Prototype Discriminative Learning for Face Image Set Classification

Wen Wang[1,2], Ruiping Wang[1,2,3(✉)], Shiguang Shan[1,2,3], and 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
wen.wang@vipl.ict.ac.cn, {wangruiping,sgshan,xlchen}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Cooperative Medianet Innovation Center, Beijing, China

Abstract. This paper presents a novel Prototype Discriminative Learning (PDL) method to solve the problem of face image set classification. We aim to simultaneously learn a set of prototypes for each image set and a linear discriminative transformation to make projections on the target subspace satisfy that each image set can be optimally classified to the same class with its nearest neighbor prototype. For an image set, its prototypes are actually "virtual" as they do not certainly appear in the set but are only assumed to belong to the corresponding affine hull, i.e., affine combinations of samples in the set. Thus, the proposed method not only inherits the merit of classical affine hull in revealing unseen appearance variations implicitly in an image set, but more importantly overcomes its flaw caused by too loose affine approximation via efficiently shrinking each affine hull with a set of discriminative prototypes. The proposed method is evaluated by face identification and verification tasks on three challenging and large-scale databases, YouTube Celebrities, COX and Point-and-Shoot Challenge, to demonstrate its superiority over the state-of-the-art.

1 Introduction

As one of the most important problems in the field of computer vision, traditional face recognition is usually posed as a single image classification problem. With development of imaging technology, multiple images can be available for one person in many real-world application scenarios such as video surveillance, multi-view camera photos or online photo albums, etc. Since multiple images usually incorporate dramatically large variations in pose, illumination, expression and other factors, it is no longer sufficient for traditional face recognition to handle such scenarios, which leads to a new research focus on face image set classification. Compared with a single image, a set of images can provide more information to describe the subjects of interest, hence image set classification

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-54187-7_23) contains supplementary material, which is available to authorized users.

S.-H. Lai et al. (Eds.): ACCV 2016, Part III, LNCS 10113, pp. 344–360, 2017.
DOI: 10.1007/978-3-319-54187-7_23

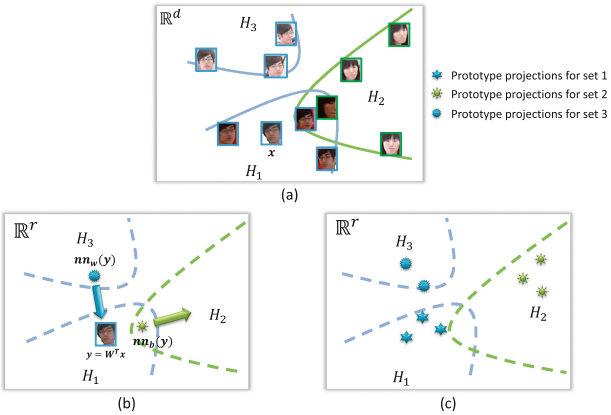

Fig. 1. Conceptual illustration. Different colors denote different subjects. $\mathbb{R}^d$ and $\mathbb{R}^r$ are respectively the original sample space and the projected subspace. (a) shows the affine hulls H_1, H_2 and H_3 of three image sets where H_1 and H_2 are overlapped which leads to a failed match. (b) illustrates the training process by taking any sample x in image set 1 as an example. The arrows imply the training objective, which is to make the projections in $\mathbb{R}^r$ satisfy that for a projection $y = W^T x$, its nearest neighbor in a prototype set from its own class (i.e., $nn_w(y)$) is closer than any other from different classes (i.e., $nn_b(y)$). (c) is an illustration of the learned target subspace and prototype sets.

is expected to achieve more appealing performance than single image classification. Generally speaking, existing image set classification methods mainly focus on how to model the image set and how to measure the dissimilarity between two sets.

In recent years, a simple but efficient affine hull model [1] is proposed to model the image set. The affine hull model tends to complement the unseen appearance variations that even do not appear in the image set via covering the affine combinations of sample images in this set. Thus the affine hull is quite appealing due to its favorable property of characterizing the implicit semantic relationship between the sample images in the set. Nevertheless, there are some fatal limitations. On the one hand, the affine hull matching fails when two hulls overlapped. This is usually caused by the over-large affine hull which usually occurs if the image set contains outliers such as incorrect or low-quality images. An illustration of such case is shown in Fig. 1(a). For shrinking the affine approximation, later methods attempt to artificially impose a tighter constraint (such as convex [1], sparse [2], regularized [3] or probabilistic [4] constraint) which, however, is a brute-force way and may lead to high time cost or missing of some representative candidate points. On the other hand, the discriminative information is ignored, while the affine hulls modeled based on original feature may not suffice to be discriminated linearly, which is iteratively learned in the form of discriminative metric in [5,6].

To address these limitations and explore a totally different and novel solution, this paper presents a Prototype Discriminative Learning (PDL) method for face image set classification. Our goal is to simultaneously learn a set of representative points (i.e. prototypes) for each image set and a linear discriminative projection. Thereinto, the learned prototypes of an image set are actually "virtual", that is, they do not certainly appear in the set but are assumed to belong to the corresponding affine hull, which aims to inherit the merit of affine hull in revealing unseen appearance variations. We expect that in the target projected subspace each image set can be optimally classified to the same class with its nearest neighbor prototype set. Figure 1(b) is an illustration of the training objective, while Fig. 1(c) shows the finally learned target subspace and prototype sets. Thus for an image set, its prototype set can be considered to shrink the corresponding affine hull discriminatively. Specifically, we estimate the loss function for classifying any image in an image set into the same class with its nearest prototype in the projected target subspace. Then by minimizing such loss, we can optimize the prototype sets and the linear projection simultaneously through gradient decent.

The rest of the paper is organized as follows: In Sect. 2, we review some existing works for face image set classification. Section 3 describes the proposed Prototype Discriminative Learning (PDL) method and then gives a discussion about some related works in the literature. In Sect. 4, we demonstrate the experimental evaluation on three challenging databases and analyze the comparison results with other works. Finally, Sect. 5 summarises the conclusion.

2 Related Work

In this section, we briefly introduce the existing works for face image set classification. To represent semantic relationship implicit in the image set, a lot of methods are proposed by exploring different kinds of image set models, for instance, one or several linear subspaces, statistical information, reconstruction model and affine hull.

In the literature, some methods tend to represent an image set as one or several linear subspaces. For example, Mutual Subspace Method (MSM) [7] and Discriminant analysis of Canonical Correlations (DCC) [8] model the image set with a single linear subspace and the difference between two subspaces is measured by principal angles. Grassmann Discriminant Analysis (GDA) [9] and Grassmann Embedding Discriminant Analysis (GEDA) [10] model the image set similarly but perform kernel discriminative learning on the Grassmann manifold where each point is a linear subspace. Besides, a series of works after Manifold-Manifold Distance (MMD) [11] propose to characterize an image set by multiple linear subspaces. Among them, MMD computes the distance between image sets by using the nearest distance between pair-wise local linear models and then Manifold Discriminant Analysis (MDA) [12] extends MMD by learning a discriminative feature subspace. Then an image set alignment method [13] is proposed to match the local linear subspaces more precisely. A later work of [14] proposes

to search joint Sparse Approximated Nearest Subspaces (SANS) and employ their distance to measure the image set dissimilarity. A Robust Structured Subspace Learning (RSSL) method [15] is proposed for data representation, which respects the locally smooth property of visual geometric structure.

Some methods propose to extract the statistical information, such as mean vector, covariance matrix, probability distribution, or a combination of them, to describe the data structure in the image set. Some earlier methods, e.g., [16,17], exploit some parametric distribution, such as Gaussian, to represent each image set and compute the similarity by Kullback-Leibler Divergence (KLD). The Covariance Discriminative Learning (CDL) method [18] exploits the covariance matrix to represent the image set and conducts kernel discriminant analysis on the Symmetric Positive Definite (SPD) manifold. Harandi et al. [19] present an SPD Manifold Learning (SPDML) method to learn an orthonormal projection from the high-dimensional SPD manifold to a low-dimensional, more discriminative one. Then Huang et al. [20] propose to learn a tangent map from the original tangent space to a new discriminative tangent space. A later work of Wang et al. [21] proposes 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).

Different from the methods above, some methods attempt to employ a reconstruction model to learn the image set representation implicitly and then compute the dissimilarity between image set by corresponding reconstruction error. For instance, face dictionary is extended from still images to videos and the sparse representation of image set can be learned through a sparse reconstruction mechanism. Specifically, Chen et al. [22,23] present a video-based dictionary method and build one dictionary for each video clip. Cui et al. [24] propose a Joint Sparse Representation (JSR) method to adaptively learn the sparse representation of an video clip with consideration of the class-level and atom-level sparsity simultaneously. Further, a simultaneous feature and dictionary learning (SFDL) method is proposed in [25] so that discriminative information can be jointly exploited. Besides, Hayat et al. [26] present an Adaptive Deep Network Template (ADNT) to learn a deep reconstruction network for each class.

In addition to the above three categories, some works present an affine hull based model to reveal the unseen appearance within an image set and the implicit semantic relationship from the view of general data geometric structure. For example, Affine Hull based Image Set Distance (AHISD) [1] is proposed to model each image set by an affine hull model and thereby defines the dissimilarity between two hulls as the distance between a pair of nearest points belonging to either hull respectively. Aiming at overcoming the disadvantage that the affine hull may be too large and overlapped, a following trend of works attempt to add some constraints to avoid too loose affine approximation. For example, Convex Hull based Image Set Distance (CHISD) [1] adds a coefficient bound to control the looseness of the convex approximation. Then Sparse Approximated Nearest Points (SANP) [2] is proposed to introduce a sparse representation constraint to the candidate points that the selected nearest points are required to be sparsely

represented by the original samples. More recently, Yang et al. [3] propose to approximate each image set by a regularized affine hull model, which exploits a constraint to regularize the affine hull. Further the work of [27] proposes to use multiple local convex hulls to approximate an image set. Chen et al. [28] propose to solve the matching problem by the minimal reconstruction error from a Dual Linear Regression Classification (DLRC) model. Wang et al. [4] propose to enhance the robustness against impure image sets by leveraging the statistical distribution of the involved image sets. To exploit the discriminative information, Zhu et al. [5] propose a Set-to-Set Distance Metric Learning (SSDML) method to learn proper metric between hulls iteratively. Besides, Leng et al. [6] extend SSDML with the strategy of prototype learning, which aims to iteratively filter out the outliers contained in original image set during SSDML.

3 Proposed Method

In this section, we first overview our proposed Prototype Discriminative Learning (PDL) method, followed by reviewing the affine hull model. Then we describe the details of the proposed method. Finally, we give a theoretical discussion about related methods.

3.1 Overview

This paper proposes a novel Prototype Discriminative Learning (PDL) method for face image set classification. As discussed in Sect. 1, it can promisingly improve the robustness of affine hull model to simultaneously learn prototypes and a linear discriminative projection, which are expected to satisfy the following two constraints.

(1) For an image set, its prototypes are a set of points belonging to the corresponding affine hull.
(2) Through the linear projection, the prototypes are mapped to a target subspace where for every sample image, its nearest neighbor in a prototype set from its own class is closer than any other from different classes.

With the first constraint, for an image set, its prototype set can be formulated as a set of combinations of the sample images in the set and to learn the prototypes, we just need to learn the corresponding affine coefficients. Hence, our proposed method inherits the favorable property of affine hull model that the unseen appearances can be revealed and employed to present the implicit semantic relationship by means of the general data geometric structure. The second constraint aims to drive that in the target subspace different image sets can be classified optimally to the same class with the nearest prototype set. We estimate the loss function similarly with the NN error estimation in [29–31]. Then by minimizing such loss function, we derive the corresponding gradients with respect to prototypes and the linear projection respectively, thus the optimized prototypes and linear discriminative projection can be learned simultaneously through gradient decent.

3.2 Affine Hull Model

Suppose there are a total of C image sets for training, the data matrix of the c-th image set is denoted by $X_c = \{x_{c,1}, x_{c,2}, ..., x_{c,n_c}\}$, where $x_{c,i}$ is a d-dimensional feature vector of the i-th image. The c-th image set can be approximated as the affine hull of the sample images [1].

$$H_c = \left\{ x = \sum_{i=1}^{n_c} \alpha_{c,i} \cdot x_{c,i} \middle| \sum_{i=1}^{n_c} \alpha_{c,i} = 1 \right\}, c = 1, ..., C. \tag{1}$$

By using the sample mean $\mu_c = \frac{1}{n_c} \sum_{i=1}^{n_c} x_{c,i}$ as a reference, we can rewrite the affine hull model as follows.

$$H_c = \{x = \mu_c + U_c v_c | v_c \in \mathbb{R}^l\}, c = 1, ..., C, \tag{2}$$

where U_c is an orthonormal basis and obtained by applying the Singular Value Decomposition (SVD) to the centered data matrix $[x_{c,1} - \mu_c, ..., x_{c,n_c} - \mu_c]$. Note that the directions corresponding to near-zero singular values are discarded, leading to l_c ($l_c < N_c$) singular vectors in U_c. As we have discussed in Sect. 1, the affine hull is a general geometric model containing all the affine combinations of sample images in the set, which can account for the unseen appearance, possible data variation, and further the semantic relationship between sample images. Nevertheless, such approximation is too loose and may lead to over-large affine hull. Therefore, in the following section we will introduce the proposed PDL method which simultaneously learns a prototype set to take place of the affine hull for each image set and a linear projection to make the prototype sets discriminative.

3.3 Prototype Discriminative Learning

Let $P = \{P_1, P_2, ..., P_C\}$ be a collection of the prototype sets to learn. Among them, for the c-th image set X_c, the prototype set can be denoted as $P_c = \{p_{c,1}, p_{c,2}, ..., p_{c,m_c}\} \subseteq H_c$, where

$$p_{c,i} = \mu_c + U_c v_{c,i}, \quad v_{c,i} \in \mathbb{R}^l. \tag{3}$$

Through a linear transformation W, we can obtain a projection in the target subspace which is denoted as

$$y = W^T x \in \mathbb{R}^r, \tag{4}$$

for each image data $x \in X_c, c = 1, ..., C$.

Our goal is to drive that for any image in each image set, it is closer to its nearest neighbor in any prototype set from the same class than that from different classes after mapped to the r-dimensional target subspace. Therefore, in reference of the NN error estimation in [29–31], we define a loss function as follows.

$$J(W, P_1, .., P_C) = \sum_{c=1}^{C} \sum_{x \in X_c} step(Q_x), \tag{5}$$

where $step(Q_x)$ is the step function, i.e.,

$$step(z) = \begin{cases} 0, & if \quad z < 1; \\ 1, & if \quad z \geq 1, \end{cases} \tag{6}$$

and

$$Q_x = \frac{d(y, nn_w^c(y))}{d(y, nn_b^c(y))}, \tag{7}$$

where $d(\cdot, \cdot)$ is the Euclidean distance. $nn_w^c(y)$ and $nn_b^c(y)$ are the nearest neighbors of y respectively from the projections of the same-class and different-class prototype sets, therefore we can formulate them as follows.

$$\begin{aligned} nn_w^c(y) &= W^T a, \quad a = \underset{\substack{a \in P \setminus P_c, \\ a \in Class(x)}}{\operatorname{argmin}} \; d(y, W^T a) \\ nn_b^c(y) &= W^T b, \quad b = \underset{\substack{b \in P \setminus P_c, \\ b \notin Class(x)}}{\operatorname{argmin}} \; d(y, W^T b) \end{aligned} \tag{8}$$

Equation (5) denotes the total loss of classifying all sample data $x \in \forall X_c$, $c = 1, ..., C$. Specifically, after mapped though W, when sample data x is nearer to a prototype of its own class than any other from a different class, the loss for classifying x is zero. On the contrary, if in the projected subspace, x is nearer to a prototype from some different class than any other from its own class, the classification of x is mistaken and a large loss of 1 is imposed. Note that here the nearest neighbor of $x \in X_c$ is searched in prototype sets except for the one corresponding to X_c.

Considering the differential property, we employ a sigmoid function with slope to approximate the step function, i.e.,

$$\mathcal{S}_\beta(z) = \frac{1}{1 + e^{\beta(1-z)}}. \tag{9}$$

Note that when β is large, $\mathcal{S}_\beta(\cdot)$ is a smooth approximation of the step function. Then the objective function can be rewritten as follows.

$$J(W, P_1, .., P_C) = \sum_{c=1}^{C} \sum_{x \in X_c} \mathcal{S}_\beta(Q_x). \tag{10}$$

3.4 Optimization

For learning optimal prototype sets $P = \{P_1, P_2, ..., P_C\}$ and linear transformation W, we need to solve the optimization problem in the following.

$$\{W^*, P_1^*, P_2^*, ..., P_C^*\} = \underset{W, P_1, P_2, ..., P_C}{\operatorname{argmin}} \; J(W, P_1, .., P_C). \tag{11}$$

In this paper, a gradient descent method is employed to solve such problem. Then we tend to derive the gradient of loss function J with respect to $W, P_1, P_2, ..., P_C$. Since the procedure to search the nearest prototype depends on the prototype sets and transformation matrix but is non-continuous and problematic, a simple approximation is usually exploited with such dependence ignored. That is to say, the same prototype neighbor is searched when the variation in the prototype sets and transformation matrix is sufficiently small [29]. Under such assumption, we can derive the gradient of J with respect to W approximately as follows:

$$\begin{aligned}\frac{\partial J}{\partial W_k} \approx & \sum_{c=1}^{C} \sum_{x \in X_c} \frac{\mathcal{S}'_\beta(Q_x) Q_x}{d^2(y, nn_w^c(y))} \cdot (x-a)(y_k - nn_w^c(y)_k) \\ & - \sum_{c=1}^{C} \sum_{x \in X_c} \frac{\mathcal{S}'_\beta(Q_x) Q_x}{d^2(y, nn_b^c(y))} \cdot (x-b)(y_k - nn_b^c(y)_k),\end{aligned} \tag{12}$$

where $W_k \in \mathbb{R}^d$ denote the k-th column of W and y_k denote the k-th element of vector y. Note that denotations a and b have been defined in Eq. (8).

According to Eq. (3), for learning the prototype sets, we just need to learn the corresponding $\{v_{c,i}\}$, $i = 1, ..., m_c$ and $c = 1, ..., C$. Thus we derive the gradient of J with respect to each vector v_{ci} as follows.

$$\begin{aligned}\frac{\partial J}{\partial v_{ci}} \approx & \sum_{c=1}^{C} \sum_{\substack{x \in X_c \\ p_{ci}=a}} \frac{\mathcal{S}'_\beta(Q_x) Q_x}{d^2(y, nn_w^c(y))} \cdot U_c^T W W^T (a-x) \\ & - \sum_{c=1}^{C} \sum_{\substack{x \in X_c \\ p_{ci}=b}} \frac{\mathcal{S}'_\beta(Q_x) Q_x}{d^2(y, nn_b^c(y))} \cdot U_c^T W W^T (b-x),\end{aligned} \tag{13}$$

For space limitation, the detailed derivations of Eqs. (12) and (13) are given in a supplementary material.

Based on the derived gradients above, we can update the prototype sets and the linear projection in an iterative procedure by using the limited-memory BFGS (L-BFGS) method [32].

3.5 Classification

After the training process, we have computed a optimal linear transformation W and prototype sets $P_1, P_2, ..., P_C$ for the total of C training image sets. Then given a total of K image sets as the gallery, we need to give a prediction of the label for a new test image set. First we optimize the prototype set for each gallery image set with W fixed by solving Eq. (11). Then we compute the projection of these gallery prototype sets and the test image set through W. Finally, the distance between the test image set and a gallery image set can be computed as the minimal distance between samples in the test image set and prototypes corresponding to the gallery image set. Thus, the test image set can be classified into the same class with its nearest gallery prototype set in the target subspace.

Algorithm 1. PDL-training

Input:

Data matrices of C image sets for training: $\{X_1, X_2, ..., X_C\}$ and their labels;
the slope for sigmoid function: β;
the initial prototype sets: $P = \{P_1, ..., P_C\}$;
the initial transformation matrix: W.

Output:

The optimal P and W

```
Initialize the value of J as zero;
while not converged do
  for c = 1 to C do
    for all x such that x ∈ X_c do
      Compute the projection y by Eq. (4);
      Solve the optimization problem in Eq. (8) to compute nn_w^c(y), nn_b^c(y);
      Compute Q_x by Eq. (7);
      Add S_β(Q_x) to the value of J;
      Add to the gradient with respect to W and P respectively by Eqs. (12) and
      (13);
    end for
  end for
  Compute the step length and seeking direction by the L-BFGS algorithm;
  Update P and W;
end while
return P*, W*;
```

The Algorithms 1 and 2 summarize the training and testing process of our proposed PDL method respectively.

3.6 Discussion About Related Works

Firstly, we analyze the differences between our proposed PDL method and the unsupervised affine hull methods, such as AHISD [1], CHISD [1], SANP [2], RNP [3], DLRC [28] and ProNN [4], etc. (1) For AHISD, the affine hull may suffer from the issue of intersection, which makes the subsequent distance computation incorrect. Later CHISD, SANP, RNP and ProNN all attempt to solve such issue by imposing a constraint (such as convex, sparse, regularized, or probabilistic constraint) to the geometry structure of affine hull or the selection criteria of nearest points. These constraints are artificially set and based on additional assumption, which may lead to high time cost or missing of some useful information. On the contrary, our method efficiently ameliorates this issue by learning more representative and discriminative prototypes from the affine hull adaptively. (2) These methods are all unsupervised, while the discriminative information has been widely considered to be very important for the object classification.

Secondly, we figure out the differences from the supervised affine hull methods SSDML [5] and SPML [6]. (1) SSDML and SPML both follow a metric learning

Algorithm 2. PDL-testing

Input:

Data matrices of K image sets as gallery and their labels $\{L_1, ..., L_K\}$;
Data matrix of an image set for test: $T = \{t_1, ..., t_{n_t}\}$, $t_i \in \mathbb{R}^d$;
the slope for sigmoid function: β;
the initial prototype sets for the gallery image sets: $G = \{G_1, ..., G_K\}$;
the initial transformation matrix: W.

Output:

The label of the test image set L^*

1: Learn the prototype sets $G^* = \{G_1^*, ..., G_K^*\}$ with W fixed similarly with Alg.1.
2: Compute the projection $\widehat{T}$ by applying Eq. (4) to each sample vector in T;
3: Compute the projection $\widehat{G^*} = \{\widehat{G_1^*}, ..., \widehat{G_N^*}\}$ by applying Eq. (4) to each sample vector in G_i, $i = 1, ..., K$;
4: $k^* = \text{argmin}_{j,k}\, d(\widehat{T}, \widehat{G_{j,k}^*})$
5: **return** $L^* = L_{k^*}$;

framework, while our PDL proposes a different strategy of learning a linear discriminative projection. (2) They both exploit a global discriminative learning, while we conduct PDL from a local view of nearest neighbor (NN), that is, only to penalize a larger distance between nearest neighbors from different classes than that from the same class. From the local view, the optimization objective is more consistent with the final NN-based classification, thus can facilitate more precise classification. (3) To solve the optimization problem, they both adopt a strategy of alternately optimizing. On the contrary, PDL presents a joint optimization mechanism, which can favorably reduce time complexity and avoid trapping in local optimum to some extent. (4) SPML can be considered as iteratively filtering out outlier samples (the remaining real samples are their so-called "prototypes") in the image set while learning discriminative metric. In contrast, our PDL aims to learn discriminative virtual prototypes, which do not necessarily appear in the original set as in SPML but are just required to belong to the corresponding affine hull. Based on this different problem formulation, the learning strategy in PDL is believed to be more direct and efficient.

Thirdly, we give a discussion about comparison with the prototype selection methods based on single sample/image [33,34]. These methods usually propose to select prototypes from the existing samples as a reference for nearest neighbor classifier, which is confined only to the existing samples. However, we argue that for the image set classification problem, the appearance variations within an image set may be too large to be matched only based on existing samples. On the contrary, our PDL first employs the affine hull to complement the unseen data variations and subsequently learns prototypes from these affine combinations, which is more specifically suited to the classification of image sets containing complex data variations.

4 Experiments

4.1 Databases and Settings

For evaluating our proposed PDL method, we used three challenging and large-scale databases: YouTube Celebrities (YTC) [35], COX [36] and Point-and-Shoot Challenge (PaSC) [37]. Examples in the three databases are shown in Fig. 2.

The YTC database is collected from YouTube and consists of 1,910 highly compressed and low-resolution video sequences belonging to 47 subjects. The face region in each image was resized into 20×20 intensity image, and was processed with histogram equalization to eliminate lighting effects. Following the similar protocols of [18,25], 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 to cover all of the 1,910 clips in the database.

The COX database is a large-scale video database and contains 3,000 video sequences from 1,000 different subjects which are captured by different camcorders. In each video, there is around 25–75 frames of low resolution and low quality, with blur, and captured under poor lighting. The face in each image was resized into 32×40 intensity image and histogram equalized. Since the database 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.

The PaSC database consists of 2,802 videos of 265 people carrying out simple actions. Half of these videos are captured by a controlled video camera, the rest are captured by one of five alternative hand held video cameras. It has a total of 280 sets for training and 1401 sets for testing. Verification experiments were conducted using control or handheld videos as target and query respectively. Since the database is relatively difficult, we followed a work of [38] to extract the state-of-the-art Deep Convolutional Neural Network (DCNN) features rather

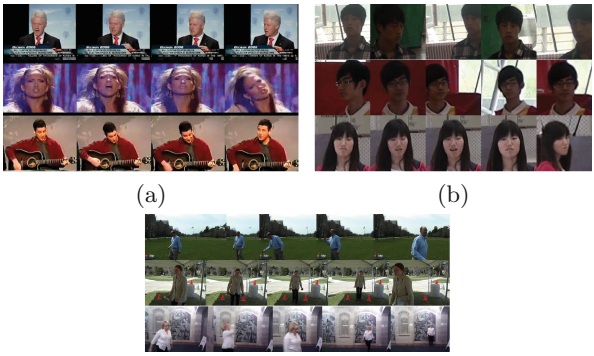

Fig. 2. Some examples of the databases. (a) YTC (b) COX (c) PaSC

than original gray features. Here the DCNN model is pre-trained on the CFW database [39] and subsequently fine-tuned on the training data of PaSC and COX database by using the Caffe [40].

4.2 Comparative Methods

To study the effectiveness of our proposed PDL method, we compared with several state-of-the-art image set classification methods. Among them, there are several affine hull based methods, including Mutual Subspace Method (MSM) [7], Affine Hull based Image Set Distance (AHISD) [1], Convex Hull based Image Set Distance (CHISD) [1], Sparse Approximated Nearest Point (SANP) [2], Regularized Nearest Points (RNP)[3], Dual Linear Regression Classification (DLRC) [28] and Set-to-Set Distance Metric Learning (SSDML) [5]. In addition, we also gave the comparison results with some other state-of-the-art supervised methods, such as Discriminative Canonical Correlations (DCC) [8], Manifold Discriminant Analysis (MDA) [12], Grassmann Discriminant Analysis (GDA) [9] and Grassmannian Graph Embedding Discriminant Analysis (GEDA) [10].

The source code of all comparative methods released by the original authors were used except that of DLRC. We carefully implemented the DLRC algorithm since downloading of its code is not available on its website now. For fair comparison, the important parameters of all the methods were carefully tuned following the recommendations in the original works: For AHISD, we retained 95% energy when learning the orthonormal basis. For CHISD, the error penalty was set to be $C = 100$ as in [1]. For SANP, the parameters were the same as [2]. Considering the high time cost in SANP, we only compared with it on YTC and COX. Note that since the SANP method is too time-consuming to run under the setting of COX, which contains large-scale data, we alternately took the image sets of 100 persons rather than all the 700 persons for testing. For RNP and DLRC, all the parameters were configured according to [3,28] respectively. For SSDML, we set $\lambda_1 = 0.001$, $\lambda_2 = 0.5$, the numbers of the positive pairs and the negative pairs per set are set to 10 and 20. For DCC, corresponding 10 maximum canonical correlations were used. For MDA, the parameters were configured according to [12]. For GDA/GEDA, the dimension of Grassmannian manifold was set to 10. For our proposed PDL, we used the PCA transformation matrix as an initialization of W and employed unit vectors to initialize the coefficients in P.[1]

4.3 Results and Analysis

The identification experiments were conducted on the YTC and the COX database. Table 1 tabulates the rank-1 identification rates on the YTC and the COX databases, where each reported rate is a mean accuracy over the ten-fold trials. Then we used the PaSC database to evaluate our performance on the verification task and Table 2 lists the verification rate at a false accept rate (FAR) of 0.01.

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

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	YTC	COX-12	COX-13	COX-23	COX-21	COX-31	COX-32
DCC [8]	0.668	0.625	0.661	0.506	0.561	0.638	0.452
MDA [12]	0.670	0.658	0.630	0.362	0.554	0.432	0.297
GDA [9]	0.659	0.723	0.807	0.744	0.714	0.820	0.776
GEDA [10]	0.668	0.767	0.838	0.766	0.726	0.828	0.800
AHISD [1]	0.637	0.530	0.361	0.175	0.435	0.350	0.188
CHISD [1]	0.665	0.569	0.301	0.148	0.444	0.264	0.137
SANP [2]	0.684	0.541	0.360	0.156	0.396	0.271	0.148
RNP [3]	0.703	0.525	0.333	0.148	0.581	0.379	0.146
DLRC [28]	0.692	0.492	0.379	0.155	0.441	0.361	0.175
SSDML [5]	0.689	0.601	0.531	0.287	0.479	0.444	0.273
PDL	**0.743**	**0.796**	**0.869**	**0.822**	**0.760**	**0.871**	**0.824**

Table 2. Verification rates on PaSC when false accept rate is 0.01 on PaSC dataset. Note that the control/handheld indicates the experiments with control/handheld videos.

Method	Control	Handheld
DCC [8]	0.389	0.375
GDA [9]	0.397	0.375
GEDA [10]	0.406	0.390
AHISD [1]	0.219	0.143
CHISD [1]	0.261	0.210
RNP [3]	0.274	0.198
DLRC [28]	0.242	0.171
SSDML [5]	0.292	0.229
PDL	**0.415**	**0.396**

As can be seen in the results, our method performs the best on all of the three databases. Firstly, our PDL achieves an impressively better result than the unsupervised affine hull based methods, such as AHISD, CHISD, SANP, RNP and DLRC. Specifically, on YTC, our method performs higher than a baseline method AHISD by 11%. On PaSC, our PDL outperforms AHISD by 19.6% for the control videos and 25.3% in the handheld scenario respectively. This supports the discussions in Sect. 3.6 that our PDL improves the affine hull model by learning prototypes discriminatively and adaptively, which is more flexible and robust than artificially imposing a tighter constraint. Secondly, our PDL is also superior over the supervised affine hull based method SSDML. As discussed in

Sect. 3.6, it mainly attributes to our innovation in learning virtual prototypes, the local discriminative learning strategy and the joint optimization mechanism. Besides, it can be generally observed that the supervised methods outperform the unsupervised methods more obviously on COX than the other two databases, due to the particularly large within-class variations on COX and the similar motions between different faces captured by the same camera.

4.4 Time Comparison

In addition, we compared the computational complexity of different methods on an Intel i7-3770, 3.40 GHz PC. Table 3 lists the time cost for the comparative methods for training and testing respectively on the YTC database. Note that only supervised methods need the training time. In practice, test time is more important for the efficiency of a method, as the training process can be conducted offline. From the table, we can see that our proposed method is very efficient and is faster than other affine hull based methods. Since For testing, our method only need to compute the projections and their distance, it is relatively efficient and is faster than other affine hull based methods.

Table 3. Time comparison (seconds) of different methods on YTC for training and testing.

Method	MSM	AHISD	CHISD	SANP	RNP	DLRC	SSDML	PDL
Training	N/A	N/A	N/A	N/A	N/A	N/A	346.33	75.30
Testing	1.31	1.58	1.71	56.77	1.56	1.91	2.35	1.15

5 Conclusions

This paper has proposed a novel Prototype Discriminative Learning method for face image set classification. We represented an image set by a prototype set learned from its basic affine hull model to shrink the loose affine hull effectively while inheriting the merit of affine hull in complementing the unseen appearance with affine combinations. Meanwhile, a linear projection was learned to drive that in the target projected subspace, the learned prototypes can be used to discriminate image sets of different classes. Our experimental evaluation has demonstrated that the proposed method can lead to state-of-the-art recognition accuracies on several challenging databases for face image set identification/verification.

In the future, we will study the regularization of W as well as prototypes more comprehensively. Further, we will explore the effect of learning representative prototypes in the large-scale and unclean image sets and study to construct dense and effective prototypes which can be easily adapted to other typical well established image set models, such as, linear subspace based set models, manifold based set models or statistical set models.

Acknowledgement. This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61379083, 61272321, and Youth Innovation Promotion Association CAS No. 2015085.

References

1. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
2. Hu, Y., Mian, A.S., Owens, R.: Sparse approximated nearest points for image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
3. Yang, M., Zhu, P., Gool, L.V., Zhang, L.: Face recognition based on regularized nearest points between image sets. In: IEEE Conference on Automatic Face and Gesture Recognition (FG) (2013)
4. Wang, W., Wang, R., Shan, S., Chen, X.: Probabilistic nearest neighbor search for robust classification of face image sets. In: IEEE Conference on Automatic Face and Gesture Recognition (FG) (2015)
5. Zhu, P., Zhang, L., Zuo, W., Zhang, D.: From point to set: extend the learning of distance metrics. In: IEEE International Conference on Computer Vision (ICCV) (2013)
6. Leng, M., Moutafis, P., Kakadiaris, I.A.: Joint prototype and metric learning for set-to-set matching: application to biometrics. In: IEEE Conference on Biometrics Theory, Applications and Systems (BTAS) (2015)
7. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: IEEE Conference on Automatic Face and Gesture Recognition (FG) (1998)
8. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **29**, 1005–1018 (2007)
9. Hamm, J., Lee, D.D.: Grassmann discriminant analysis: a unifying view on subspace-based learning. In: International Conference on Machine Learning (ICML) (2008)
10. Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C.: Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
11. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
12. Wang, R., Chen, X.: Manifold discriminant analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
13. Cui, Z., Shan, S., Zhang, H., Lao, S., Chen, X.: Image sets alignment for video-based face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
14. Chen, S., Sanderson, C., Harandi, M.T., Lovell, B.C.: Improved image set classification via joint sparse approximated nearest subspaces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
15. Li, Z., Liu, J., Tang, J., Lu, H.: Robust structured subspace learning for data representation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **37**, 2085–2098 (2015)

16. Shakhnarovich, G., Fisher, J.W., Darrell, T.: Face recognition from long-term observations. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 851–865. Springer, Heidelberg (2002). doi:10.1007/3-540-47977-5_56
17. Arandjelović, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005)
18. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: a natural and efficient approach to image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
19. Harandi, M.T., Salzmann, M., Hartley, R.: From manifold to manifold: geometry-aware dimensionality reduction for SPD matrices. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 17–32. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_2
20. Huang, Z., Wang, R., Shan, S., Li, X., Chen, X.: Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In: International Conference on Machine Learning (ICML) (2015)
21. Wang, W., Wang, R., Huang, Z., Shan, S., Chen, X.: 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)
22. Chen, Y.-C., Patel, V.M., Phillips, P.J., Chellappa, R.: Dictionary-based face recognition from video. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 766–779. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_55
23. Chen, Y.C., Patel, V.M., Shekhar, S., Chellappa, R., Phillips, P.J.: Video-based face recognition via joint sparse representation. In: IEEE Conference on Automatic Face and Gesture Recognition (FG) (2013)
24. Cui, Z., Chang, H., Shan, S., Ma, B., Chen, X.: Joint sparse representation for video-based face recognition. Neurocomputing **135**, 306–312 (2014)
25. Lu, J., Wang, G., Deng, W., Moulin, P.: Simultaneous feature and dictionary learning for image set based face recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 265–280. Springer, Cham (2014). doi:10.1007/978-3-319-10590-1_18
26. Hayat, M., Bennamoun, M., An, S.: Learning non-linear reconstruction models for image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
27. Chen, S., Wiliem, A., Sanderson, C., Lovell, B.C.: Matching image sets via adaptive multi convex hull. arXiv preprint arXiv:1403.0320 (2014)
28. Chen, L.: Dual linear regression based classification for face cluster recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
29. Paredes, R., Vidal, E.: Learning prototypes and distances: a prototype reduction technique based on nearest neighbor error minimization. Pattern Recogn. **39**, 180–188 (2006)
30. Paredes, R., Vidal, E.: Learning weighted metrics to minimize nearest-neighbor classification error. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **28**, 1100–1110 (2006)
31. Villegas, M., Paredes, R.: Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)

32. Le, Q.V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Ng, A.Y.: On optimization methods for deep learning. In: International Conference on Machine Learning (ICML) (2011)
33. Garcia, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **34**, 417–435 (2012)
34. Ma, M., Shao, M., Zhao, X., Fu, Y.: Prototype based feature learning for face image set classification. In: IEEE Conference on Automatic Face and Gesture Recognition (FG) (2013)
35. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
36. Huang, Z., Wang, R., Shan, S., Chen, X.: Learning euclidean-to-riemannian metric for point-to-set classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
37. Ross, B., Phillips, J., Bolme, D., Draper, B., Givens, G., Lui, Y.M., Teli, M.N., Zhang, H., Scruggs, W.T., Bowyer, K., Flynn, P., Cheng, S.: The challenge of face recognition from digital point-and-shoot cameras. In: IEEE Conference on Biometrics Theory, Applications and Systems (BTAS) (2013)
38. Beveridge, J.R., Zhang, H., Draper, B.A., Flynn, P.J., Feng, Z., Huber, P., Kittler, J., Huang, Z., Li, S., Li, Y., Kan, M., Wang, R., Shan, S., Chen, X.: Report on the FG 2015 video person recognition evaluation. In: IEEE Conference and Workshops on Automatic Face and Gesture Recognition (FG) (2015)
39. Zhang, X., Zhang, L., Wang, X.J., Shum, H.Y.: Finding celebrities in billions of web images. IEEE Trans. Multimedia **14**, 995–1007 (2012)
40. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia (2014)