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

For face recognition with image sets, modeling image sets by statistical models [2, 3, 4] have achieved great performance due to their capacity in characterizing the set data distribution more flexibly and faithfully. Among many others, Gaussian Mixture Model (GMM) can precisely capture the data variations with a multi-modal density. 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, especially when the gallery and probe sets have weak statistical correlations.

To address the above problem, in this paper we propose to learn a discriminative and compact 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. Since Gaussian distributions lie on a specific Riemannian manifold according to information geometry [1], 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)**. 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 (KDA) is utilized to discriminate the Gaussians from different subjects with their prior probabilities incorporated.

Kernels for Gaussian distributions. Among the derived kernels, here we take the best performing *kernel based on Mahalanobis distance and Log-Euclidean distance* as an example.

We 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, resides on the SPD manifold. Formally, given two Gaussian distributions $g_i = (\mu_i, \Sigma_i)$ and $g_j = (\mu_j, \Sigma_j)$, we choose Mahalanobis distance (MD) for means

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

and Log-Euclidean distance (LED) for covariance matrices

$$LED(\Sigma_i, \Sigma_j) = \|\log(\Sigma_i) - \log(\Sigma_j)\|_F. \quad (2)$$

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

$$K_{MD}(\mu_i, \mu_j) = \exp\left(-\frac{MD^2(\mu_i, \mu_j)}{2t^2}\right), \quad (3)$$

while the kernel based on LED is formulated by

$$K_{LED}(\Sigma_i, \Sigma_j) = \exp\left(-\frac{LED^2(\Sigma_i, \Sigma_j)}{2t^2}\right). \quad (4)$$

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

$$K_{MD+LED}(g_i, g_j) = \gamma_1 K_{MD}(\mu_i, \mu_j) + \gamma_2 K_{LED}(\Sigma_i, \Sigma_j), \quad (5)$$

where γ_1 and γ_2 are the combination coefficients.

Weighted kernel discriminative learning. Formally, suppose we have n image sets belonging to c classes for training. From their GMM models, we collect all the N Gaussian components $g_1, g_2, ..., g_N$, 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 derived kernels) 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 $k_j^i = [k(g_j^i, g_1), ..., k(g_j^i, g_N)]^T \in \mathbb{R}^N$.

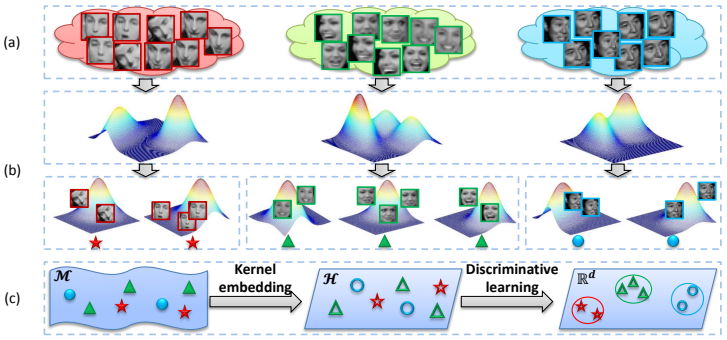

Figure 1: Conceptual illustration of the proposed approach. (a) Training image sets in the gallery. Different colors denote different subjects. (b) Modeling each image set with GMM. Different legends (*i.e.* star, circle and triangle) denote the component Gaussians of different subjects. (c) Discriminant 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 $\mathbb{R}^d$.

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 $J(\alpha)$.

$$J(\alpha) = \frac{|\alpha^T B \alpha|}{|\alpha^T W \alpha|}, \quad (6)$$

where

$$B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T, \quad W = \sum_{i=1}^{c} \frac{1}{w_i} \sum_{j=1}^{N_i} (k_j^i - m_i)(k_j^i - m_i)^T, \quad (7)$$

and

$$m_i = \frac{1}{N_i w_i} \sum_{j=1}^{N_i} w_j^i k_j^i, \quad m = \frac{1}{N} \sum_{i=1}^{c} \frac{1}{w_i} \sum_{j=1}^{N_i} w_j^i k_j^i, \quad w_i = \sum_{j=1}^{N_i} w_j^i \quad (8)$$

Then the optimization problem can be reduced to solving a generalized eigenvalue problem: $B\alpha = \lambda W \alpha$. After solving its $c-1$ leading eigenvectors $\alpha_1, \alpha_2, ..., \alpha_{c-1}$, 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 these discriminative representations of the test set and those of all the training sets.

[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 *Proc. CVPR*, 2005.

[3] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *Proc. ICCV*, 2013.

[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. CVPR*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.