Prototype Discriminative Learning for Image Set Classification

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

***Abstract*—This letter presents a prototype discriminative learning (PDL) method for image set classification. We aim to simultaneously learn prototypes and a linear discriminative projection to drive that in the target subspace each image set can be discriminated with its nearest neighbor prototype. To reveal the unseen appearance variations implicitly in an image set, the prototypes are actually "virtual," which do not certainly appear in the set but are searched in the corresponding affine hull. Moreover, to enhance the stability and robustness of the learned target subspace, an orthogonality constraint is imposed on the projection. Thus, to optimize the prototypes and the projection jointly, we design a specific gradient descent mechanism by updating the projection on Stiefel manifold and the prototypes in Euclidean space in an alternative optimization manner. Experimental results on four challenging databases demonstrate the superiority of the proposed PDL method.**

***Index Terms*—Discriminative learning, image set classification, prototype learning.**

I. Introduction

SINCE a set of images can provide more information to more effectively describe the subjects of interest than a single image, there has been a growing research focus on image set classification [1]–[9]. However, multiple images usually incorporate dramatically large variations in pose, illumination, expression, etc., which poses a new challenge on modeling the useful information contained implicitly in an image set.

In recent years, a simple but efficient affine hull model [1] is proposed to model the image set. 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, there are some fatal limitations: 1) The affine hull may be overlarge. An illustration of such case is shown in Fig. 1(a) where two hulls H_1 and H_2 from different subjects are overlapped and it leads to a failed match. For shrinking the affine approximation, later methods attempt to artificially impose a tighter constraint (such as convex [1], sparse [10], regularized [11], or probabilistic [12] constraint) which, however, is a brute-force way and may lead to high time cost or missing of some representative candidate points. 2) The discriminative information is ignored. To tackle this problem, Zhu *et al.* [13] propose to alternatively optimize a discriminative metric and nearest affine points, which is hard to solve with the possibility of trapping in local optimum and high time complexity. A later work [14] extends [13] by iteratively filtering out outlier samples, which may suffer from information missing and computational difficulty.

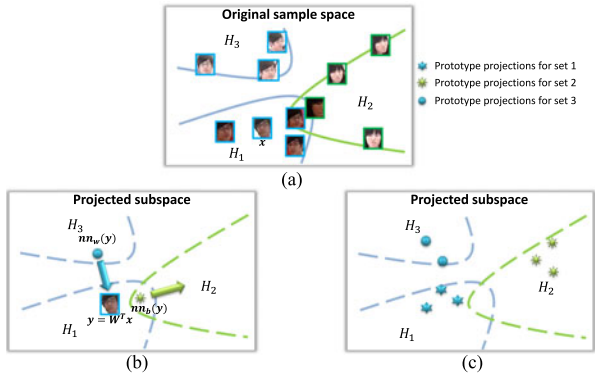

Fig. 1. Conceptual illustration. Different colors denote different subjects. (a) Shows a failed case caused by overlarge affine hulls. (b) Illustrates the training process. The arrows imply the training objective that for a projected point y, its nearest prototype from different classes $nn_b(y)$ leaves away and the one from the same class $nn_w(y)$ approaches near. (c) Depicts the learned target subspace and prototype sets.

To address these limitations, this letter explores a totally different and novel solution of shrinking the corresponding affine hull discriminatively. We present a prototype discriminative learning (PDL) method aiming at learning a set of representative points (i.e., prototypes) for each image set and a linear discriminative projection simultaneously. To favorably inherit the merit of affine hull in revealing unseen appearance variations, the learned prototypes of an image set are actually "virtual," that is, they do not certainly appear in the set but are searched in the corresponding affine hull. Moreover, as shown in Fig. 1, the training objective is estimated from a local view of nearest neighbor (NN) to penalize a larger distance between NN from different classes than that from the same class. This allows that in the target subspace each image set can be optimally classified to the same class with its nearest prototype set. To further guarantee the stability and robustness of the

Manuscript received April 1, 2017; revised June 7, 2017; accepted June 20, 2017. Date of publication July 4, 2017; date of current version July 24, 2017. This work was supported in part by 973 Program under Grant 2015CB351802, in part by the Natural Science Foundation of China under Grant 61390511, Grant 61379083, Grant 61650202, Grant 61402443, and Grant 61672496, and in part by Youth Innovation Promotion Association CAS under Grant 2015085. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vishal Monga. *(Corresponding author: Ruiping Wang.)*

The authors are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wen.wang@vipl.ict.ac.cn; ruiping.wang@vipl.ict.ac.cn; sgshan@ict.ac.cn; xlchen@ict.ac.cn).

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

Digital Object Identifier 10.1109/LSP.2017.2723084

target subspace, the projection is regularized to be orthonormal. Hence, to optimize the prototypes and the projection jointly, we design a specific gradient descent algorithm, which updates the projection on Stiefel manifold and the prototypes in Euclidean space in an alternative optimization manner. A preliminary conference version has been published in [15], and compared with it, this letter has made three major extensions.

1) Further improvement for the stability and robustness of the target subspace.
2) More comprehensive investigation for the main factors.
3) More extensive experiments to evaluate the method and compare with other state-of-the-art algorithms.

The contributions of the proposed PDL method mainly lie in the following four aspects.

1) PDL learns the prototypes and a linear discriminative projection with a joint optimization mechanism.
2) Prototypes are not limited by the existing image samples but complement the unseen data variations with affine combinations.
3) PDL is designed focusing on the interclass and intraclass NN prototypes, which is consistent with NN-based testing, leading to more efficient training and more precise classification.
4) We study an orthonormal projection constraint to retain the geometric property favorably and present a specific gradient descent algorithm.

II. Prototype Discriminative Learning

In this section, we give a detailed formulation of the proposed PDL and introduce the optimization algorithm for jointly learning the prototypes and the linear projection.

A. Prototype Representation

We start with reviewing the affine hull model [1]. Suppose there are a total of C image sets for training, the cth one is denoted by $X_c = \{x_{c,i}\}_{i=1}^{n_c}$, where $x_{c,i}$ is a d-dimensional feature vector of the ith image. The cth image set can be approximated by the affine hull of the sample images:

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

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:

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

where U_c is an orthonormal basis and obtained by applying singular value decomposition to the centered data. Since the directions corresponding to near-zero singular values are discarded, U_c contains l_c ($l_c < n_c$) singular vectors.

Let $P = \{P_1, P_2, \ldots, P_C\}$ be a collection of the prototype sets to learn. Among them, for the cth image set X_c, the prototype set can be denoted as $P_c = \{p_{c,i}\}_{i=1}^{m_c}$. To make the learned prototypes more flexible and representative, we propose to search the prototypes from the corresponding affine hull, rather than from the existing samples, i.e., $P_c \subseteq H_c$ and according to (2), we have

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

B. Loss Function

Besides the prototypes, we also need to learn a linear projection W to guarantee its discrimination. For each image sample $x \in X_c$, its projection through W is formulated as follows:

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

Our goal is to drive that for any image in each image set, it is closer to its NN in any prototype set from the same class than that from different classes after mapped with W. Therefore, in reference of the NN error estimation in [16]–[18], we define a loss function as follows:

$$J(W, P) = \sum_{c=1}^{C} \sum_{x \in X_c} \mathcal{S}_\beta(Q_x) \tag{5}$$

where $\mathcal{S}_\beta(z) = \frac{1}{1+e^{\beta(1-z)}}$ is a smooth approximation of the step function when β is large:

$$Q_x = \frac{\|y - nn_w^c(y)\|_2}{\|y - nn_b^c(y)\|_2} \tag{6}$$

where $nn_w^c(y)$ and $nn_b^c(y)$ are the NNs of y respectively from the projections of the same-class and different-class prototype sets, and we can formulate them as follows:

$$\begin{aligned} nn_w^c(y) &= W^T a, \quad a = \underset{\substack{a \in P \setminus P_c, \\ a \in \text{Class}(x)}}{\operatorname{argmin}} \left\| y - W^T a \right\|_2 \\ nn_b^c(y) &= W^T b, \quad b = \underset{\substack{b \in P \setminus P_c, \\ b \notin \text{Class}(x)}}{\operatorname{argmin}} \left\| y - W^T b \right\|_2. \end{aligned} \tag{7}$$

To further facilitate numerical stability and achieve a more robust target subspace, an orthogonality constraint is imposed on the projection W, i.e., we need to solve the optimization problem $\min_{W,P} J(W, P)$ with a constraint $W^T W = I_r$.

C. Optimization

For learning the optimal prototype sets $P = \{P_1, P_2, \ldots, P_C\}$ and the linear projection W, a gradient descent method is employed to minimize the loss function $J(W, P)$. Then we tend to derive the gradient of loss function J with respect to W and P. The procedure to search the nearest prototype depends on the prototype sets and the projection, but it is noncontinuous and problematic. Thus, a simple approximation is usually exploited with such dependence ignored [16]. Under such assumption, we can derive the gradient of J with respect to W approximately as follows:

$$\begin{aligned} \frac{\partial J}{\partial W_k} \approx & \sum_{c=1}^{C} \sum_{x \in X_c} \frac{\mathcal{S}'_\beta(Q_x) Q_x}{\|y - nn_w^c(y)\|_2^2} \cdot (x - a)(y_k - nn_w^c(y)_k) \\ & - \sum_{c=1}^{C} \sum_{x \in X_c} \frac{\mathcal{S}'_\beta(Q_x) Q_x}{\|y - nn_b^c(y)\|_2^2} \cdot (x - b)(y_k - nn_b^c(y)_k) \end{aligned} \tag{8}$$

where $W_k \in \mathbb{R}^d$ denote the kth column of W and y_k denote the kth element of vector y.

According to (3), for learning the prototype sets, we just need to learn the corresponding affine coefficients $V = \{V_1, \ldots, V_C\}, V_c = \{v_{c,i}\}_{i=1}^{m_c}$. Thus, we derive the gradient of

J with respect to each vector v_{ci} as follows:

$$\frac{\partial J}{\partial v_{ci}} \approx \sum_{c=1}^{C} \sum_{\substack{x \in X_c \\ p_{ci}=a}} \frac{\mathcal{S}'_\beta(Q_x)Q_x}{\|y - nn_w^c(y)\|_2^2} \cdot U_c^T W W^T (a - x) - \sum_{c=1}^{C} \sum_{\substack{x \in X_c \\ p_{ci}=b}} \frac{\mathcal{S}'_\beta(Q_x)Q_x}{\|y - nn_b^c(y)\|_2^2} \cdot U_c^T W W^T (b - x). \quad (9)$$

To solve the optimization problem without orthonormal projection constraint, we can update W and V in an iterative procedure based on the derived gradients in the above-mentioned equation by using limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [19] to control the step size. As for the optimization problem with orthonormal projection constraint, it is nontrivial to jointly optimize W and V as the feasible set of W is on a Stiefel manifold. Considering its favorable property of low computational cost, we choose the curvilinear searching method in [20] to search the step size and the path of W. Furthermore, to guarantee the optimization of W and V to be performed jointly and consistently, at each step, we update W along curvilinear retraction by using [20] and V along straight lines by using L-BFGS. Although the proof of convergence is not given for the proposed optimization algorithm, its convergence to a global minimum was confirmed experimentally in Section III-C.

D. Classification

After the training process, we have computed an optimal linear projection W and prototype sets $P_1, P_2, \ldots, P_C$ 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 minimizing (5). 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 of its nearest gallery set.

III. Experiments

A. Databases and Settings

For evaluating our proposed PDL method, we used four challenging and large-scale databases: YouTube celebrities (YTC) [21], Chinese Academy of Sciences-OMRON Social Solutions-Xinjiang University (COX) [22], multiple biometric grand challenge (MBGC) [23], and point-and-shoot challenge (PaSC) [24].

The YTC database consists of 1910 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 [2] and [25], we conducted tenfold cross validation experiments and randomly selected three clips for training and six for testing in each fold. The COX database contains 3000 video sequences from 1000 subjects, which are captured by different camcorders. The face in each image was resized into 32×40 intensity image and histogram equalized. Similar with the protocol in [5], we conducted leave-one-out testing.

The MBGC database consists of 143 subjects walking toward a camera in a variety of illumination conditions, and the number of videos per subject ranges from 1 to 5. Following the similar settings in [26], we resized the face images to 100×100 and conducted leave-one-out testing.

The PaSC database consists of 2802 videos of 265 people carrying out simple actions. Verification experiments were conducted using control or handheld videos as target and query, respectively. Since the database is relatively difficult, we followed [27] to extract the state-of-the-art deep convolutional neural network (DCNN) features and all comparison methods are performed based on the DCNN features on PaSC. Here, the DCNN model was pretrained on the Celebrities on the Web (CFW) database [28] and subsequently fine-tuned on the training data of PaSC and COX database by using the Caffe [29].

B. Comparison With the State of the Art

1) Comparative Methods and Parameter Settings: To study the effectiveness of our proposed PDL method, we compared with several state-of-the-art image set classification methods, including affine/convex hull based image set distance (AHISD/CHISD) [1], sparse approximated nearest point (SANP) [10], regularized nearest points (RNP) [11], dual linear regression classification (DLRC) [4], pairwise linear regression classification (PLRC) [7] , and set-to-set distance metric learning (SSDML) [13].

The source code of all comparative methods released by the original authors were used except that of DLRC that is carefully implemented according to [4]. 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 [10]. Note that since the SANP method is too time consuming to run for COX, we alternately took the image sets of 100 persons rather than all the 1000 persons. For RNP, DLRC, and PLRC, all the parameters were configured according to [4], [7], [11], respectively. For SSDML, we set $\lambda_1 = 0.001$ and $\lambda_2 = 0.5$, the numbers of the positive pairs and the negative pairs per set are set to 10 and 20.

For our proposed PDL, we used the Principal Component Analysis (PCA) projection matrix as an initialization of W. Considering the varying numbers of images contained in different image sets, we did not set a fixed number m_c for prototypes of different image sets but employed the maximal linear patch (MLP) algorithm in [30] to compute the proper value of m_c and a stable initialization of P_c is accordingly calculated by the centers of local models. See Section III-C for more detailed experimental comparison and analysis of different m_c. For simplicity, we denote PDL without orthonormal projection constraint as PDL-NOP and use PDL-OP to represent PDL when constraining the projection to be orthonormal.

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

As can be seen in the results, our method performs the best on all of the four databases. First, our PDL achieves an impressively better result than the unsupervised affine-hull-based methods, such as AHISD, CHISD, SANP, RNP, and DLRC. This supports

TABLE I
COMPARISON WITH THE STATE OF THE ART

Method \ Dataset	YTC	COX	MBGC	PaSC	
				control	handheld
AHISD [1]	0.637	0.386	0.181	0.219	0.143
CHISD [1]	0.665	0.398	0.193	0.261	0.210
SANP [10]	0.684	0.494	0.247	0.232	0.157
RNP [11]	0.703	0.518	0.284	0.274	0.198
DLRC [4]	0.692	0.526	0.314	0.242	0.171
PLRC [7]	0.692	0.592	0.358	0.281	0.242
SSDML [13]	0.689	0.578	0.365	0.292	0.229
PDL-NOP	0.743	0.658	0.402	0.415	0.386
PDL-OP	0.761	0.692	0.443	0.447	0.401

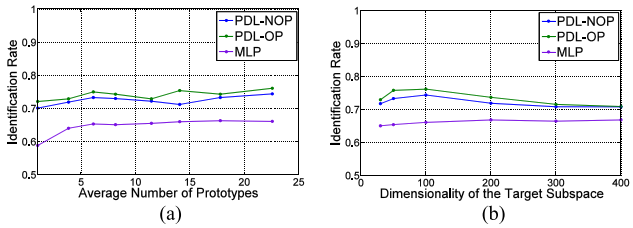

Fig. 2. (a) Comparison of different prototype numbers. (b) Effect of different dimensionalities of the target subspace.

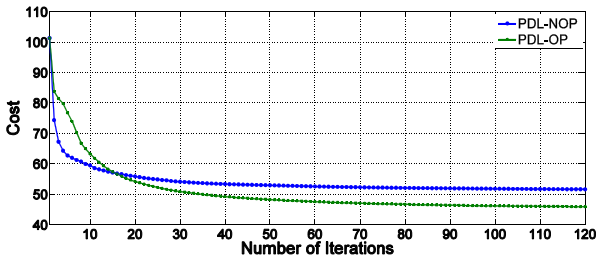

Fig. 3. Convergence of the optimization algorithm on YTC.

the motivation 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 to the geometric structure of affine hull or the selection criteria of nearest points. Second, our PDL is also superior over the supervised affine-hull-based method SSDML. It mainly attributes to our innovation in three aspects.

1) Learning virtual prototypes flexibly and efficiently.
2) Training from a local view of NN.
3) Solving with a joint optimization mechanism.

C. Evaluations of Main Factors

In this section, we conducted further experiments to evaluate main factors, which may affect the accuracy as well as the stability of the proposed PDL method.

1) Parameter Comparison: Experiments were performed to investigate the influence of the main parameters, i.e., the number of prototypes m_c for each image set and the dimension r of the target subspace. Fig. 2(a) illustrates the accuracy of PDL-NOP, PDL-OP, and the initial clustering obtained by MLP [30] according to the average number of prototypes in each set with $r = 100$. Note that different image sets may contain different numbers of prototypes and thus the average number of prototypes is not necessarily integer value. Then as shown in Fig. 2(b), we explore the effect of different dimensionalities of the target subspace with the average number of prototypes fixed to be about 23 and here the curve of MLP is obtained by performing MLP on the projected subspace of PCA. From these experiments, we can see that the proposed PDL method shows favorable stability with changing of the prototype number and the target subspace dimension.

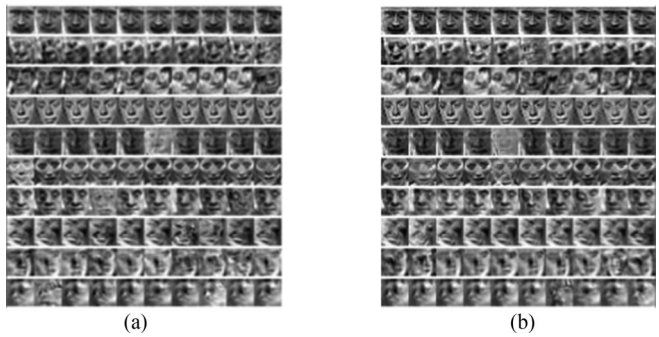

Fig. 4. Some examples for the learned prototypes. (a) PDL-NOP. (b) PDL-OP.

TABLE II
TIME COMPARISON (SECONDS) OF DIFFERENT METHODS ON YTC FOR TRAINING AND TESTING

Method	AHISD	CHISD	SANP	RNP	DLRC	PLRC	SSDML	PDL-NOP	PDL-OP
Training	N/A	N/A	N/A	N/A	N/A	N/A	346.33	75.30	84.00
Testing	1.58	1.71	56.77	1.56	1.91	2.12	2.35	1.15	1.15

2) Convergence: In Fig. 3, we illustrate the convergence of the optimization process in PDL-NOP and PDL-OP by taking the YTC database as an example. We can analyze from the practice that the values of the loss function both become stable in less than 100 iterations. From the same initialization, we experimentally find that PDL-NOP converges faster than PDL-OP and yet it reaches a suboptimal solution prematurely while PDL-OP perhaps avoids certain local minimizers with a further decline to lead to a better and more stable (local) solution. Besides, Fig. 4 gives some examples of the learned prototypes in different image sets on YTC.

3) Time Comparison: In addition, we compared the computational complexity of different methods on an Intel i7-3770, 3.40 GHz PC. Table II lists the time cost of the comparative methods for training and testing respectively on the YTC database. Note that only supervised methods need the training time. We can see that since training is performed offline, the online matching for PDL testing is very efficient and is faster than other affine-hull-based methods.

IV. CONCLUSION

This letter has proposed a PDL method for image set classification. We represented an image set by a prototype set learned from its affine hull to shrink the loose affine approximation effectively. 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. To enhance the stability and robustness of the target subspace, an orthogonality constraint is further studied to impose on the projection. Accordingly, a gradient descent algorithm is employed to solve such optimization problem. Our experimental evaluation has demonstrated the superiority of the proposed PDL on several challenging databases.

REFERENCES

[1] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2567–2573.

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

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

[4] L. Chen, "Dual linear regression based classification for face cluster recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2673–2680.

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

[6] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, "Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2048–2057.

[7] Q. Feng, Y. Zhou, and R. Lan, "Pairwise linear regression classification for image set retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4865–4872.

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

[9] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Joint face representation adaptation and clustering in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 236–251.

[10] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 121–128.

[11] M. Yang, P. Zhu, L. V. Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proc. IEEE Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–7.

[12] W. Wang, R. Wang, S. Shan, and X. Chen, "Probabilistic nearest neighbor search for robust classification of face image sets," in *Proc. IEEE Conf. Workshops Autom. Face Gesture Recognit.*, 2015, pp. 1–7.

[13] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2664–2671.

[14] M. Leng, P. Moutafis, and I. A. Kakadiaris, "Joint prototype and metric learning for set-to-set matching: Application to biometrics," in *Proc. IEEE Conf. Biometrics Theory, Appl. Syst.*, 2015, pp. 1–8.

[15] W. Wang, R. Wang, S. Shan, and X. Chen, "Prototype discriminative learning for face image set classification," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 344–360.

[16] R. Paredes and E. Vidal, "Learning prototypes and distances: A prototype reduction technique based on nearest neighbor error minimization," *Pattern Recognit.*, vol. 39, no. 2, pp. 180–188, 2006.

[17] R. Paredes and E. Vidal, "Learning weighted metrics to minimize nearest-neighbor classification error," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1100–1110, Jul. 2006.

[18] M. Villegas and R. Paredes, "Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[19] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 265–272.

[20] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, nos. 1/2, pp. 397–434, 2013.

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

[22] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian metric for point-to-set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1677–1684.

[23] P. J. Phillips *et al.*, "Overview of the multiple biometrics grand challenge," in *Proc. Int. Conf. Biometrics*, 2009, pp. 705–714.

[24] B. Ross *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE Conf. Biometrics Theory, Appl. Syst.*, 2013, pp. 1–8.

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

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

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

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

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

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