.98	14.30	28.90	31.72	19.30
CDL [34]	69.70	78.37	85.25	79.74	75.59	85.83	81.87
DARG-KLD	72.21	71.93	80.11	73.65	70.87	81.03	76.99
DARG-BD	72.49	77.55	85.02	79.11	76.01	85.13	82.12
DARG-HD	68.33	74.98	82.34	75.11	70.96	81.34	78.08
DARG-LGD	68.72	76.74	84.99	78.02	72.93	83.88	81.54
DARG-MD+LED	**77.09**	**83.71**	**90.13**	**85.08**	**81.96**	**89.99**	**88.35**

Table 1: 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	MSM [38]	AHISD [6]	CHISD [6]	SANP [18]	MMD [36]	CDL [34]	DARG-MD+LED
Result	62.54	66.50	66.24	63.74	64.96	69.74	**73.01**

Table 2: Comparisons on YTF. The performance is evaluated by the area under ROC curve (AUC) in this table.

Setting \ Method	DCC [23]	GDA [14]	GEDA [16]	MDA [33]	CDL [34]	DARG-MD+LED
Control	4.44	13.45	10.84	5.75	13.87	**18.73**
Handheld	4.10	10.80	8.52	2.88	12.40	**18.32**

Table 3: Comparisons on PaSC. Note that the verification rates (%) at a false accept rate (FAR) of 0.01 on PaSC is reported in this table. Here, " Control " denotes the experiment using the control videos as target and query, while " Handheld " implies that the handheld videos are used as target and query.

For face verification task, Tab. 2 shows the area under ROC curve (AUC) on YTF. The comparisons on PaSC are shown in Tab. 3 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) As shown in Tab. 1, among the non-discriminant methods, compared with the single modeling methods MSM, AHISD, CHISD, SGM, most of 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) In the discriminant methods, kernel-based methods GDA, GEDA and CDL including our proposed method yield better results than DCC and MDA. This is because DCC and MDA learn the discriminant metrics in Euclidean space, whereas most of them classify the sets in non-Euclidean spaces. In contrast, these kernel-based 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, GEDA and CDL, our method achieves the best performance. This is because that they only utilize the relatively weak information of set variations while our method attempt to model the data distribution and jointly fuse both mean and covariance information.

(4) Among four databases, all methods have relatively poor performances on PaSC, due to the low-quality and large motion blur of face region images on PaSC. Besides, we did not exploit external data to expand the training set, which also makes recognition more difficult on PaSC.

(5) As shown in Tab. 1, the kernel based on MD and 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.

6.3. Comparison of computation time

In Tab. 4, 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-MD+LED as an example 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.

6.4. Comparison of different Gaussian component numbers

Fig. 3 shows how the identification rate changes for our proposed DARG method with kernel based on MD and LED when using different average numbers of Gaussian components on YTC. Note that the number of Gaussian components is different for each image set. Hence, the average number of Gaussian components is not necessarily integer value. The results show favorably stability within a proper range of Gaussian numbers. We get the best result with an average of about 7 Gaussian components in GMM, which balances the accuracy of image set modeling and local Gaussian statistics estimating.

Process \ Method	AHISD [6]	DCC [23]	GDA [14]	MDA [33]	CDL [34]	DARG-MD+LED
Training	N/A	44.85	3.86	11.34	4.15	114.70
Testing	0.28	0.32	0.42	0.31	0.32	0.80

Table 4: Computation time (seconds) of different methods on YTC for training and testing (classification of one image set).

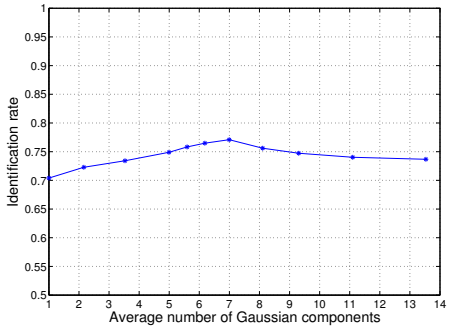

Figure 3: Comparison of different average Gaussian component numbers using " DARG-MD+LED " on YTC for face identification task.

7. Conclusion

This paper contributes a discriminant analysis on the Riemannian manifold of Gaussian distributions for face recognition with image sets. Our method differs from tradition methods in learning for Gaussian distributions on manifold rather than vectors in Euclidean space. We utilize GMM to represent each image set by a number of Gaussian components with prior probabilities. Then, a series of simple but valid probabilistic kernels were derived from various distances between Gaussians. Through these kernels, a weighted Kernel Discriminant Analysis technique was devised to maximize the margin between Gaussians from different classes. The experiments have demonstrated the superiority of our proposed approach over state-of-the-art methods. For future work, we are studying more probabilistic kernels for Gaussians and more conventional learning methods will be extended to Riemannian manifold of Gaussian distributions.

Acknowledgements

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

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. Translations of Mathematical monographs. Oxford University Press, 2000.

[2] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[3] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 2006.

[4] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 2000.

[5] M. Calvo and J. M. Oller. A distance between multivariate normal distributions based in an embedding into the siegel group. *Journal of multivariate analysis*, 1990.

[6] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[7] A. B. Chan, N. Vasconcelos, and P. J. Moreno. A family of probabilistic kernels based on information divergence. *Univ. California, San Diego, CA, Tech. Rep. SVCL-TR-2004-1*, 2004.

[8] L. Chen. Dual linear regression based classification for face cluster recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[10] S. Chen, A. Wiliem, C. Sanderson, and B. C. Lovell. Matching image sets via adaptive multi convex hull. *arXiv preprint arXiv:1403.0320*, 2014.

[11] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *European Conference on Computer Vision (ECCV)*, 2012.

[12] Y.-C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips. Video-based face recognition via joint sparse representation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG).*, 2013.

[13] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. *IEEE Computer Society on Computer Vision and Pattern Recognition (CVPR).*, 2012.

[14] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning (ICML)*, 2008.

[15] M. T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li. Expanding the family of grassmannian kernels: An embedding perspective. In *European Conference on Computer Vision (ECCV)*, 2014.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[17] M. Hayat, M. Bennamoun, and S. An. Learning non-linear reconstruction models for image set classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[18] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] Z. Huang, R. Wang, S. Shan, and X. Chen. Learning euclidean-to-riemannian metric for point-to-set classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. T. Harandi. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In *IEEE Computer Society on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[21] M. Kan, S. Shan, D. Xu, and X. Chen. Side-information based linear discriminant analysis for face recognition. *British Machine Vision Conference (BMVC)*, 2011.

[22] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[23] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2007.

[24] P. Li, Q. Wang, and L. Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[25] C. Liu, J. yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: dense correspondence across difference scenes. In *European Conference on Computer Vision (ECCV)*, 2008.

[26] M. Lovrić, M. Min-Oo, and E. A. Ruh. Multivariate normal distributions parametrized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 2000.

[27] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *European Conference on Computer Vision (ECCV)*, 2014.

[28] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2013.

[29] P. J. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[30] B. Ross, J. Phillips, D. Bolme, B. Draper, G. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. Bowyer, P. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2013.

[31] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *European Conference on Computer Vision (ECCV)*. 2002.

[32] P. Viola and M. J. Jones. Robust real-time face detection. In *International Journal of Computer Vision (IJCV)*, 2008.

[33] R. Wang and X. Chen. Manifold discriminant analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[34] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[35] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao. Maximal linear embedding for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(9):1776–1792, 2011.

[36] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[37] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[38] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG).*, 1998.

[39] M. Yang, P. Zhu, L. V. Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. *IEEE Computer Society on Automatic Face and Gesture Recognition (FG).*, 2013.