Discriminative Covariance Oriented Representation Learning for Face Recognition with Image Sets

Wen Wang[1,2], Ruiping Wang[1,2,3], Shiguang Shan[1,2,3], Xilin Chen[1,2,3]
[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
[3]Cooperative Medianet Innovation Center, China
wen.wang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Abstract

For face recognition with image sets, while most existing works mainly focus on building robust set models with hand-crafted feature, it remains a research gap to learn better image representations which can closely match the subsequent image set modeling and classification. Taking sample covariance matrix as set model in the light of its recent promising success, we present a Discriminative Covariance oriented Representation Learning (DCRL) framework to bridge the above gap. The framework constructs a feature learning network (e.g. a CNN) to project the face images into a target representation space, and the network is trained towards the goal that the set covariance matrix calculated in the target space has maximum discriminative ability. To encode the discriminative ability of set covariance matrices, we elaborately design two different loss functions, which respectively lead to two different representation learning schemes, i.e., the Graph Embedding scheme and the Softmax Regression scheme. Both schemes optimize the whole network containing both image representation mapping and set model classification in a joint learning manner. The proposed method is extensively validated on three challenging and large scale databases for the task of face recognition with image sets, i.e., YouTube Celebrities, YouTube Face DB and Point-and-Shoot Challenge.

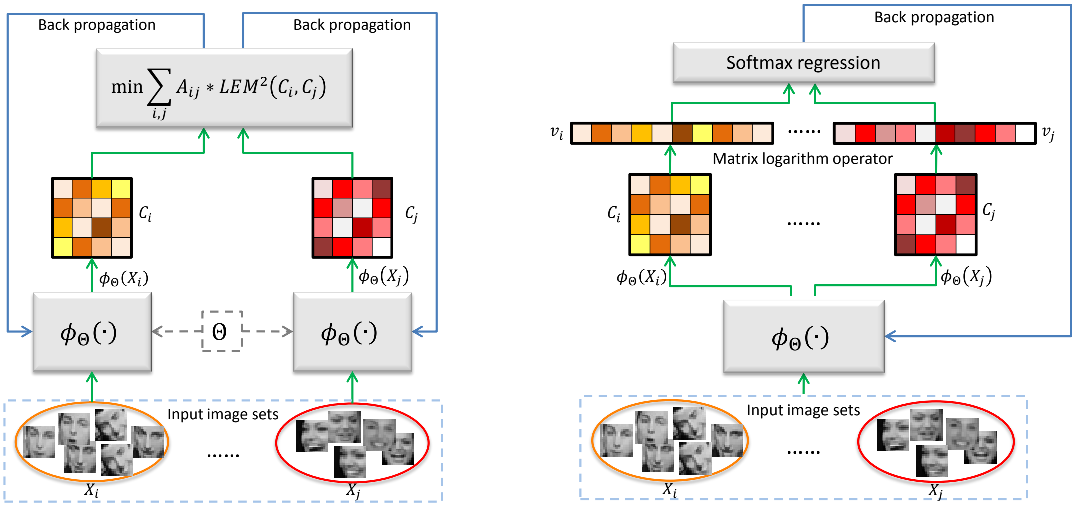
1. Introduction

With the rapid development of multimedia technologies, increasing interest has been attracted on face recognition with image sets in real-world applications such as video surveillance, classification with images from multi-view camera networks or online face books, etc. Different from traditional single-shot image based face recognition task, for face recognition with image sets, both the gallery and probe samples are image sets, each of which contains a large number of images belonging to the same class. Nevertheless, large intra-class variations are usually contained in the image sets due to changes in pose, expression, illumination and other factors, thus how to represent these variations using robust set models and further discover invariance from them is considered a key issue.

Among the existing set models, set covariance matrix has gained a promising success [39, 37, 26, 15]. As a natural second-order statistics, set covariance matrix characterizes each image set with different number of samples as a fixed-dimensional and comparable representation compactly and effectively, thus is chosen to represent image set in this paper. Since non-singular covariance matrices lie on a specific Symmetric Positive Definite (SPD) Riemannian manifold, their distances are usually measured with the Riemannian metrics, e.g., the Log-Euclidean metric (LEM) [2]. Alternatively, several works [5, 35, 36] propose to apply the matrix logarithm operator to convert a covariance matrix from the SPD manifold to a vector in the tangent space at the identity matrix where Euclidean geometry applies.

For face recognition with image sets, most existing works, e.g., [39, 26, 24, 15, 41, 13], etc., mainly focus on building robust set models with hand-crafted feature. Moreover, motivated by the proven success of deep network, class-specific neural networks are trained in [17, 25] to extract image representations in different image sets respectively, which learn the within-class structure merely using the class-specific mechanism yet without an explicit set representation. Thus, we argue that it still remains a research gap to learn more desirable image representations which can closely match the subsequent image set modeling and classification.

To address this issue, we present a Discriminative Covariance oriented Representation Learning (DCRL) method which aims to learn a feature learning network, such as



(a) Graph Embedding scheme

(b) Softmax Regression scheme

Figure 1: Conceptual illustration. The basic idea is to find a shared mapping $\phi_\Theta(\cdot)$ which projects the images of different sets into a target feature space such that the set model (i.e., set covariance matrix) calculated in this target space has maximum discriminative ability. More precisely, since $\phi_\Theta(\cdot)$ is a CNN which is parameterized by Θ, we seek to find a value of Θ to meet such optimization objective. Note that the green and blue arrows denote feeding forward and back propagation respectively. (a) Given a pair of image sets (X_i, X_j), the **Graph Embedding** scheme optimizes Θ through minimizing the Log-Euclidean metric (LEM) weighted by an affinity matrix A , whose entries correspond to pairs of set covariance matrices calculated in the target image feature space. (b) For the **Softmax Regression** scheme, we seek to find a Θ which ensures that log-covariance vectors corresponding to different sets can be classified by a softmax regression machine.

Convolutional Neural Network (CNN), to project the images into a target feature space such that the set covariance matrices calculated in this target space have maximum discriminative ability. Under the DCRL framework, we propose two different representation learning schemes, i.e., the Graph Embedding scheme and the Softmax Regression scheme, based on two elaborately designed loss functions respectively. In particular, the Graph Embedding scheme, inspired by the general graph embedding framework for dimensionality reduction [44], takes pairs of image sets as training input and tries to minimize the LEM weighted by an affinity matrix whose entries correspond to pairs of set covariance matrices calculated in the target space. While the Softmax Regression scheme has its initial motivation from the multi-class classification layer in conventional fine-tuned neural networks [3], here we conduct softmax regression on log-covariance vectors which are ensured to follow the Euclidean geometry as mentioned above. Note that for both representation learning schemes, the two stages contained in the whole network, i.e., image feature learning and set model classification, are jointly optimized with back-propagation. Fig. 1 shows the overall schematic flowchart of the two representation learning schemes.

2. Related Work

In this section, we give a review of recent methods for face recognition with image sets. Generally speaking, most of these methods tend to represent the image set by some model and then study the dissimilarity measure between the models of different sets. Regarding the type of image set modeling, the existing methods can be categorized into four general classes, which respectively characterize image set with linear/affine subspace, nonlinear manifold, face dictionary and statistical models.

Linear/affine subspaces are used to model the image set for their simplicity and efficiency. For instance, methods of [43, 23] measure the image set similarity by the principal angles of two linear subspaces. Then following a pioneering work of [6], a series of methods [18, 45, 9, 7, 13] are proposed to approximate each image set with one or multiple convex geometric regions (affine or convex hull). Additionally, methods of [14, 16] model image sets as points (i.e. linear subspaces) on Grassmann manifold and define Grassmann kernels to conduct discriminative learning on the manifold.

Considering the complicated variations contained in the image set, some methods model the image set by more sophisticated nonlinear manifold. Among them, some methods attempt to partition the image set manifold into several local linear models. For instance, in [40], Manifold-Manifold Distance (MMD) is computed by integrating the distances between pair-wise local linear models. Manifold Discriminant Analysis (MDA) [38] further extends MMD to solve the supervised between-manifold distance. Cui et al. [12] present an image set alignment method for more precise local model matching. In [8], the nearest pair of local linear subspaces is searched by joint sparse approximation and then the image set dissimilarity is measured by distance between such nearest pair.

Recently, face dictionary is explored as a representation for image set and the works in face dictionary and sparse representation are extended from still images to videos. For example, Chen et al. [10, 11] present video-based dictionary learning methods, and then classification is conducted based on integrating the minimum residuals from the learned models. Further, a simultaneous feature and dictionary learning (SFDL) method is proposed in [24] so that discriminative information can be jointly exploited.

In the literature, statistical models have also been employed for image set modeling due to their capacity in characterizing the set data distribution. Some methods represent each image set with some parametric distribution, such as single Gaussian [31] or Gaussian Mixture Model (GMM) [1], and use the Kullback-Leibler Divergence (KLD) as the dissimilarity measure. Wang et al. [39] present a Covariance Discriminative Learning (CDL) method to model the image set by its covariance matrix, and further derive a Riemannian kernel to conduct discriminative learning on the Symmetric Positive Definite (SPD) manifold. Lu et al. [26] propose to combine multiple order statistics as features of image sets, and develop a localized multi-kernel metric learning (LMKML) algorithm for classification. Harandi et al. [15] attempt to learn an orthonormal projection from the high-dimensional SPD manifold to a low-dimensional, more discriminative one. Wang et al. [41] propose to model the image set with a GMM and derive a series of kernels for Gaussians to conduct Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG).

In addition to the above methods, motivated by the success of deep learning, an Adaptive Deep Network Template (ADNT) [17] is presented to learn a deep reconstruction network for each class, thereby classification is conducted based on minimum reconstruction error from the learned class-specific models. To further take discriminative information into account, Lu et al. [25] propose a Multi-Manifold Deep Metric Learning (MMDML) method. Despite the relatively good performance of these methods, they also suffer from some limitations as the class-specific models need to be passed through by images in each test image set and thereby bring growingly high complexity with large number of classes in the gallery. More recently, Zhang et al. [47] present a deep learning based method to jointly conduct face representation adaptation and clustering in videos.

3. Basic Idea

In this paper, we propose a new method named Discriminative Covariance oriented Representation Learning (DCRL) for face recognition with image sets. In the following, we start with illustrating the basic idea of the proposed method.

As mentioned above, we aim to learn more discriminative image representations which are consistent with image set modeling as well as subsequent classification. Specifically, we learn a feature learning network to project the images into a target feature space such that the set covariance matrices calculated in this target space can be better discriminated. Therefore, there are two pivotal challenges: 1) constructing the feature learning network structure, 2) designing loss functions to jointly optimize the two stages of image feature learning and set model classification.

As for the first challenge, since Convolutional Neural Network (CNN) has been extensively studied as feature extractor in the problem of single-shot image based face recognition, we could refer to the structure of some successful examples, such as DeepID [33, 32, 34], FaceNet [30] and VGG-Face [28], etc. By passing all the images through the feature learning network, we can project them into a target feature space and subsequently calculate the sample covariance matrix for each image set. Thereby each image set containing variable number of face images are merged as a fixed-size unified representation which is compact, efficient and more desirably, discriminative. Then how to define the optimization objective and to train this network remains to be tackled as the second challenge.

To cope with the second challenge, we define the loss function to characterize the training goal of set model classification. Specifically, we design two loss functions exactly in allusion to the two representative strategies for discriminative learning, i.e., learning discriminative metric and learning classification function. Thus minimizing these loss functions respectively leads to two different representation learning schemes, that is, the Graph Embedding scheme and Softmax Regression scheme. For both of the two schemes, the architecture is optimized as a whole and the parameters for the feature learning network and the set model classification are trained jointly through back-propagation.

4. Discriminative Covariance oriented Representation Learning (DCRL)

Suppose there are n training image sets $\{X_i\}_{i=1}^n$ with corresponding class labels $\{y_i\}_{i=1}^n$, where $y_i \in [1, m]$. Among them, $X_i \in \mathbb{R}^{N_i \times D}$ contains D-dimensional original feature vectors of N_i images. The original feature vector of a face image is the concatenation of intensity of all pixels.

4.1. Image Feature Extraction

Due to the property of compact structure and thereby relatively low computation complexity, we choose a feature learning network which has similar structure with that in [32]. It should be noted that it is natural to extend to other network designs for boosting performance, which allows our DCRL framework to benefit from the progress of state-of-the-art deep models for single-shot image based face recognition.

To perform image feature extraction, all the images in each input image set are passed through the feature learning network. We denote $\phi_\Theta(\cdot)$ as the mapping from the original image feature to the target feature space, where Θ is the parameters of the network. Therefore, for an image set X_i, its projection h_i in the target feature space is denoted as

$$h_i = \phi_\Theta(X_i), \tag{1}$$

where h_i consists of N_i row feature vectors, which are computed by passing X_i through the feature learning network image by image.

4.2. Set Covariance Modeling

Having extracted the image features in each image set, it is desirable to merge them to a compact, fixed-size and discriminative set representation, which facilitates performance improvement while saving the time and space cost. A lot of set models have been extensively researched in recent years and among them, the set covariance model has achieved promising success [39, 37, 26, 15] due to its favorable properties of integrating the variation information of an image set with any number of image samples into a comparable representation.

We thus choose to model an image set X_i with the set covariance matrix C_i of the extracted image features h_i, and C_i is calculated as follows.

$$C_i \quad = h_i^T J_{N_i} h_i \tag{2}$$

where the centering of h_i is performed by a constant matrix J_{N_i}.

$$J_{N_i} = \frac{1}{N_i - 1}(I_{N_i} - \frac{1}{N_i} 1_{N_i}) \tag{3}$$

is determined by N_i, which is the number of images in the i-th image set. I_{N_i} and 1_{N_i} denote the identity matrix and the matrix of all ones, whose size are $N_i \times N_i$. Besides, to guarantee the non-singularity, a standard regularizer is imposed by adding a small positive perturbation to the covariance matrix C_i.

4.3. Network Optimization

In this paper, we present two different representation learning schemes, i.e., the Graph Embedding scheme and Softmax Regression scheme. Next, we introduce the two schemes and explain how to jointly optimize the two stages of representation mapping and set model classification.

4.3.1 Graph Embedding Scheme

The Graph Embedding scheme is trained on pairs of image sets. Given pairs of set covariance matrices computed by the extracted image features, we expect that their Log-Euclidean metric (LEM) to be small for pairs from the same class and large for those from different classes. Inspired by the graph embedding framework [44], we define an affinity matrix to encode the data structure and semantic relationship in the original image space. Then, minimizing the distance between set covariance matrices weighted by the affinity matrix is thus defined as the optimization objective. The Graph Embedding scheme is illustrated as Fig. 1a.

Firstly, we measure the distance between two set covariance matrices C_i and C_j calculated in the target space. S-ince the non-singular covariance matrices lie on the Symmetric Positive Definite (SPD) manifold, their distance can be measured with the Log-Euclidean metric (LEM) [2]:

$$LEM(C_i, C_j) = \|log(C_i) - log(C_j)\|_F, \tag{4}$$

where $log(\cdot)$ is the ordinary matrix logarithm operator. Let $C_i = U_i \Sigma_i U_i^T$ be the eigen-decomposition of C_i, its log-covariance matrix is formulated as:

$$log(C_i) = U_i log(\Sigma_i) U_i^T. \tag{5}$$

Secondly, to drive the set covariance matrices to be discriminative, we utilize a graph embedding framework with the graph defined by a real-valued symmetric affinity matrix $A \in \mathbb{R}^{n \times n}$, where each element A_{ij} measures the relationship between two sets X_i and X_j. While A can be constructed via several strategies, here we simply define it as follows:

$$A_{ij} = \begin{cases} d_{ij}, & \text{if } X_i \in N_w(X_j) \text{ or } X_j \in N_w(X_i) \\ -d_{ij}, & \text{if } X_i \in N_b(X_j) \text{ or } X_j \in N_b(X_i) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $d_{ij} = \exp(-LEM^2(C_i, C_j)/\sigma^2)$. $N_w(X_i)$ consists of K_w nearest neighbors of X_i that share the same label as y_i, while $N_b(X_i)$ contains K_b nearest neighbors of X_i belonging to different classes from X_i.

Finally, we define the optimization objective in the following:

$$\Theta = \arg\min_\Theta J(\Theta) \tag{7}$$

where

$$J(\Theta) = \frac{1}{4} \sum_{i,j} A_{ij} LEM^2(C_i, C_j). \tag{8}$$

4.3.2 Softmax Regression Scheme

For the Softmax Regression scheme, we refer to the conventional strategy for fine-tuning a neural networks, i.e., wiring a softmax regression machine to the output layer. Nevertheless, it is non-trivial as the set covariance matrices lie on a specific SPD Riemannian manifold. In view of the fact that the SPD matrices can be transformed to log-covariance vectors in the tangent space at the identity matrix, we develop an optimization objective such that log-covariance vectors corresponding to different image sets can be classified by a softmax regression machine. Fig. 1b gives an illustration of its conceptual architecture.

We start with transforming C_i to a log-covariance vector v_i via vectorizing the log-covariance matrix $\log(C_i)$ in Equ. 5. Since the log-covariance vectors follow the Euclidean geometry, we present an optimization objective that log-covariance vectors calculated in the target space can be classified by a softmax regression machine. Formally, the optimization objective is of the form:

$$\Theta' = \arg\max_{\Theta'} J(\Theta') \tag{9}$$

where

$$J(\Theta') = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} 1\{y_i = j\}\log(o_{ij}) \tag{10}$$

is the logistic regression cost function and $\Theta' = \{\Theta, W, b\}$. n is the number of training image sets and m denotes the number of classes. $1\{\cdot\}$ is the indicator function, that is, $1\{true\} = 1, 1\{false\} = 0$. o_{ij} is the estimated probability of v_i for the i-th class, i.e., $o_{ij} = P(y_i = j|v_i; W, b)$.

$$o_{ij} = \frac{e^{W_j v_i + b_j}}{\sum_l e^{W_l v_i + b_l}} \tag{11}$$

where W_j denotes the j-th row of W, and b_j is the j-th element of b.

Due to space limitation, the detailed derivation of the gradients in Sec. 4.3.1 and 4.3.2 is given in our supplementary materials. Accordingly, the optimization problems defined in Equ. 7 and 9 are solved with Stochastic Gradient Descent.

4.4. Classification

In the testing stage, given a test image set, we first pass the images through the learned image feature learning network $\phi_{\Theta^*}(\cdot)$. Its set covariance matrix is obtained based on the learned image features, hence the problem is casted as the discrimination of set covariance matrices. The set covariance matrices can be naturally classified by the corresponding Riemannian metrics, e.g., LEM used in this paper. Alternatively they can also serve as the extracted image set representation on which existing covariance discriminative learning methods are performed to further enhance performance.

4.5. Implementation Details

For network training, we perform the data augmentation to generate more image sets by random sampling in existing image sets and to enlarge each image set via randomly cropping, flip and color shuffle. Through these strategies, we can further enrich the variation to the training process, which leads to better robustness against noises in the data.

The structure of the feature learning network is similar with the one in [32] which takes the RGB image of size 55×47 as input. It consists of four convolutional layers and a fully-connected layer. The first three convolutional layers are followed by max-pooling. By passing through the feature learning network, a 160-dimensional feature vector is extracted for each sample image. For more details of network structure, please refer to the supplementary materials.

The two proposed schemes are implemented in Titan-X GPU with 12GB memory using a modified Caffe deep learning toolbox [20]. The learning rate is initially set to 0.01 and reduced following polynomial curve with gamma value equal to 0.5. The momentum is set as 0.9 and the weight decay is set as 0.0005.[1]

4.6. Discussion

4.6.1 Differences from Related Methods

Though both the SPD Manifold Learning (SPDML) method [15] and our proposed Graph Embedding scheme aim to learn discriminative covariance models, they differ essentially in the following aspects: 1) SPDML is an SPD manifold learning method, while our method aims to learn image representations to drive the corresponding set model to be discriminative. 2) In SPDML, two types of metric, i.e., AIM and Stein metric, are chosen due to the affine invariant constraint. While this extra constraint is not necessarily imposed in our DCRL framework, our method is open to a wider range of SPD measures. Here we choose the more efficient LEM which favorably overcomes the computational limitation of AIM and the failure of Stein metric in defining the true manifold geodesics.

One of the main innovations of our method is that we develop a deep architecture for image set classification. In the literature, there are only two similar works which explore the application of deep learning technique in image set classification, i.e., ADNT [17] and MMDML [25]. However, their ideas are totally different from ours in that both of them expect a deep structure itself to model the variation in each image set implicitly, which is nevertheless hard to be captured. On the contrary, our method utilizes the set

[1] The source code is available at http://vipl.ict.ac.cn/resources/codes.

covariance model to characterize the variation within each image set compactly and effectively. Besides, both ADNT and MMDML learn one deep structure for one class, which brings a large amount of parameters to learn and makes the training as well as the testing process reduplicative. On the contrary, our method trains a common mapping, which not only avoids redundant parameters but also explores the implicit common pattern in different sets.

4.6.2 Differences between the Two Proposed Schemes

While under the framework of DCRL, we give two representation learning schemes, i.e., the Graph Embedding scheme and Softmax Regression scheme, it should be noticed that both of them are efficient and different from the following perspectives.

1) The usage of discriminative information. The Graph Embedding scheme works through learning discriminative metric, while the Softmax Regression scheme back propagates the discriminative information by using a softmax classification function. This makes them have different focuses which complement and reinforce the whole framework of DCRL.

2) The form of the objective functions. The Graph Embedding scheme exploits an affinity matrix to make the target space reserve the data structure and semantic relationships which are implicit in raw image space. In contrast, the Softmax Regression scheme directly connects the target space with the class label space by assigning each data point a probability of each class.

5. Experiments

We evaluate our proposed method on three challenging and large datasets: YouTube Celebrities (YTC) [22], YouTube Face DB (YTF) [42] and Point-and-Shoot Challenge (PaSC) [29]. Among them, the YTC dataset is used to evaluate our performance in the task of face identification with image sets, and the YTF and PaSC datasets are used to evaluate our performance in the task of face verification with image sets. In our experiments, the protocol and performance measure all follow the original literature.

5.1. Databases Description and Settings

The YTC dataset [22] contains 1,910 YouTube videos of 47 subjects. Most videos are low resolution which leads to noisy and low-quality image frames. The number of frames for these videos varies from 8 to 400. We conducted ten-fold cross validation experiments and randomly selected three clips for training and six for testing for each subject in each of the ten folds, which is similar with protocol in [38, 39, 12, 24].

The YTF dataset [42] contains 3,425 videos of 1,595 subjects. There are large variations in pose, illumination, and expression, and resolution in these videos. We followed the same settings with benchmark tests in [42]. 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. Thus we conducted ten-fold cross validation on these splits.

The PaSC dataset [29] 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.

5.2. Comparison Methods and Parameter Settings

We first compared our performance with several state-of-the-art image set classification methods including: Discriminant Canonical Correlation analysis (DCC) [23], Manifold-to-Manifold Distance (MMD) [40], Manifold Discriminant Analysis (MDA) [38], Affine Hull based Image Set Distance (AHISD) [6], Convex Hull based Image Set Distance (CHISD) [6], Covariance Discriminative Learning (CDL) [39], Localized Multi-kernel Metric Learning (LMKML) [26], SPD Manifold Learning (SPDML-AIRM,SPDML-Stein) [15], Adaptive Deep Network Template (ADNT) [17], Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG) [41] and Multi-manifold Deep Metric Learning (MMDML) [25].

We employed the implementations provided by the original authors, except for MMDML. Since the source code of MMDML has not been released, we implemented it according to the description in [25]. For fair comparison, the referred parameters of each method were followed by the original references. In MMD and MDA, we used the default parameters in [40, 38]. For AHISD, CHISD, we searched the PCA energy when learning the linear subspace through cross validation, and reported the best result for each method. In CDL, we used KDA for discriminative learning and the same setting as [39] on the YTC and PaSC datasets. For LMKML, we utilized the same setting as [26]. For SPDML-AIRM and SPDML-Stein, we searched v_w and v_b following the direction of [15]. For DARG, we reported the results of the "MD+LED" kernel which, as in [41], outperforms other kernels. Note that on the YTF dataset we cannot get the exact label, but only know whether an image pair belong to the same subject. Therefore, for DCC, we utilized a pairwise version and for CDL and DARG we used a kernel version of SILD [21] rather than KDA. While MMDML cannot deal with verification task, we only tested it by face identification task on the YTC dataset. In our method, parameters for constructing the affinity matrix were tuned via cross validation on different datasets.

On the YTC and YTF datasets, the gray intensity fea-

Method	Accuracy
DCC [23]	66.81±3.25
MMD [40]	65.30±3.23
MDA [38]	66.98±3.19
AHISD [6]	63.69±3.13
CHISD [6]	66.46±2.41
CDL [39]	69.70±1.13
LMKML [26]	70.31±2.52
SPDML-AIRM [15]	62.87±4.47
SPDML-Stein [15]	60.25±3.29
ADNT [17]	66.75±3.82
DARG [41]	77.09±2.65
MMDML [25]	69.81±4.62
FN-CDL	85.21 ±1.57
GE-CDL	**91.03±1.22**
SR-CDL	**94.85±0.96**

Table 1: Comparisons on the YTC dataset. Note that average accuracy ± standard deviation (%) is reported.

Method	Accuracy
DCC (pair) [23]	70.84
MMD [40]	64.96
AHISD [6]	66.50
CHISD [6]	66.24
CDL(pair) [39]	69.74
SPDML-AIRM [15]	66.11
SPDML-Stein [15]	65.85
DARG [41]	73.01
FN-CDL	79.07
GE-CDL	81.34
SR-CDL	82.25

Table 2: Comparisons on YTF. The performance is evaluated by the mean accuracy in this table.

Method	Control	Handheld
DCC [23]	39.87	37.05
CDL [39]	46.07	44.53
ADNT [17]	41.72	38.68
DARG [41]	47.73	44.02
FN-CDL	52.31	50.15
GE-CDL	55.65	53.06
SR-CDL	56.43	55.73

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.

tures are used as the original references for the comparison image set classification methods except ADNT where a Local Binary Pattern (LBP) [27] feature is extracted as in [17]. Moreover, on the PaSC dataset, we followed a work of [4] to extract the state-of-the-art deep features for the comparison image set classification methods, and only reported the results of some methods which are representative and have achieved good performance on YTC and YTF due to the large scale of PaSC.

For our proposed DCRL framework, we tested the performances of the two proposed schemes. Hereinafter, the Graph Embedding scheme is denoted by "GE" and the Softmax Regression scheme is called as "SR", where classification is performed by using CDL on the learned covariance matrices and computing the similarities on the resulting discriminative space. Besides, we also give the result by using the image features from a feature learning network, which is trained for single-shot image classification, to compute set covariance matrices. This comparison network is denoted by "FN" and is a DeepID2 [32] network with one single CNN. On all the three datasets, we pre-trained the proposed feature learning network by the CFW dataset [46] containing about 150,000 images from 1,580 subjects. For the training stage, on YTF the corresponding training data was employed to fine-tune the network; on PaSC, the COX face dataset [19] was used as auxiliary training data together with the training set of PaSC; on YTC, we directly extracted image features with a well-trained model for PaSC.

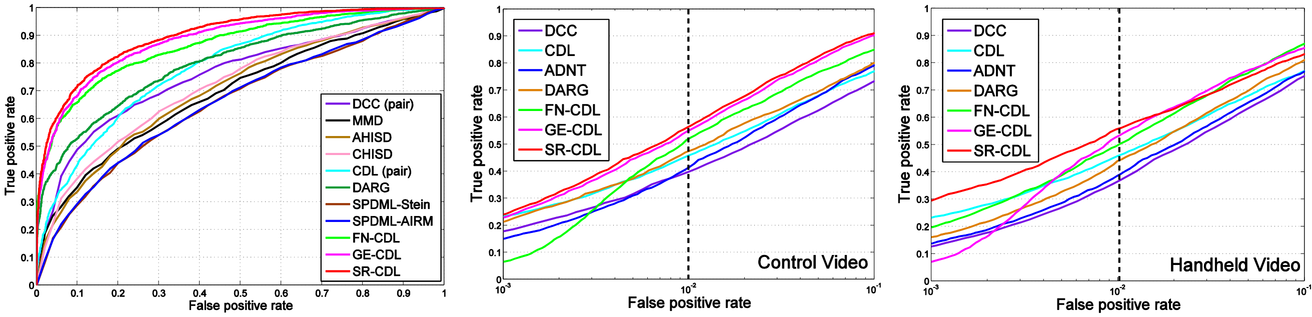
5.3. Results and Analysis

5.3.1 Comparison with the State-of-the-art

Firstly, we give a comparison result with the above state-of-the-art image set classification methods. We conducted face identification experiments on the YTC dataset and face verification experiments on the YTF and PaSC dataset. Tab. 1 tabulates the average recognition accuracy and standard deviation over ten-fold trials on the YTC dataset. The comparisons on the YTF dataset are shown in Tab. 2 and the performance is evaluated by the mean accuracy over ten-fold trials. Tab. 3 lists the verification rates (%) at a false accept rate (FAR) of 0.01 on the PaSC dataset. Besides, Fig. 2a and 2b shows the Receiving Operating Characteristic (ROC) curves on the YTF and PaSC dataset respectively. Note that in Fig. 2b, we highlight the verification rate at a false accept rate (FAR) of 0.01 by a vertical dotted line as its performance is usually reported on the PaSC dataset in previous literatures.

We can see from the results that the two proposed schemes both achieve superior performances in most tests. In the following, we give a brief analysis about the experimental results and try to conclude the probable causes.

1) Set covariance matrix model, which is used in CDL, LMKML, SPDML and our proposed approach, outperform-



(a) YTF

(b) PaSC

Figure 2: Comparisons of ROC curves on the YTF and PaSC dataset. On the PaSC dataset, the performance of the verification rate at a false accept rate (FAR) of 0.01 (often reported in previous literature) is shown by the vertical dotted line.

	FN	GE	SR
-LEM	33.63	35.74	34.52
-CDL	52.31	55.65	56.43

Table 4: Comparisons of different feature learning strategies (classified using LEM or CDL) on the control videos of the PaSC dataset.

s other set models promisingly in our experiments. This demonstrates the capability of covariance matrix to characterize the data structure of the image set.

2) Compared with other methods based on covariance matrix model, our proposed method shows better performance. This can be attributed to the learned image representations which more closely match the classification of the set covariance models.

3) In comparison with other deep model based methods ADNT and MMDML, our method yields better results, which implies the explicit use of discriminative information and set covariance model can facilitate the classification.

4) While on YTC our method has obtained a much higher performance than other methods which shows again the potential of image feature learning on this dataset, it should be noted that the comparison is not in a totally fair setting since external data were exploited to train our network for feature extraction and the RGB input contains more information than the gray intensity one.

5) Performances of all the methods on PaSC are relatively poor due to the low-quality and bad-alignment of face region images on PaSC. Even so, the Softmax Regression scheme with CDL still achieved an accuracy of 56.63% for control videos and 55.48% for handheld videos.

5.3.2 Comparisons of Feature Learning Strategies

To further manifest the superiority of our DCRL framework, we conducted experiments to employ different feature learning strategies before classifying with LEM or CDL. The comparisons of their performance are shown in Tab. 4. Clearly, our proposed method outperforms FN, which can be attributed to the benefit that the learned image representation is consistent with the set covariance model and also supports our motivation to facilitate classification of set covariance model.

6. Conclusion

This paper develops a Discriminative Covariance oriented Representation Learning (DCRL) framework for face recognition with image sets. Our method mainly contributes an early attempt in learning deep representations for images which is exactly suitable with both set covariance modeling and classification with such models. Our contributions mainly lie in three folds: 1) Our method extracts and organizes the discriminative information implicit in the images and image sets jointly which significantly facilitates the image set classification. 2) Flexible and efficient representation can be learned automatically due to favorable capability of deep learning in modeling nonlinearity and extracting discriminative information. 3) The proposed method can be not only directly used, but also seamlessly integrated with existing covariance modeling methods. Extensive experiments show that our method achieves impressive performance for face recognition with image sets.

In the future, the appealing idea of learning image representations consistent with both image set modeling and subsequent classification can be extended to other image set models as well as other deep model designs.

Acknowledgements

This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61379083, and Youth Innovation Promotion Association CAS No. 2015085.

References

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

[2] 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.

[3] Y. Bengio. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2009.

[4] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, et al. Report on the fg 2015 video person recognition evaluation. In *IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[5] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision (ECCV)*, 2012.

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

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

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

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

[10] 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.

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

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

[13] Q. Feng, Y. Zhou, and R. Lan. Pairwise linear regression classification for image set retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *International Conference on Machine learning (ICML)*, 2008.

[15] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: geometry-aware dimensionality reduction for SPD matrices. 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 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 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing (TIP)*, 2015.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International conference on Multimedia*, 2014.

[21] M. Kan, S. Shan, D. Xu, and X. Chen. Side-information based linear discriminant analysis for face recognition. In *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 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] 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.

[25] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

[27] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002.

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[29] 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 *IEEE International Conference on Biometrics: Theory, Applications and Systems (B-TAS)*, 2013.

[30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

[32] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[33] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[34] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[35] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani. Multi-class classification on Riemannian manifolds for video surveillance. In *European Conference on Computer Vision (ECCV)*, 2010.

[36] R. Vemulapalli and D. W. Jacobs. Riemannian metric learning for symmetric positive definite matrices. *arXiv preprint arXiv:1501.02393*, 2015.

[37] R. Vemulapalli, J. K. Pillai, and R. Chellappa. Kernel learning for extrinsic classification of manifold features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

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

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

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

[41] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

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

[44] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2007.

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

[46] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *IEEE Transactions on Multimedia*, 2012.

[47] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Joint face representation adaptation and clustering in videos. In *European Conference on Computer Vision (ECCV)*, 2016.