Simultaneous Feature and Dictionary Learning for Image Set Based Face Recognition

Jiwen Lu, *Senior Member, IEEE*, Gang Wang, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— In this paper, we propose a simultaneous feature and dictionary learning (SFDL) method for image set-based face recognition, where each training and testing example contains a set of face images, which were captured from different variations of pose, illumination, expression, resolution, and motion. While a variety of feature learning and dictionary learning methods have been proposed in recent years and some of them have been successfully applied to image set-based face recognition, most of them learn features and dictionaries for facial image sets individually, which may not be powerful enough because some discriminative information for dictionary learning may be compromised in the feature learning stage if they are applied sequentially, and vice versa. To address this, we propose a SFDL method to learn discriminative features and dictionaries simultaneously from raw face pixels so that discriminative information from facial image sets can be jointly exploited by a one-stage learning procedure. To better exploit the nonlinearity of face samples from different image sets, we propose a deep SFDL (D-SFDL) method by jointly learning hierarchical non-linear transformations and class-specific dictionaries to further improve the recognition performance. Extensive experimental results on five widely used face data sets clearly shows that our SFDL and D-SFDL achieve very competitive or even better performance with the state-of-the-arts.**

***Index Terms*— Face recognition, feature learning, dictionary learning, deep learning, image set classification.**

I. Introduction

IMAGE set based face recognition has attracted increasing interest in computer vision and pattern recognition in recent years [2], [5], [8]–[12], [14], [15], [18]–[20], [24], [27], [28], [33], [40], [45], [47], [50], [58], [67]–[69], [72], [77]. Unlike conventional image based face recognition systems where each training and testing example is a single still image, for image set based face recognition, each training and testing example contains a set of face images which were captured from different poses, illuminations, expressions and resolutions. While more information can be exploited to describe the identity information of persons from facial image sets, image set based face recognition is still challenging because there are large intra-class variations among face samples within each set, especially when they are captured in unconstrained environments.

There have been a number of studies on image set based face recognition over the past two decades [2], [5], [6], [9]–[12], [14], [15], [18]–[20], [23], [24], [27], [28], [33], [40], [45], [47], [50], [58], [67]–[69], [72], [77], and dictionary-based methods have achieved very competitive performance [9], [10] because variations of the pose, illumination and expression information within face image sets can be implicitly encoded into the learned dictionaries. However, most existing dictionary-based face recognition with image sets methods are unsupervised [9], [10], which are not discriminative enough to classify face sets. Moreover, these methods learn dictionaries using the original raw pixels, which may contain some noisy components that are irrelevant to dictionary learning. Since face images usually lie on a low-dimensional manifold, it is desirable to seek the most discriminative features in a low-dimensional subspace and suppress the useless information to promote learning robust and discriminative dictionaries for facial image sets.

In this paper, we propose a new simultaneous feature and dictionary learning (SFDL) method for image set based face recognition, where the basic idea is illustrated in Fig. 1. The goal of our SFDL is to jointly learn a feature projection matrix and a structured dictionary, where each frame within a set is projected into a low-dimensional subspace and encoded with a discriminative coding coefficient, and face image sets from each person are represented as a sub-dictionary so that person-specific dictionaries can be learned to extract more discriminative information, simultaneously. To better model the nonlinearity of face samples, we develop a deep SFDL (D-SFDL) method by jointly learning multiple hierarchical non-linear transformations and class-specific dictionaries to further improve the recognition performance. Extensive experimental results on five widely used face datasets clearly show that our SFDL and D-SFDL achieve very competitive or even better performance than state-of-the-art image set based face recognition methods.

This paper is an extended version of [44]. New contributions include the newly proposed D-SFDL method, more analysis of the proposed methods including convergence analysis and

Manuscript received March 15, 2016; revised October 3, 2016 and December 27, 2016; accepted June 6, 2017. Date of publication June 8, 2017; date of current version June 23, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61225008, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shiguang Shan. *(Corresponding Author: Jiwen Lu.)*

J. Lu and J. Zhou are with the Department of Automation, State Key Laboratory of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

G. Wang is with the Alibaba AI Labs, Hangzhou 310052, China (e-mail: gangwang6@gmail.com).

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

Digital Object Identifier 10.1109/TIP.2017.2713940

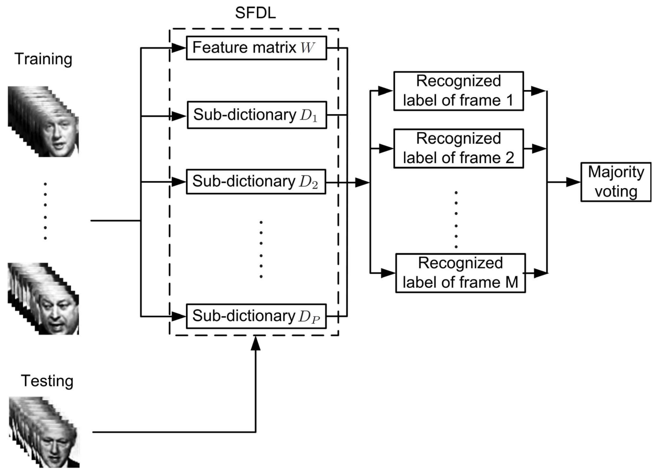


Fig. 1. The basic idea of our proposed SFDL approach to image set based face recognition, where discriminative features and dictionaries are learned simultaneously to encode the pose, illumination and expression information in face image sets, so that it is more robust to noise. In the training stage, we learn a feature projection matrix W and a structured dictionary $D = [D_1, D_2, \cdots, D_P]$ (one sub-dictionary per class) by using the proposed SFDL method, where P is the number of subjects in the training set. Given a testing face image set containing M image frames, we first apply the learned feature projection matrix W to project each sample into a feature and recognize its label by using the smallest reconstruction error corresponding to the associated sub-dictionary. Lastly, the majority voting strategy is used to classify the whole testing face image set.

parameter settings, comparisons with more state-of-the-art image set based face recognition methods, and experimental evaluation on more challenging video face datasets.

II. Related Work

In this section, we briefly review three related topics: 1) feature learning, 2) dictionary learning, and 3) image set based face recognition.

A. Feature Learning

Feature learning has been widely used in computer vision and a variety of feature learning methods have been proposed in recent years [3], [21], [26], [30], [43], [55]. Unlike most existing hand-crafted feature representation methods which are heuristics and require strong prior knowledge, feature learning methods can learn representations from raw data directly. Representative feature learning methods are restricted Boltzmann machine [21], sparse auto-encoders [3], independent subspace analysis [30], denoising auto-encoders [55], reconstruction independent component analysis [37], convolutional neural networks [26], and recurrent convolutional neural networks [52]. These methods have achieved many state-of-the-art performance in various visual analysis tasks such as human detection [53], action recognition [38], image classification [37], visual tracking [13], and person re-identification [42].

Recently, feature learning has also been successfully used in face recognition and achieved encouraging performance. For example, Cao *et al.* [4] proposed a feature learning approach by clustering neighboring face pixels into different groups to extract histogram features under a bag-of-word framework. Hussain *et al.* [29] proposed a local quantized pattern method by modifying the local binary pattern method with a codebook learning approach. Lei *et al.* [41] presented a discriminant face descriptor method by learning discriminative image filters. Lu *et al.* [43] proposed a compact binary face descriptor method by employing learning to hashing over local face patches to obtain histogram features for face images. Compared to hand-crafted face descriptors, feature learning methods usually achieve better performance in face recognition because it is data-driven and can automatically exploit more high-level information.

More recently, a variety of deep learning methods [26], [54], [57], [59]–[63] have been proposed for face feature learning, and some of them have achieved excellent performance in many face recognition systems. For example, Huang *et al.* [26] proposed a convolutional deep belief network approach for hierarchical feature learning for face verification. Sun *et al.* [59]–[62] presented several deep learning based feature learning methods for unconstrained face representation and recognition. Taigman *et al.* [63] introduced a DeepFace method which learn supervised face representation with 4,000,000 labeled face samples by using the deep convolutional neural networks. Parkhi *et al.* [63] and Schroff *et. al.* [57] presented two deep convolutional neural networks methods by learning a very deep network with the triplet loss function, where 2.6M and 200M labeled face images were used to train the deep model, respectively.

B. Dictionary Learning

There have been extensive work on dictionary learning in the literature [1], [31], [35], [46], [48], [64], [70], [73], [76]. Dictionary learning aims to seek a collection of atoms for sparse representation of the input samples, where each data is linearly represented by a small number of atoms. Existing dictionary learning methods can be mainly classified into two categories: unsupervised [1] and supervised [35], [73]. In recent years, dictionary learning has been extensively used in face recognition and also shown good performance [35], [51], [73], [75]. However, most existing dictionary learning methods have been developed for image based face recognition and little progress has been made for image set based face recognition. More recently, Chen *et al.* [9] presented a discretionary learning method for video-based face recognition, where each face video is first clustered into several clusters and then the dictionary is learned for each cluster. However, their method is unsupervised, which may not be discriminative enough for classification.

C. Image Set Based Face Recognition

Over the past recent years, we have witnessed a considerable interest in developing new methods for image set based face recognition [2], [5], [7], [9]–[12], [14], [15], [18]–[20], [24], [27], [28], [33], [40], [45], [47], [49], [50], [56], [58], [67]–[69], [72], [77]. These methods can be mainly categorized into two classes: parametric and non-parametric. Parametric methods first model each face image set as a distribution function and then compute the divergence between two distributions as the similarity of two face image sets. The key shortcoming of parametric methods is that if there are not strong correlations between two face image sets, the estimated model cannot well characterize the sets and may fail to measure their similarity.

Non-parametric methods represent each face image set as a single or mixture of linear subspaces, and then use the subspace distance to measure the similarity of face image sets. Representative subspace distance methods include principal angle [22], affine/convext hull similarity [5], and nearest points distance [25]. Recently, Chen *et al.* [9], [10] presented a dictionary-based approach for image set based face recognition by building one dictionary for each class of face image sets and using these dictionaries to measure the similarity of face image sets. While reasonably good recognition performance can be obtained, their approach is a generative model and the dictionaries are learned from the original raw pixels, which may contain some noisy and irrelevant components. In this work, we model face image sets as discriminative dictionaries with the learned feature space, so that more discriminative information can be exploited.

III. Proposed Methods

In this section, we first introduce the proposed SFDL method, and detail the extended D-SFDL. Then, we show how to use our SFDL and D-SFDL methods for image set based face recognition. Lastly, we discuss the key difference between our proposed SFDL approach and several recently proposed image set based face recognition methods.

A. SFDL

Generally, there are two key stages in a practical image set based face recognition system [12], [24], [25], [28], [45], [50], [67]: image set representation and image set matching. Previous work [66] has shown that feature learning is an effective tool for image set representation because it can extract discriminative information from face image sets. Recent study [9] has also shown that dictionary learning is a promising solution to image set matching because face images with varying poses, illuminations and expressions within a set can be encoded as dictionaries so that the noise can be effectively alleviated and better matching performance can be obtained. However, most previous image set based face recognition methods learned features and dictionaries individually, which may not be powerful enough because some discriminative information for dictionary learning may be compromised in the feature learning stage, and vice versa. This is because the objective of feature learning is usually inconsistent to that of dictionary learning because feature learning is essentially a feature selection problem while dictionary learning is intrinsically a clustering problem. Hence, it is suboptimal to apply feature learning and dictionary learning for image set based face recognition. To address this shortcoming, we propose a SFDL method to learn discriminative features and dictionaries simultaneously in the following.

Let $X = [X_1, X_2, \cdots, X_P]$ be the training set of face image sets from P different subjects. Assume there are N images in total in the training set by concatenating all the frames from image sets, we rewrite X as $X = [x_1, x_2, \cdots x_N] \in R^{d \times N}$, where x_i is a d-dimensional vector of the cropped face image. To extract more discriminative and robust information from the training set, SFDL aims to simultaneously learn a feature projection matrix and a discriminative structured dictionary to project each image frame in all image sets into a low-dimensional subspace, under which each image frame is encoded by a discriminative coding coefficient. To achieve this, we formulate the following optimization problem:

$$\begin{aligned}\min_{W,D,A} J &= J_1 + \lambda_1 J_2 + \lambda_2 J_3 \\ &= \sum_{i=1}^{N} (\|Wx_i - D\alpha_i\|_2^2 + \eta_1 \|\alpha_i\|_1) \\ &\quad + \sum_{p=1}^{P} \sum_{i=1}^{N_p} \|Wx_{ip} - D_p \alpha_{ip}^p\|_2^2 \\ &\quad + \lambda_1 \sum_{i=1}^{N} (\|W^T W x_i - x_i\|_2^2 + \eta_2 \sum_{j=1}^{k} h(W_j x_i)) \\ &\quad + \lambda_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \|\alpha_i - \alpha_j\|_2^2 S_{ij} \qquad (1)\end{aligned}$$

where $W \in R^{q_1 \times d}$ is the feature projection matrix, q_1 is the dimension of the learned feature space, $D = [D_1, D_2, \cdots, D_P] \in R^{q_1 \times c}$ is the structured dictionary, $D_p \in R^{q_1 \times c_p}$ is the sub-dictionary for the pth class in D, c and c_p are the number of atoms in the learned dictionary D and sub-dictionary D_p, $\alpha_i \in R^c$ is the representation coefficient vector for x_i, $c = \sum_{p=1}^{P} c_p$, $x_{ip} \in R^d$ is the ith raw pixel sample from the pth class, N_p is the number of samples in the pth class, $A = [\alpha_1, \alpha_2, \cdots, \alpha_N] \in R^{c \times N}$ is the sparse representation of the training samples in X, α_i is the coefficient vector of x_i, $\alpha_{ip}^p \in R^{c_p}$ is the representation coefficient vector of α_i from the pth class, h is a nonlinear convex function which is defined as a smooth l_1 penalty: $h(\cdot) = \log(\cosh(\cdot))$ [37], λ_1, λ_2, η_1 and η_2 are four parameters to balance the importance of different terms, $S \in R^{N \times N}$ is an affinity matrix to measure the similarity of the sparse codes α_i and α_j according to their label and appearance information, which is defined as follows:

$$S_{ij} = \begin{cases} 1, & \text{if } x_i \in N_{k_1}(x_j) \text{ or } x_j \in N_{k_1}(x_i) \\ & \text{and } l(x_i) = l(x_j) \\ -1, & \text{if } x_i \in N_{k_2}(x_j) \text{ or } x_j \in N_{k_2}(x_i) \\ & \text{and } \ l(x_i) \neq l(x_j) \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

where $N_k(x_i)$ and $l(x_i)$ denote the k-nearest neighbors and the label of x_i, respectively.

The first term J_1 in (1) is to enforce that for each face sample x_i from the pth class in the low-dimensional feature subspace, it is not only well reconstructed by the whole dictionary D, but also the sub-dictionary D_p of the pth class.

The second term J_2 in (1) is to enforce that the feature projection matrix W can preserve the energy of each x_i as much as possible and each column in W is to be as sparse as possible. The third term J_3 in (1) is to enforce that the difference of the sparse codes of two face images is minimized if they are from the same class and look similar, and the difference of the sparse codes of two face images is maximized if they are from different classes and also look similar, such that discriminative information can be discovered when learning sparse representation coefficients.

We rewrite A as $A = [A_1, A_2, \cdots, A_P] \in R^{c \times N}$, where $A_p \in R^{c \times N_p}$ denotes the sub-matrix from the pth class containing the coding coefficients of $X_p \in R^{d \times N_p}$ over D. Let $A_p^p \in R^{c_p \times N_p}$ be the coding coefficient of X_p over the sub-dictionary D_p. Then, J_1 in (1) can be re-written as follows:

$$\begin{aligned} J_1 &= \sum_{p=1}^{P} (\|WX_p - DA_p\|_F^2 + \|WX_p - D_p A_p^p\|_F^2) \\ &\quad + \eta_1 \|A\|_1 \\ &= \sum_{p=1}^{P} G_p(W, X_p, D, A_p) + \eta_1 \|A\|_1 \end{aligned} \tag{3}$$

where

$$\begin{aligned} G_p(W, X_p, D, A_p) &\triangleq \|WX_p - DA_p\|_F^2 \\ &\quad + \|WX_p - D_p A_p^p\|_F^2 \end{aligned} \tag{4}$$

We can also simplify J_2 and J_3 in (1) as follows:

$$J_2 = \|W^T WX - X\|_2^2 + \eta_2 H(WX) \tag{5}$$

$$J_3 = tr(A^T CA) - tr(A^T SA) = tr(A^T LA) \tag{6}$$

where $H(Z)$ is the sums of the outputs of the nonlinear convex function h which is applied on all elements in the matrix Z, $C = diag\{c_1, c_2, \cdots, c_N\} \in R^{N \times N}$ is a diagonal matrix whose diagonal elements are the sums of the row elements of $S \in R^{N \times N}$, and $L = C - S$, $L \in R^{N \times N}$.

Combining (4)-(6) into (1), we have the following SFDL model:

$$\begin{aligned} \min_{W,D,A} J &= \sum_{p=1}^{P} G_p(W, X_p, D, A_p) + \eta_1 \|A\|_1 \\ &\quad + \lambda_1 (\|W^T WX - X\|_2^2 \\ &\quad + \eta_2 H(WX)) + \lambda_2 tr(A^T LA) \end{aligned} \tag{7}$$

While the objective function in (7) is not convex for W, D and A simultaneously, it is convex to one of them when the other two are fixed. Following the work in [31], [73], [17], [35], we iteratively optimize W, D and A by using the following three-stage method.

Step 1 (Learn W With Fixed D and A): when D and A are fixed, (7) can be rewritten as

$$\begin{aligned} \min_{W} J &= \sum_{p=1}^{P} (\|WX_p - DA_p\|_F^2 + \|WX_p - D_p A_p^p\|_F^2) \\ &\quad + \lambda_1 (\|W^T WX - X\|_2^2 + \eta_2 H(WX)) \end{aligned} \tag{8}$$

(8) is an unconstrained optimization problem and many existing fast unconstrained optimizers can be applied to solve this problem. In our implementations, we use the conjugate gradient decent method in [37] to get W.

Step 2 (Learn A With Fixed W and D): when W and D are fixed, (7) can be rewritten as

$$\begin{aligned} \min_{A} J &= \sum_{p=1}^{P} (\|Y_p - DA_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2) \\ &\quad + \eta_1 \|A\|_1 + \lambda_2 tr(A^T LA) \end{aligned} \tag{9}$$

where $Y_p = WX_p \in R^{d_1 \times N_p}$ is the projection of X_p in the feature space. We compute A_p sequentially by fixing the other coefficient matrices A_q ($q \neq p$, and $1 \leq q \leq P$). Then, (9) can be simplified as

$$\begin{aligned} \min_{A_p} J &= \|Y_p - DA_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2 \\ &\quad + \eta_1 \|A_p\|_1 + \lambda_2 tr(A_p^T LA_p) \end{aligned} \tag{10}$$

Following the work in [39], we optimize each α_{ip} in A_p alternatively. To obtain each α_{ip}, we fix the encoding coefficients α_{jp} ($j \neq i$) for other samples, and rewrite (10) as

$$\begin{aligned} \min_{\alpha_{ip}} J &= \|Y_p - D\alpha_{ip}\|_F^2 + \|Y_p - D_p \alpha_{ip}^p\|_F^2 \\ &\quad + \eta_1 \|\alpha_{ip}\|_1 + \lambda_2 F(\alpha_{ip}) \end{aligned} \tag{11}$$

where

$$F(\alpha_{ip}) = \lambda_2 (\alpha_{ip}^T (A_p L_i) + (A_p L_i)^T \alpha_{ip} - \alpha_{ip}^T L_{ii} \alpha_{ip}) \tag{12}$$

$L_i \in R^N$ is the ith column of L, and L_{ii} is the entry in the ith row and ith column of L. We apply the feature sign search algorithm [39] to solve α_{ip}.

Step 3 (Learn D With Fixed W and A): when W and A are fixed, (7) can be rewritten as

$$\min_{D} J = \sum_{p=1}^{P} (\|Y_p - DA_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2) \tag{13}$$

We update D_p sequentially by fixing the other sub-dictionaries D_q ($q \neq p$, and $1 \leq q \leq P$). Then, (13) can be reduced as

$$\min_{D_p} J = \|Y_p - D_p A_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2 \tag{14}$$

We restrict that each column d_p^i in D_p is a unit vector, where $1 \leq i \leq K_p$, K_p is the number of atoms in D_p. (14) is a quadratic programming problem and can be solved by using the algorithm in [74], which updates D_p atom by atom.

We repeat the above steps steps until the algorithm converges, and summarize the proposed SFDL in **Algorithm 1**.

B. D-SFDL

Since SFDL only learns a linear dictionary for image set representation, it is not powerful enough to model the nonlinearity of face samples, especially when face images are captured in unconstrained environments. Even if some nonlinear operation such as the major voting strategy may be employed to make use of the nonlinearity of samples in the final decision procedure, it is still desirable to exploit such nonlinearity information at both the dictionary learning and final decision stages so that the nonlinearity of face samples

```
Algorithm 1 SFDL
Input: Training set $X = [X_1, X_2, \cdots, X_P]$, affinity
       matrix $S$, parameters $\lambda_1$, $\lambda_2$, $\eta_1$ and $\eta_2$, iteration
       number $T$, convergence error $\epsilon$.
Output: Feature weighting matrix $W$, dictionary $D$, and
        coding coefficient matrix $A$.
Step 1 (Initialization):
    1.1: Initialize each column $d_p^i$ in $D_p$ as a random vector
         with unit l2-norm.
    1.2: Initialize each column in $A$ as a random vector.
Step 2 (Local optimization):
    For $t = 1, 2, \cdots, T$, repeat
    2.1. Solve $W^t$ with fixed $D^{t-1}$ and $A^{t-1}$ via (8).
    2.2. Solve $A^t$ with fixed $W^t$ and $D^{t-1}$ via (11).
    2.3. Solve $D^t$ with fixed $W^t$ and $A^t$ via (14).
    2.3. If $|D^t - D^{t-1}| < \epsilon$ or $|W^t - W^{t-1}| < \epsilon$
         and $t > 2$, go to Step 3.
Step 3 (Output):
    Output $W = W^t$, $D = D^t$, and $A = A^t$.
```

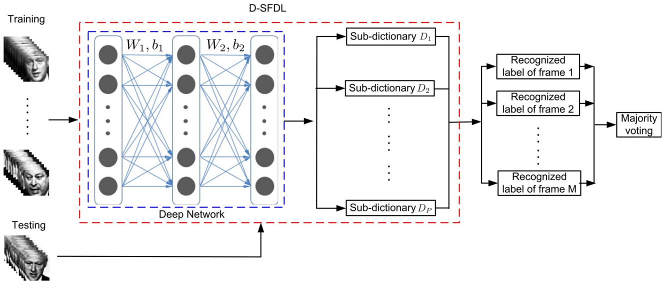


Fig. 2. The basic idea of our proposed D-SFDL approach to image set based face recognition, where multiple hierarchical non-linear transformations and class-specific dictionaries are simultaneously learned. In the training stage, we learn multiple hierarchical non-linear transformations and class-specific dictionaries by using the proposed D-SFDL method. Given a testing face image set containing M image frames, we first apply the learned multiple hierarchical non-linear transformations to project each sample into a feature and recognize its label by using the smallest reconstruction error corresponding to the associated sub-dictionary. Lastly, the majority voting strategy is used to classify the whole testing face image set.

within image sets can be better modeled and exploited. To address this, most previous dictionary learning methods [1], [35], [70], [73] employed the kernel methods to learn kernel-based dictionaries of samples to handle non-linearities in the data samples, so that the class identity in the face dataset can be better separated in the kernelized nonlinear space. However, kernel-based methods implicitly map samples into a high-dimensional feature space and then learn a dictionary in the high-dimensional space, which cannot explicitly obtain the nonlinear mapping functions and usually suffer from the scalability problem. Unlike these methods, we develop a deep SFDL (D-SFDL) method by jointly learning multiple hierarchical non-linear transformations and class-specific dictionaries to further improve the recognition performance. Specifically, D-SFDL uses a feed-forward neural network to exploit the nonlinearity of samples for simultaneous feature and dictionary learning. Fig. 2 illustrates the basic idea of the proposed D-SFDL method.

Unlike SFDL which only learns one linear projection matrix W, our D-SFDL constructs a feed-forward neural network to represent each face image with multiple layers of nonlinear transformations. Assume there are $M+1$ layers in our deep network, and the mth layer contains $p^{(m)}$ units, where $m = 1, 2, \cdots, M$. For each face sample $x_i \in R^d$, the output at the first layer is $h_{x_i}^{(1)} = s(W_1 x_i + b_1) \in R^{p_1}$, where $W_1 \in R^{p_1 \times d}$ is a projection mapping in the first layer, $b_1 \in R^{p_1}$ is a bias vector, and s is an activation function, p_1 is the feature dimension of the output of the first layer. The output of the first layer $h_{x_i}^{(1)}$ is employed as the input of the second layer. Similarly, the output of the second layer can be computed as $h_{x_i}^{(2)} = s(W_2 h_{x_i}^{(1)} + b_2) \in R^{p_2}$, where $W_2 \in R^{p_2 \times p_1}$ is a projection matrix to be learned in the second layer, $b_2 \in R^{p_2}$ is a bias vector, and p_2 is the feature dimension of the output of the second layer, respectively. Similarly, the output of the mth layer is $h_{x_i}^{(m)} = s(W_m h_{x_i}^{(m-1)} + b^m) \in R^{p^{(m)}}$, where $W_m \in R^{p_{m+1} \times p_m}$ is a projection matrix to be learned in the mth layer, $b_m \in R^{p_m}$ is a bias vector, and p_m is the feature dimension of the output of the mth layer, respectively.

Assume $\Omega = \{W_1, b_1, \cdots, W_M, b_M\}$ be the parameter set of the designed network, we formulate the D-SFDL method as the following optimization problem:

$$\begin{aligned}\min_{\Omega, D, A} \; O &= O_1 + \lambda_1 O_2 + \lambda_2 O_3 \\ &= \sum_{i=1}^{N} (\| h_{x_i}^{(M)} - D\alpha_i \|_2^2 + \eta_1 \|\alpha_i\|_1) \\ &\quad + \sum_{p=1}^{P} \sum_{i=1}^{N_p} \| h_{x_{ip}}^{(M)} - D_p \alpha_{ip}^p \|_2^2 \\ &\quad + \lambda_1 \sum_{m=1}^{M} \left(\|W_m\|_F^2 + \|b_m\|_2^2 \right) \\ &\quad + \lambda_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \|\alpha_i - \alpha_j\|_2^2 S_{ij} \qquad (15)\end{aligned}$$

where $h_{x_i}^{(M)} \in R^{p_{M+1}}$, $D \in R^{p_{M+1} \times c}$, and $D_p \in R^{p_{M+1} \times c_p}$.

The first term O_1 in (15) is to ensure that for each face sample x_i from the pth class, it is not only well reconstructed by the whole dictionary D which is learned at the last layer of our network, but also the sub-dictionary D_p of the pth class. The second term O_2 in (15) represents the regularizer of the projection matrix and bias. The third term O_3 in (15) is to ensure that the difference of the sparse codes of two face images at the last layer of the network is minimized if they are from the same class and look similar, and the difference of the sparse codes of two face images at the last layer of the network is maximized if they are from different classes and also look similar, so that discriminative information can be discovered when learning these sparse representation coefficients.

The optimization problem in (15) is not convex with respect to Ω, D, and A. Hence, it is non-trivial to obtain a global optimal solution. Alternatively, we use an iterative method to obtain a local solution which alternately optimizes Ω, D, and A.

Step 1 (Learn Ω With Fixed D and A): when D and A are fixed, (15) can be rewritten as

$$\min_{\Omega} O = \sum_{p=1}^{P} (\|H_p^{M+1} - DA_p\|_F^2 + \|H_p^{M+1} - D_p A_p^p\|_F^2) + \lambda_1 \sum_{m=1}^{M} \left(\|W_m\|_F^2 + \|b_m\|_2^2 \right) \tag{16}$$

where $H_p^{M+1} \in R^{p_{M+1} \times N}$ is the matrix form of the output of the training samples at the last layer of our network.

We use the back-propagation method with the batch gradient descent algorithm to update the parameter Ω. Specifically, W_m and b_m can be updated as follows:

$$W_m = W_m - \beta \frac{\partial O}{\partial W_m}, \tag{17}$$

$$b_m = b_m - \beta \frac{\partial O}{\partial b_m}, \tag{18}$$

where β is the learning rate controlling the convergence speed.

Step 2 (Learn A With Fixed Ω and D): when Ω and D are fixed, (15) can be rewritten as

$$\min_{A} O = \sum_{p=1}^{P} (\|H_p^{M+1} - DA_p\|_F^2 + \|H_p^{M+1} - D_p A_p^p\|_F^2) + \eta_1 \|A\|_1 + \lambda_2 tr(A^T LA) \tag{19}$$

We use the same optimization strategy in (11) of SFDL to obtain A.

Step 3 (Learn D With Fixed Ω and A): when Ω and A are fixed, (15) can be rewritten as

$$\min_{D} O = \sum_{p=1}^{P} (\|H_p - DA_p\|_F^2 + \|H_p - D_p A_p^p\|_F^2) \tag{20}$$

Similarly, we obtain D by sequentially updating D_p with other fixed sub-dictionaries D_q ($q \neq p$, and $1 \leq q \leq P$), and obtain D_p with the same optimization strategy in (20) of SFDL.

We repeat the above three steps until the algorithm is convergent. The proposed D-SFDL algorithm is summarized in **Algorithm 2**.

Algorithm 2 D-SFDL

```
Input: Training set $X = [X_1, X_2, \cdots, X_P]$, affinity
    matrix $S$, parameters $\lambda_1$, $\lambda_2$, $\eta_1$, $\eta_2$, $\beta$, and $M$,
    number of nodes in each layer of the network $p_1$,
    $p_2$, $\cdots$, $p_M$, iteration number $T$, convergence
    error $\epsilon$.
Output: Network parameter $\Omega$, dictionary $D$, and coding
    coefficient matrix $A$.
Step 1 (Initialization):
    1.1: Initialize each column $d_p^i$ in $D_p$ as a random vector
        with unit l2-norm.
    1.2: Initialize each column in $A$ as a random vector.
    1.3: Initialize $b_m$ as 0 and $W_m$ as a random matrix.
Step 2 (Local optimization):
    For $t = 1, 2, \cdots, T$, repeat
    2.1. Solve $\Omega^t$ with fixed $D^{t-1}$ and $A^{t-1}$ via
        (16)-(18).
    2.2. Solve $A^t$ with fixed $\Omega^t$ and $D^{t-1}$ via (11).
    2.3. Solve $D^t$ with fixed $\Omega^t$ and $A^t$ via (14).
    2.3. If $t > 1$ and $|O_t - O_{t-1}| < \varepsilon$, go to Step 3.
Step 3 (Output):
    Output $\Omega = \Omega^t$, $D = D^t$, and $A = A^t$.
```

C. Recognition

In this subsection, we show how to use SFDL and D-SFDL for image set based face recognition. Given a testing face video $X^q = [x_1^q, x_2^q, \cdots, x_n^q] \in R^{d \times n}$, where x_j^q is the jth ($1 \leq j \leq n$) frame of this video and n is the number of image frames in this video, we first apply the learned feature projection matrix W in SFDL or the parameter set Ω in D-SFDL to project each frame x_j^q in this video into a feature and recognize its label by using the smallest reconstruction error corresponding to each sub-dictionary D_p ($q \leq p \leq P$), which is computed as follows:

$$\text{SFDL}: \quad p' = \arg\min_p \|Wx_j^q - D_p D_p^{\dagger} x_j^q\|_2 \tag{21}$$

$$\text{D-SFDL}: \quad p' = \arg\min_p \|h_{x_j^q} - D_p D_p^{\dagger} h_{x_j^q}\|_2 \tag{22}$$

where $D_p^{\dagger} = (D_p^T D_p)^{-1} D_p^T$ is the pseudo-inverse of D_p, $h_{x_j^q}$ is the last layer output of the x_j^q in the deep network of D-SFDL.

Then, we adopt the majority voting strategy to classify the whole testing face video:

$$p^* = \arg\max_p Z_p \tag{23}$$

where Z_p is the total number of votes from the pth class.

IV. Experimental Results

We evaluated our proposed methods on five publicly available face databases including the Honda [40], Mobo [16], YouTube Celebrities (YTC) [32], PubFig [36], and IJB-A [34] datasets.

A. Datasets

The Honda dataset [40] contains 59 face videos of 20 subjects, where there are large pose and expression variations and the average length of these videos are approximately 400 frames.

The MoBo dataset [16] contains 96 videos from 24 subjects. For each subject, four videos corresponding to different walking patterns on a treadmill such as slow, fast, inclined and carrying a ball were captured and each video corresponds to one walking pattern. For each video, there are around 300 frames covering pose and expression variations.

The YTC dataset [32] contains 1910 video sequences of 47 celebrities (actors, actresses and politicians) which are collected from YouTube. Most videos are low resolution which leads to noisy and low-quality image frames. The number of frames for these videos varied from 8 to 400.

The PubFig dataset [36] contains 58797 face images of 200 persons, which were captured in the wild from the internet, so that there are also large variations of pose, illumination, and expression on these images.

The IJB-A dataset [34] contains 2085 face videos from 500 subjects, where each subject has at least one video and 4.2 videos per subject on average are available. The original videos were not provided in this dataset, and the sampled frames from the videos were provided. The pose and wide variations of face images in this dataset are much larger than other datasets such as PubFig and YTF so that face recognition on this dataset is more challenging. In general, each image set contains approximately 10 samples.

For face videos in the Honda, Mobo and YTC datasets, each image frame was first automatically detected by applying the face detector method proposed in [65] and then resized to a 30×30 intensity image. For face images in the PubFig dataset, we cropped face region of each face image according to the provided bounding box position, and resized it into 30×30. Face videos in the IJBA dataset were detected and aligned with the provided results from the original datasets, where each face image was also cropped and resized into 30×30 according to the provided eye coordinates. Hence, each video is represented as an image set. For each image frame in all these these datasets, we only performed histogram equalization to remove the illumination effect.

B. Experimental Settings

To make a fair comparison with state-of-the-art image set based face recognition methods, we followed the same protocol used in [5], [25], [66]–[68], [71], On the Honda, MoBo, and YTC datasets, we conducted experiments 10 times by randomly selecting training and testing sets, computed and compared the average identification rate. For both the Honda and MoBo datasets, we randomly selected one face video per person to construct the training set and used the remaining videos as the testing set. For the YTC dataset, we equally divided the whole dataset into five folds (with minimal overlapping), and each fold contains 9 videos for each person. For each fold, we randomly selected 3 face videos for each person for training and used the remaining 6 for testing. For the PubFig dataset, we divided face images of each person into three folds with an equal size, where three different image sets were constructed for each person. We used one fold for training and the remaining two for testing. For the IJB-A and PaSC datasets, we followed their standard protocols and computed the verification rates for different methods.

We used the cross-validation strategy to select and determine these parameters in our experiments. In our implementations, we tuned the parameters in a training set which combines the training sets of the Honda, Mobo, YTC, and PubFig datasets. Specifically, we randomly split the training set into 10 folds where each fold contains the same number of image sets. We selected 1 fold as the validation set and employed the other 9 folds to train the model, where the validation set was used to determine the parameters of our SFDL and D-SFDL. The parameters which yielded the best recognition performance were selected for the testing sets. Since there are multiple parameters to be tuned and it is generally difficult to determine them simultaneously, we employed a stepwise selection strategy. Specifically, we first pre-fixed other parameters and seek the optimal value for one parameter and then update these parameters with the newly learned parameters. These parameters were determined when the recognition performance reached to a stable rate. Finally, the feature dimension of W and the parameters λ_1, λ_2, η_1 and η_2 of our SFDL were empirically specified as 200, 1, 1, 0.05, and 0.2, respectively, and the number of atoms per person (K_p) for different datasets are summarized in Table I. For the D-SFDL method, we employed a three-layer feed-forward network ($M = 2$) and the threshold τ, learning rate μ and regularization parameter λ were empirically set as 3, 10^{-3}, and 10^{-2} for all experiments, respectively. The number of nodes for these three layers were set as 400, 200 and 100, respectively. For each layer, we used a nonlinear active function to nonlinearly map the output of the previous layer to another range space. In our experiments, we used the *tanh* function as the activation function because we empirically found it achieved better performance than others in our work. The *tanh* function and its derivative are computed as follows:

TABLE I

SUMMARY OF NUMBER OF ATOMS PER PERSON (K_p) FOR DIFFERENT FACE DATASETS IN OUR EXPERIMENTS

Dataset	Honda	MoBo	YTC	PubFig	IJB-A
K_p	20	25	35	40	40

$$s(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{24}$$

$$s'(x) = \tanh'(x) = 1 - \tanh^2(x). \tag{25}$$

The initializations of W_m and b_m ($1 \leq m \leq M$) are important to our D-SFDL method because we use the gradient descent method to seek the parameters. In this work, we used a normalized random initialization method, where the bias b_m is initialized as $\mathbf{0}$, and the weight W_m is initialized as follows:

$$W_m \sim U\left[-\frac{\sqrt{6}}{\sqrt{p_m + p_{m-1}}}, \frac{\sqrt{6}}{\sqrt{p_m + p_{m-1}}}\right], \tag{26}$$

where $p^{(0)}$ is the dimension of the input layer, $1 \leq m \leq M$.

C. Results and Analysis

1) Comparison With Existing State-of-the-Art Image Set Based Face Identification Methods: We compared our approach with twelve state-of-the-art image set based face identification methods, including Mutual Subspace Method (MSM) [72], Discriminant Canonical Correlation analysis (DCC) [33], Manifold-to-Manifold Distance (MMD) [68], Manifold Discriminant Analysis (MDA) [66], Affine Hull based Image Set Distance (AHISD) [5], Convex Hull based Image Set Distance (CHISD) [5], Sparse Approximated Nearest Point (SANP) [25], Covariance Discriminative Learning (CDL) [67], Dictionary-based Face Recognition from Video (DFRV) [9], Local Multi-Kernel Metric Learning (LMKML) [45], Projection Metric Learning (PML) [27], and Discriminative Analysis on Riemannian manifold of Gaussian distributions (DARG) [69]. Since some of these methods were not tested on some face datasets which were used in our

TABLE II

AVERAGE RANK-ONE IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
MSM [72]	92.5*	85.5*	61.5*	57.4*	18.6*
DCC [33]	94.9*	88.1*	64.8*	45.5*	19.8*
MMD [68]	96.9	93.6	66.7*	46.3*	20.2*
MDA [66]	100.0	96.5	67.2	48.6*	20.6*
AHISD [5]	97.4	92.7	66.5*	62.1*	19.4*
CHISD [5]	94.9	94.2	67.4*	64.5*	19.2*
SANP [25]	**100.0**	97.1	65.0	78.5*	20.4*
CDL [67]	98.0	86.7	67.5	65.5*	21.6*
DFRV [9]	97.4	94.4*	74.5*	74.5*	23.6*
LMKML [45]	98.5	96.3*	78.2	76.7	22.2
PML [27]	**100.0***	96.3*	70.3	76.7	23.6
DARG [69]	**100.0***	96.3*	77.1	76.7	24.7
SFDL	**100.0**	96.7	76.7	78.5	26.6
D-SFDL	**100.0**	**98.5**	**79.5**	**83.5**	**28.2**

*The results with ∗ were from the original papers and those without ∗ were obtained by our re-implementations. Moreover, our results on IJB-A cannot be compared with those in [7], [49] because their experimental settings are different from ours.

TABLE III

AVERAGE RECOGNITION RATES (%) OF DIFFERENT METHODS ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
SFDL	98.3	94.1	74.3	78.5	26.6
D-SFDL	100.0	98.5	79.5	83.5	28.2
CNN+DL	100.0	100.0	86.3	84.6	36.2
CNN+SFDL	100.0	100.0	88.8	86.7	38.6
CNN+D-DL	100.0	100.0	89.6	88.2	39.4
CNN+D-SFDL	100.0	100.0	90.4	89.4	40.2

experiments, we provided our evaluations for these methods with the standard implementations of those compared methods because the codes of most of these methods were provided by the original authors except the DFRV method because its codes have not been publicly available. We carefully implemented their methods by following their settings in [9]. We tuned the parameters of different methods as follows: For MSM and DCC, we performed PCA to learn a linear subspace for each face image set where each subspace dimension was set as 10 to preserve 90% of the energy to compute the similarity of two image sets. For MMD and MDA, the parameters were configured according to [68] and [66], respectively. Specifically, the maximum canonical correlation was used to compute MMD, and the number of connected nearest neighbors for computing geodesic distance in both MMD and MDA was fixed as 12. No parameter is required in AHISD. For CHISD and SANP, we followed the same parameter settings as those in [5] and [25]. For CDL, the KLDA was employed for discriminative learning and the regularization parameter was set the same as that in [67]. For DFRV, we followed the parameter settings in [9]. For the DCC, CDL and LMKML methods, if there is a single video from each class in the Honda, MoBo, YTC, and PubFig datasets, we randomly and equally divided each video clip into two image sets to model the within-class variation.

Table II tabulates the average rank-one recognition rates of different image set based face identification methods on the Honda, MoBo, YTC, PubFig, and IJB-A datasets. We see that our SFDL and D-SFDL perform better than the other twelve compared image set based face identification methods on the Honda, MoBo, YTF, PubFig, and IJB-A datasets, respectively. Compared with the existing unsupervised image set based face recognition methods such as MSM, DCC, MMD, AHISD, CHISD, SANP, and DFRV, our SFDL and D-SFDL can exploit more discriminative information in the learned feature projection matrix and dictionary. Compared with the existing supervised image set based face recognition methods such as MDA, CDL, and LMKML, our SFDL and D-SFDL can project each image frame into a discriminative feature subspace and encode it with a class-specific dictionary, so that more person-specific information can be extracted. Moreover, D-SFDL consistently outperforms SFDL in all experiments as it can better exploit the nonlinear information of different face videos, which is helpful to separate different persons in the nonlinear feature space.

2) Comparison With Existing Deep Learning-Based Face Recognition Methods: We conducted experiments by first extracting features using deep CNNs for each frame and then learning dictionaries for image set-based face recognition. Specifically, we compared our SFDL and D-SFDL with the following four methods:

- CNN+DL (Dictionary Learning): We employed the VGG-Face CNN model provided by [54] to compute the CNN feature descriptor for each frame. Specifically, for each face image in the image set, we first resized it into 224 × 224 pixel image to compute a 4096-dimensional CNN feature vector. Then, each CNN feature vector was reduced to the size of 200 by PCA. Then, we performed dictionary learning on these CNN features for image set based face recognition.
- CNN+SFDL: We first extracted the CNN feature for each frame and then performed SFDL on the extracted feature rather than the original raw pixel for image set based face recognition.
- CNN+D-DL (Deep Dictionary Learning): We first extracted the CNN feature for each frame and then performed deep dictionary learning on these CNN features for image set based face recognition.
- CNN+D-SFDL: We first extracted the CNN feature for each frame and then performed D-SFDL on these CNN features rather than the original raw pixel for image set based face recognition.

Table III shows the image set-based face identification performance of our methods when different face representations on the frame level on different face datasets. We see that the CNN feature representation indeed improves the performance our SFDL framework when the raw pixels are used for recognition, so that CNN+DL achieves better performance than SFDL. This is because SFDL cannot learn deep feature at the frame level for image set representation. However, our method can also been employed on the CNN feature and better performance can be further obtained. Hence, joint learning still achieves better performance than individual learning on all image set based recognition experiments. We find that the improvement of the VGG CNN feature over the conventional shallow models on PaSC is not so significant. There are two possible reasons: 1) the quality of videos in the PaSC dataset

TABLE IV

AVERAGE RECOGNITION RATES (%) OF DIFFERENT FEATURE AND DICTIONARY LEARNING STRATEGIES ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
IFDL	98.3	94.1	74.3	78.5	24.8
SFDL	**100.0**	**96.7**	**76.7**	**80.2**	**26.6**
D-IFDL	99.5	96.7	78.3	80.5	27.2
D-SFDL	**100.0**	**98.5**	**79.5**	**83.5**	**28.2**

TABLE V

AVERAGE RECOGNITION RATES (%) OF THE STRUCTURED AND SHARED DICTIONARY LEARNING METHODS ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
Shared SFDL	98.3	95.3	74.7	78.9	25.2
Structured SFDL	**100.0**	**96.7**	**76.7**	**80.2**	**26.6**
Shared D-SFDL	98.5	97.3	76.7	81.3	27.6
Structured D-SFDL	**100.0**	**98.5**	**79.5**	**83.5**	**28.2**

is much worse that those in other face images and videos such as LFW and YTF, so that the power of the deep features on such low-quality datasets is not so strong; 2) the VGG CNN feature is a pre-trained face feature model and we have not fine-tuned it on the PaSC dataset.

3) Simultaneous vs. Individual Feature and Dictionary Learning: The feature learning and dictionary learning can also be learned in an individual manner. To show the effectiveness of SFDL, we compare our SFDL method with the individual feature and dictionary learning (IFDL) method, and D-SFDL with the deep individual feature and dictionary learning (D-IFDL) method, respectively. IFDL means the feature projection matrix and the structured dictionaries are learned from the training set individually, and D-IFDL means the feature projection matrix is learned first and then the structured deep dictionaries are learned from the training set subsequently. Table IV tabulates the average recognition rates of these four methods. We observe that our simultaneous methods achieve higher recognition rate than the compared individual methods, which shows that jointly learning the feature and dictionary is better because some useful information for dictionary learning may be lost in the feature learning phase in the individual learning methods.

4) Structured vs. Shared SFDL: To demonstrate the advantage of the structured dictionary in our approch, we also compared SFDL with a shared SFDL method which learns a common dictionary in SFDL rather than a structured dictionary, and D-SFDL with a shared D-SFDL method which learns a common dictionary in D-SFDL rather than a structured dictionary, respectively. Table V tabulates the average recognition rates of these four types of SFDL methods. We observe that the structured SFDL and D-SFDL achieve higher recognition rate than the shared SFDL and D-SFDL methods. This is because the structured SFDL and D-SFDL can characterize more class-specific information than the corresponding shared versions.

5) Kernel-Based SFDL vs. D-SFDL: We employed the kernel trick on the DSML method and developed another baseline method (kernel SFDL: K-SFDL) to map the data vectors into a high dimensional space without an explicit nonlinear mapping. In our experiments, we used the RBF kernel $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$. Table VI shows performance comparisons with K-SFDL and D-SFDL on different face datasets. We see that our deep SFDL outperforms the corresponding kernel-based SFDL method on all datasets

TABLE VI

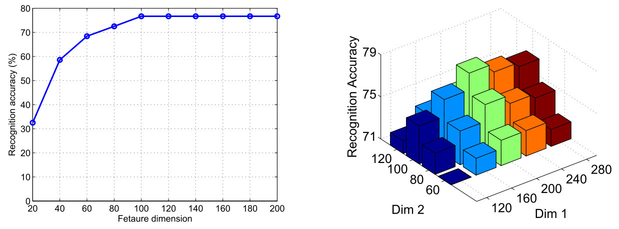
AVERAGE RECOGNITION RATES (%) OF THE KERNEL-BASED AND DEEP SFDL METHODS ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
K-SFDL	98.5	97.5	77.5	81.6	27.6
D-SFDL	**100.0**	**98.5**	**79.5**	**83.5**	**28.2**

TABLE VII

AVERAGE RECOGNITION RATES (%) OF D-SFDL WITH DIFFERENT ACTIVE FUNCTIONS ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
SFDL-1	98.5	95.7	74.3	78.8	24.8
SFDL-2	99.5	96.3	76.0	79.3	25.4
SFDL	**100.0**	**96.7**	**76.7**	**80.2**	**26.6**



(a) (b)

Fig. 3. Average recognition rate (%) of our (a) SFDL and (b) D-SFDL versus different feature dimension of the learned feature spaces on the YTC dataset, where $Dim1$ and $Dim2$ denote the feature dimensions of the second and third layers, respectively.

6) Influence of Different Learning Strategies: We investigated the contributions of different terms in our SFDL model. We defined the following two alternative baselines to study the importance of different terms in our SFDL model:

1) SFDL-1: learning the model without J_2.
2) SFDL-2: learning the model without J_3.

Table VII shows performance comparisons of SFDL when J_2 and J_3 were not used to learn the model, respectively. We see that both J_2 and J_3 in our SFDL exploit discriminative information, and J_2 contributes more than J_3 in terms of the final performance. Moreover, the highest recognition rate can be obtained when both J_2 and J_3 are used together to learn the model.

7) Parameter Analysis: We first evaluated the effect of the feature dimension of the learned feature spaces of our SFDL and D-SFDL on the recognition performance. Fig. 3 shows the recognition accuracy of our SFDL and D-SFDL versus different feature dimensions on the YTC dataset. We see that our proposed SFDL and D-SFDL achieve stable performance when the feature dimension reaches around 100 for SFDL and D-SFDL at the last layer of the network, respectively.

Then, we investigated the performance of our SFDL versus different parameters. Figs. 4-8 shows the recognition accuracy of our SFDL versus different parameters on the YTC dataset. We see that our proposed SFDL achieves stable performance when these parameters were set in suitable ranges.

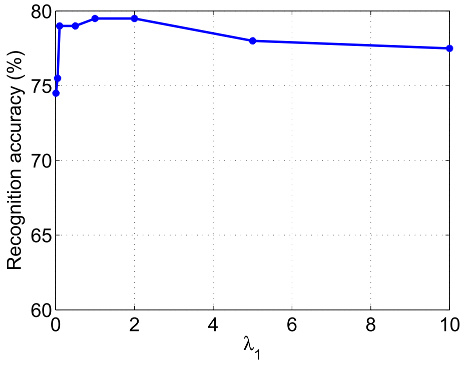


Fig. 4. Average recognition rate (%) of our SFDL versus different values of λ_1 on the YTC dataset.

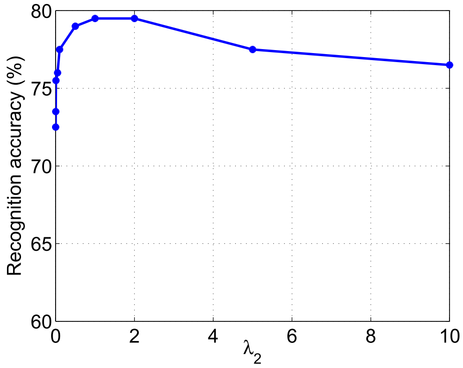


Fig. 5. Average recognition rate (%) of our SFDL versus different values of λ_2 on the YTC dataset.

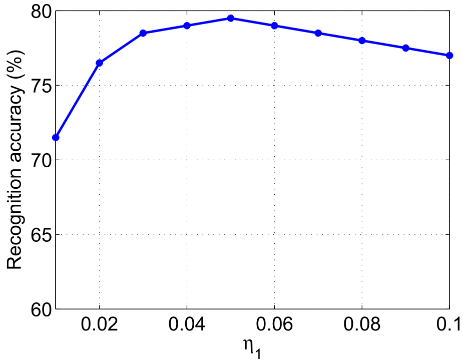


Fig. 6. Average recognition rate (%) of our SFDL versus different values of η_1 on the YTC dataset.

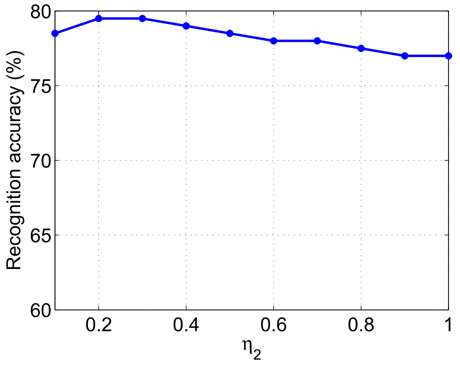


Fig. 7. Average recognition rate (%) of our SFDL versus different values of η_2 on the YTC dataset.

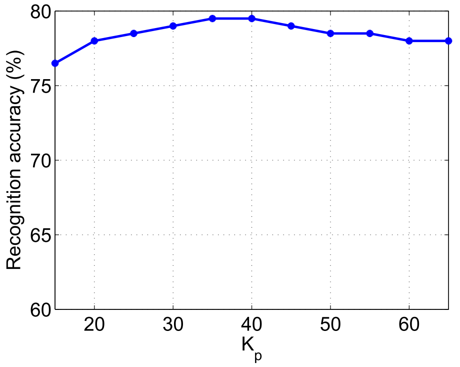


Fig. 8. Average recognition rate (%) of our SFDL versus different values of K_p on the YTC dataset.

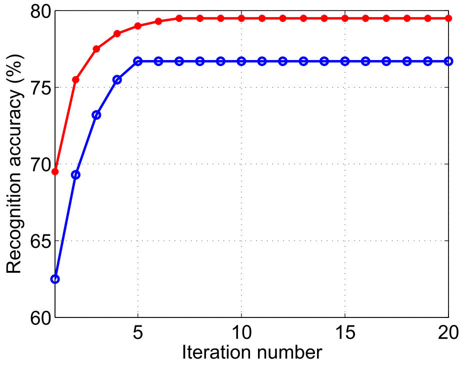


Fig. 9. Average recognition rate (%) of our approach versus different number of iterations on the YTC dataset.

We also investigated the performance of our SFDL and D-SFDL versus different number of iterations. Fig. 9 shows the recognition accuracy of our SFDL over different number of iterations on the YTC dataset. We see that our proposed SFDL can achieve stable performance in several iterations.

We evaluated the effect of the activation function in our D-SFDL method. Table VIII shows the performance of our D-SFDL method where different activation functions are employed on different face datasets. We see that the *tanh* function performs the best and the *sigmoid* function performs the worst. The main reason is that the output of the *tanh* function is symmetrical and within the range $(-1, 1)$, but the output of the standard *sigmoid* function is asymmetrical and within the range $(0, 1)$. The asymmetry in the *sigmoid* function introduces a bias that pushes the activations of the later layers towards saturation, which is not a good position to obtain good training.

Lastly, we investigated the performance of our D-SFDL versus different number of layers. Table IX shows the performance of our D-SFDL method where different number of layers were employed in our D-SFDL on the YTC dataset. We see that D-SFDL achieves the best performance when the number of layer was set to 2. Specifically, the performance first increased and then dropped as the increasing of the number of layers. The reason is that as the increasing of the number of layers, the nonlinearity of samples can be better exploited

TABLE VIII

AVERAGE RECOGNITION RATES (%) OF D-SFDL WITH DIFFERENT ACTIVE FUNCTIONS ON DIFFERENT FACE DATASETS

Method	Honda	MoBo	YTC	PubFig	IJB-A
sigmoid	98.3	96.5	76.7	80.9	27.2
ns-sigmoid	98.3	95.3	74.7	78.9	26.8
tanh	**100.0**	**98.5**	**79.5**	**83.5**	**28.2**

TABLE IX

AVERAGE RECOGNITION RATES (%) OF D-SFDL WITH DIFFERENT NUMBER OF LAYERS ON DIFFERENT FACE DATASETS

Layer number	Honda	MoBo	YTC	PubFig	IJB-A
1	**100.0**	96.7	76.7	78.5	26.6
2	**100.0**	**98.5**	**79.5**	**83.5**	**28.2**
3	98.3	96.7	78.4	82.4	27.9
4	98.3	96.3	77.3	81.7	27.5
5	97.5	96.0	77.0	81.3	27.1

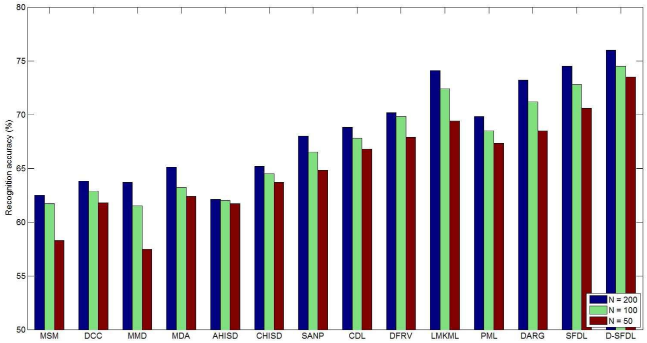


Fig. 10. The average recognition rate (%) of different image set based face identification methods on the YTC dataset where different number of image frames are used for evaluation.

in the network. However, since we don't have large number of training samples, the network could be over-fit when the number of layers increased.

8) Robustness Analysis: We first tested the robustness of our proposed approach versus different amount of noisy data in face videos. We followed the settings in [5], [45], [67] and conducted three experiments where the training and/or testing face image sets were corrupted by adding one image from each of the other classes. The original data and three noisy scenarios are called as "original", "NTR" (only training videos have noisy data), "NTE" (only testing videos have noisy data), and "NTT" (both training and testing videos have noisy data), respectively. Table X records the recognition accuracy of different image set based face recognition methods with different amounts of noisy data on the YTC dataset.

We also evaluated the performance of our approach when face videos contain varying number of image frames. We randomly selected N frames from each face image set (both training and testing) and used them for recognition. If there are less than N image frames for one face image set, all image frames within this image set were used for recognition. Fig. 10 shows the performance of different methods on the YTC dataset with varying image frames.

From Table X and Fig. 10, we observe that our approach demonstrates strong robustness with some slight performance drop than the other compared methods. This is because we use dictionaries to represent each face image set and such dictionary-based methods are robust to noise and the number of samples in face image set. Hence, the effects of the noisy samples and varying data size can be alleviated in our proposed approach.

TABLE X

AVERAGE RECOGNITION RATES (%) OF DIFFERENT IMAGE SET BASED FACE RECOGNITION METHODS WITH DIFFERENT AMOUNTS OF NOISY DATA ON THE YTC DATASET

Method	Original	NTR	NTE	NTT
MSM [72]	62.8	59.7	45.3	52.2
DCC [33]	64.8	58.7	49.9	54.2
MMD [68]	66.7	62.5	46.4	55.4
MDA [66]	68.1	65.8	52.5	53.4
AHISD [5]	66.5	62.5	44.5	35.6
CHISD [5]	67.4	66.8	42.5	38.5
SANP [25]	68.3	67.2	47.5	39.4
CDL [67]	69.7	68.4	54.5	58.4
DFRV [9]	74.5	71.1	60.8	62.1
LMKML [45]	78.2	76.1	64.5	66.1
PML [27]	70.3	69.5	49.6	52.6
DARG [69]	77.1	72.5	61.4	59.8
SFDL	**76.7**	**76.3**	**64.8**	**67.2**
D-SFDL	**79.7**	**78.5**	**67.9**	**69.5**

TABLE XI

COMPARISON OF THE COMPUTATION TIME (SECONDS) OF DIFFERENT METHODS ON THE YTC DATASET

Method	Training	Testing
MSM	N.A.	2.7
DCC	98.6	2.5
MMD	N.A.	3.5
MDA	185.3	3.2
AHISD	N.A.	8.7
CHISD	N.A.	6.7
SANP	N.A.	48.6
CDL	68.5	12.8
DFRV	8656.5	5.4
LMKML	4232.8	210.6
PML	124.6	5.4
DARG	112.3	2.8
SFDL	7532.5	6.5
D-SFDL	9632.8	7.8

9) Computational Time: Lastly, we reported the computational time of different methods using the YTC dataset. Our hardware configuration is a 2.8-GHz CPU and a 24GB RAM. Table XI shows the computational time for different methods under the Matlab paltform. It is to be noted that training time is only required for some discriminative learning and dictionary learning methods such as DCC, MDA, CDL, DFRV, LMKML and our methods. We see that the computational time of our SFDL and D-SFDL are generally larger than many other compared methods for training and are comparable to them for testing.

D. Discussion

The above experimental results suggest the following four key observations:

1) Our SFDL and D-SFDL outperform most state-of-the-art image set based face identification methods. This is because our SFDL and D-SFDL project each image frame into a discriminative feature subspace and encode it with a class-specific dictionary, so that more person-specific information can be extracted.
2) D-SFDL consistently outperforms SFDL in all experiments, which shows that hierarchical non-linear

transformations can better exploit the nonlinearity of face samples to improve the recognition performance.

3) Our SFDL and D-SFDL show stronger robustness to noise than existing image set based face recognition methods. This is because variations of the pose, illumination and expression information within face image sets can be implicitly encoded into the learned dictionaries, which are robust to noise variations.
4) Both the simultaneous learning and the structured learning strategies contribute the performance improvement of our SFDL method, which further shows the effectiveness of exploiting discriminative features and dictionaries in a one-step framework and learning class-specific dictionaries in our approach.

V. Conclusion

In this paper, we have proposed a new simultaneous feature and dictionary learning (SFDL) method for image set based face recognition. By jointly learning the feature projection matrix and the structured dictionary, our approach extracts more discriminative information for image set based face representation. To better exploit the nonlinearity of face samples from different image sets, we have developed a deep SFDL (D-SFDL) method by jointly learning hierarchical nonlinear transformations and class-specific dictionaries to further improve the recognition performance. Experimental results on five widely used face datasets have shown the superiority of our approach over the state-of-the-art image set based face recognition methods in terms of accuracy and robustness.

References

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proc. CVPR*, 2005, pp. 581–588.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. NIPS*, 2007, pp. 153–160.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. CVPR*, 2010, pp. 2707–2714.

[5] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. CVPR*, 2010, pp. 2567–2573.

[6] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. CVPR*, 2013, pp. 3025–3032.

[7] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[8] L. Chen, "Dual linear regression based classification for face cluster recognition," in *Proc. CVPR*, 2014, pp. 2673–2680.

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

[10] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face and person recognition from unconstrained video," *IEEE Access*, vol. 3, no. 3, pp. 1783–1798, 2015.

[11] T. Chin, K. Schindler, and D. Suter, "Incremental kernel SVD for face recognition with image sets," in *Proc. FG*, 2006, pp. 461–466.

[12] M. Du, A. C. Sankaranarayanan, and R. Chellappa, "Robust face recognition from multi-view videos," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1105–1117, Mar. 2014.

[13] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1610–1623, Oct. 2010.

[14] W. Fan and D. Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in *Proc. CVPR*, 2006, pp. 1384–1390, 2006.

[15] A. Fitzgibbon and A. Zisserman, "Joint manifold distance: A new approach to appearance based clustering," in *Proc. CVPR*, 2003, pp. 26–33.

[16] R. Gross and J. Shi, "The CMU motion of body (MOBO) database," Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 2001.

[17] H. Guo, Z. Jiang, and L. S. Davis, "Discriminative dictionary learning with pairwise constraints," in *Proc. ACCV*, 2012, pp. 328–342.

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

[19] M. Hayat, M. Bennamoun, and S. An, "Reverse training: An efficient approach for image set classification," in *Proc. ECCV*, 2014, pp. 784–799.

[20] 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.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[23] H. Hu, "Face recognition with image sets using locally Grassmannian discriminant analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1461–1474, Sep. 2014.

[24] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. CVPR*, 2014, pp. 1875–1882.

[25] Y. Hu, A. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1992–2004, Oct. 2012.

[26] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. CVPR*, 2012, pp. 2518–2525.

[27] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on Grassmann manifold with application to video based face recognition," in *Proc. CVPR*, 2015, pp. 140–149.

[28] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, "Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification," in *Proc. ICML*, 2015, pp. 720–729.

[29] S. U. Hussain *et al.*, "Face recognition using local quantized patterns," in *Proc. BMVC*, 2012, pp. 1–12.

[30] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," in *Proc. Natural Image Statist.*, 2009, pp. 151–175.

[31] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. CVPR*, 2011, pp. 1697–1704.

[32] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. CVPR*, 2008, pp. 1–8.

[33] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[34] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A," in *Proc. CVPR*, 2015, pp. 1931–1939.

[35] S. Kong and D. Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *Proc. ECCV*, 2012, pp. 186–199.

[36] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. ICCV*, 2009, pp. 365–372.

[37] Q. V. Le, A. Karpenko, J. Ngiam, and A. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. NIPS*, 2011, pp. 1017–1025.

[38] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. CVPR*, 2011, pp. 3361–3368.

[39] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proc. NIPS*, 2006, pp. 801–808.

[40] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. CVPR*, 2003, pp. 313–320.

[41] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 289–302, Feb. 2014.

[42] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, 2014, pp. 152–159.

[43] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

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

[45] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. ICCV*, 2013, pp. 329–336.

[46] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. CVPR*, 2012, pp. 2586–2593.

[47] A. Mahmood, A. Mian, and R. Owens, "Semi-supervised spectral clustering for image set classification," in *Proc. CVPR*, 2014, pp. 121–128.

[48] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. CVPR*, 2008, pp. 1–8.

[49] I. Masi, A. T. Trân, J. T. Leksut, T. Hassner, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proc. ECCV*, vol. 5. 2016, pp. 579–596.

[50] A. Mian, Y. Hu, R. Hartley, and R. Owens, "Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5252–5262, Dec. 2013.

[51] H. Nguyen, V. Patel, N. Nasrabadi, and R. Chellappa, "Sparse embedding: A framework for sparsity promoting dimensionality reduction," in *Proc. ECCV*, 2012, pp. 414–427.

[52] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. ICML*, 2014, pp. 82–90.

[53] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. CVPR*, 2012, pp. 3258–3265.

[54] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–11.

[55] S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. ICML*, 2011, pp. 833–840.

[56] W. J. Scheirer *et al.*, "Report on the BTAS 2016 video person recognition evaluation," in *Proc. BTAS*, 2016, pp. 1–8.

[57] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[58] G. Shakhnarovich, J. Fisher, and T. Darrell, "Face recognition from long-term observations," in *Proc. ECCV*, 2006, pp. 361–375.

[59] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.

[60] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10000 classes," in *Proc. CVPR*, 2014, pp. 1891–1898.

[61] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. CVPR*, 2015, pp. 2892–2900.

[62] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. ICCV*, 2013, pp. 1489–1496.

[63] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, 2014, pp. 1–8.

[64] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[65] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[66] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proc. CVPR*, 2009, pp. 1–8.

[67] R. Wang, H. Guo, L. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. CVPR*, 2012, pp. 2496–2503.

[68] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proc. CVPR*, 2008, pp. 1–8.

[69] 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. CVPR*, 2015, pp. 2048–2057.

[70] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," in *Proc. ICML*, 2013, pp. 846–854.

[71] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, 2011, pp. 529–534.

[72] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proc. FG*, 1998, pp. 318–323.

[73] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. ICCV*, 2011, pp. 543–550.

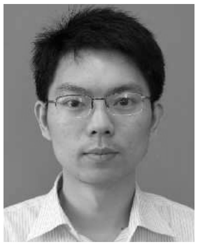
[74] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. ICIP*, 2010, pp. 1601–1604.

[75] L. Zhang, M. Yang, Z. Feng, and D. Zhang, "On the dimensionality reduction for sparse representation based face recognition," in *Proc. ICPR*, 2010, pp. 1237–1240.

[76] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. CVPR*, 2010, pp. 2691–2698.

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

Jiwen Lu (S'10–M'11–SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2003, 2006, and 2012, respectively. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. From 2011 to 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He has authored or co-authored over 150 scientific papers in these areas, where 45 were the IEEE Transactions papers. His current research interests include computer vision, pattern recognition, and machine learning. He serves/has served as an elected member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He was a recipient of the National 1000 Young Talents Plan Program in 2015. He is/was a Workshop Chair/Special Session Chair/Area Chair for over ten international conferences. He serves/has served as an Associate Editor of *Pattern Recognition Letters*, *Neurocomputing*, and the IEEE ACCESS, a Managing Guest Editor of *Pattern Recognition* and *Image and Vision Computing*, and a Guest Editor of *Computer Vision and Image Understanding*.



Gang Wang received the B.Eng. degree in electrical engineering from the Harbin Institute of Technology and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. He was an Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. He had a joint appointment with the Advanced Digital Science Center, Singapore, as a Research Scientist from 2010 to 2014. He is currently a Researcher of Alibaba and a Chief Scientist of Alibaba AI Labs. He was a recipient of MIT technology review innovator under 35 awards (Asia). He is an Associate Editor of TPAMI and an Area Chair of ICCV 2017.

Jie Zhou (M'01–SM'04) received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as PAMI, TIP, and CVPR. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the *International Journal of Robotics and Automation* and two other journals.