# Learning Stacked Image Descriptor for Face Recognition

Zhen Lei, *Member, IEEE,* Dong Yi and Stan Z. Li, *Fellow, IEEE*

***Abstract*—Learning based face descriptors have constantly improved the face recognition performance. Compared to the hand-crafted features, learning based features are considered to be able to exploit information with better discriminative ability for specific tasks. Motivated by the recent success of deep learning, in this paper, we extend the original 'shallow' face descriptors to 'deep' discriminant face features by introducing a stacked image descriptor (SID). With deep structure, more complex facial information can be extracted and the discriminant and compactness of feature representation can be improved. The SID is learned in a forward optimization way, which is computational efficient compared to deep learning. Extensive experiments on various face databases are conducted to show that SID is able to achieve high face recognition performance with compact face representation, compared with other state-of-the-art descriptors.**

***Index Terms*—Face recognition, stacked image descriptor, deep discriminant face representation, learning based descriptor**

## I. Introduction

Face recognition has attracted much attention due to its potential value for applications and its theoretical challenges. In real world, the face images are usually affected by different expressions, poses, occlusions and illuminations, so that the difference of face images from the same person could be larger than that from different ones. Therefore, how to extract robust and discriminant features which make the intra-person faces compact and enlarge the margin among different persons is a critical and difficult problem in face recognition.

In the early period of face recognition, researchers try to extract global feature transformation from the whole face image. Subspace learning like principal component analysis (PCA) [42], linear discriminant analysis (LDA) [2] etc. are representative methods in this category. Because of the holistic property, global feature transformation is usually not robust to local appearance variations caused by expression, occlusion, pose etc. The local face descriptors like Gabor [29], [21] and local binary pattern (LBP) [1], then emerge. However, all these descriptors are designed in a hand-crafted way, which may not exploit the discriminative information from face images sufficiently.

Recently, researchers propose to learn face descriptors from face images in a data-driven way. Yang et al. [55] learn the optimal dictionary in sparse representation using Fisher discrimination. [4], [13], [30], [31] propose to learn pattern encoders of LBP like features. Cao et al. [4] adopt an unsupervised method (random-projection and PCA tree) to make the distribution of encoded LBP value uniform. Guo et al. [13] propose a supervised learning method based on Fisher separation criterion to learn a discriminative encoder of LBP like feature. In [30], authors propose to construct a decision tree for each region to encode the pixel comparison result and in [31], a heuristic algorithm is used to find the optimal pixel comparison pairs for discriminative face representation. Lei et al. [22] propose a discriminant face descriptor by learning discriminant image filters and optimal sampling strategy to better discriminate face images. Zhang et al. [62] propose multiple random faces guided sparse single-hidden-layer neural networks to extract pose-invariant features and meanwhile keep identity information.

Sparse representation classifier (SRC) is another important branch for face representation and recognition. Since the first successful application of SRC to face recognition [50], a lot of variants of sparse representation have been proposed [49]. Gao et al. [11] extends the sparse representation to kernel space and propose kernel sparse representation to capture the nonlinear similarity of features. Yang el al. [57] propose a robust sparse coding method by seeking the maximum likelihood estimation (MLE) solution of the sparse coding problem, which is more robust to occlusions and corruptions. Yang et al. [56] further apply SRC with a Gabor based occlusion dictionary to improve the robustness to occlusion. Wagner et al. [47] apply sequential $l^1$-minimization to achieve pixel-accurate face alignment result and adopt SRC for face recognition. Zhang et al. [59] propose joint dynamic sparse representation to address the multi-view face recognition problem by promoting shared joint sparsity patterns among the multiple sparse representation vectors. Mousavi et al. [32] introduce the spike-and-slab priors into sparse representation and treat the multi-view face recognition as a multi-task image classification problem. Xu et al. [53] propose a two-phase sparse representation method to classify the test samples to its class accurately. A novel sparse manifold subspace learning method by incorporating locality sparse coding and sparse eigen-decomposition is recently proposed [35]. It avoids the parameter tuning of neighbors selection and improves the robustness of the solution to data noise. More face recognition methods using sparse representation can be seen in [16], [15], [9], [27].

Most of the above face descriptors are learned in a 'shallow' way. Very recently, with the development of computational resource and data collection, deep structure feature learning



Z. Lei, D. Yi, and S. Li are with Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Donglu, Beijing 100190, China. Email: {zlei, dyi, szli}@nlpr.ia.ac.cn.

has achieved great success in object representation field. With the help of deep structure, the capacity of model representation is greatly improved and it is possible to learn complex (non-linear) information from data effectively. Simonyan et al. [37] show that deep Fisher networks by simply stacking Fisher vectors improve the performance of standard Fisher vector representation significantly. Chen et al. [8] propose marginalized stacked denoising autoencoders (SDAs) to learn the optimal solution in closed-form, thus the high computational cost is reduced and the scalability to high-dimensional features is improved. In the following, we mainly review the related work in face recognition field. Sun et al. [38] propose a hybrid convolutional network with restricted bolzmann machine for face verification. The local relational visual features from face pairs are learned and extracted by the deep network. Zhu et al. [63] design a neural network consisting of face identity-preserving feature layer and reconstruction layer to learn features that reduce the intra variance and preserve the discriminative information between identities. Taigman et al. [41] from Facebook propose a DeepFace model by learning a locally connected nine-layer deep neural network without weighting sharing from four million facial images. Sun et al. [40] learn multiple convolutional networks (ConvNets) by distinguishing more than 10,000 subjects. The learned ConvNets generalizes well to face verification task. They improve the ConvNet by incorporating face identification and verification tasks and the face recognition performance is further enhanced [39]. Hu et al. [17] propose deep metric learning using fully connected neural network, which outperforms the traditional metric learning. Cai et al. [3] stack several independent subspace analysis (ISA) with sparse constraint to build up deep network to extract identity related representation.

Motivated by the success of deep learning, in this work, we propose to stack the traditional shallow descriptor to deep structure and propose a framework of stacked image descriptor (SID) learning. The SID consists of image descriptor layer and pooling layer alternatively. A sub-optimal forward learning method is adopted to learn SID in a layer-wise way. For each image descriptor layer, we try to minimize the specific loss function to learn the best descriptor based on the output of the last layer. We experimentally show that SID, which stacks image descriptors, does improve the face discriminative ability and achieve better face recognition performance compared to the traditional shallow learning based descriptors.

The contribution of this paper mainly includes: 1) A stacked image descriptor is proposed, which can be learned in a layer-wise way with a close-form solution. The computational complexity is lower than deep learning. It can be optimized without using specific acceleration hardware like GPU. 2) The structure of proposed SID is compatible with convolutional neural network (CNN) and the solution of SID can be considered as a pre-training result for CNN. For descriptor learning at each layer, the responses of different channels from previous layers are concatenated, which provide sufficient and complementary information to learn discriminant image descriptors. 3) Four SID implementations (PCA-SID, LDA-SID, Tensor-SID and DFD-SID) are presented. PCA-SID and LDA-SID involves 2nd order information, while Tensor-SID utilizes higher order information to obtain the optimal solution. DFD-SID takes into account not only the discriminative convolutional filters, but also the optimal sampling strategy to extract effective face representation.

The rest of the paper is organized as follows. We detail the SID learning in Section II. Four image descriptor learning methods are described. We discuss the time complexity of SID learning and its relation to previous methods in Section III and IV. In Section V, we examine the performance of SID on both constrained and unconstrained face databases, compared with state-of-the-art methods. In Section VI, we conclude the paper.

## II. Stacked Image Descriptor Learning Framework

Learning based image descriptors have demonstrated its superiority over hand-crafted ones in previous work [55], [4], [13], [30], [31], [22]. On the other hand, the advantage of deep feature representation has been shown in many works [37], [38], [41], [40], [39] recently. In this work, we incorporate the image descriptor learning (IDL) and deep structure to propose a stacked image descriptor learning method. Fig. 1 illustrates the structure of SID used in this paper. The image descriptor and max-pooling layers are concatenated alternatively. At each image descriptor layer, the optimal descriptor is learned based on the output of the last layer. Finally, the responses from the last image descriptor layer and the max-pooling layer are concatenated and the linear discriminant analysis (LDA) is applied to derive the compact discriminant feature representation, i.e., the representation of stacked image descriptor.

There have been several image descriptor learning methods for face images. Chan et al. [5] apply PCA to learn optimal image filters. Lei et al. [23] apply LDA to learn discriminant image filters. Furthermore, they propose discriminant face descriptor (DFD) to learn the discriminant and optimal sampling strategy respectively. In the following, we describe the image descriptor learning methods using PCA, LDA, and DFD at a single layer, which are abbreviated as PCA-IDL, LDA-IDL and DFD-IDL, respectively. Besides, we also introduce discriminant tensor analysis (DTA) [54] based image descriptor learning method, denoted as DTA-IDL.

### A. PCA-IDL

In this part, we try to learn optimal image filters that extract useful information for face recognition. The convolution of the filter and the local image region consists of element-wise multiplication and a summation operator. The pixels in a local image region and the filter can be represented in vectors form, denoted as $x$ and $w$ respectively. The convolution result, $f$, can then be computed as $f = w^T x$, where $f$ is the convolution result. The image descriptor learning aims to find optimal filter $w$, so that after convolutional filtering, specific properties of the responses are optimized. With PCA formulation, the covariance of the responses is expected to be maximized so that most energy of signals can be preserved and the irrelative noise is removed. Denoting the pixel vector of patch at position $p$ from the $i$-th sample as $x_i^p$ and filter vector as $w$, the

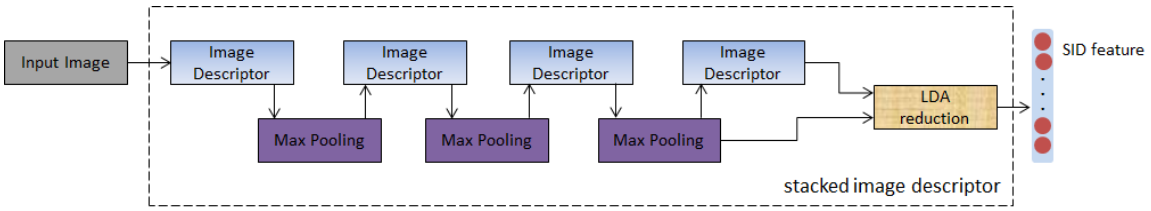



Fig. 1. Structure of stacked image descriptor for feature extraction.

convolution operation can be represented as $f_i^p = w^T x_i^p$, where $f_i^p$ is the convolution response. The objective of PCA based IDL is to find a filter $w$ to maximize the covariance of response $f^p$, which can be formulated as

$$\begin{aligned} w &= \arg\max_w \sum_{i=1}^{N} (f_i^p - \bar{f^p})(f_i^p - \bar{f^p})^T \\ &= \arg\max_w \sum_{i=1}^{N} w^T (x_i^p - \bar{x^p})(x_i^p - \bar{x^p})^T w, \end{aligned} \tag{1}$$

where $\bar{f^p}$ and $\bar{x^p}$ are mean vectors over all the samples at position $p$; $N$ is the number of total samples. If we share the filter over the whole image, that is, the filter at different position is the same, the objective of PCA-CFL can be re-formulated as

$$w = \arg\max_w \sum_{p=1}^{P} \sum_{i=1}^{N} w^T (x_i^p - \bar{x^p})(x_i^p - \bar{x^p})^T w, \tag{2}$$

where $P$ is the number of positions. Denoting $C = \sum_{p=1}^{P} \sum_{i=1}^{N} (x_i^p - \bar{x^p})(x_i^p - \bar{x^p})^T$, the optimal filter vector $w$ can be obtained by solving the eigen-value problem of $C$ corresponding to the leading eigenvalue.

### B. LDA-IDL

PCA based filter learning does not consider class information, thus the discriminant information is not sufficiently exploited. We can intuitively use LDA instead of PCA to exploit more discriminant information. With LDA formulation, after image filtering, the response differences of filtered images from the same class are expected to be minimized and the response differences from different classes are maximized. Let $x_{ij}^p$ be an image patch vector at position $p$ of $j$-th image from the $i$-th class, and the learned filter vector be $w$, the objective of LDA based IDL can be formulated as

$$w = \arg\max_w \frac{|S_b(w)|}{|S_w(w)|}, \tag{3}$$

where $S_b(w)$ and $S_w(w)$ are between-class and within-class scatter matrices, defined as

$$\begin{aligned} S_w &= \sum_{p=1}^{P} \sum_{i=1}^{L} \sum_{j=1}^{C_i} w^T (x_{ij}^p - m_i^p)(x_{ij}^p - m_i^p)^T w \\ &= w^T (\sum_{p=1}^{P} \sum_{i=1}^{L} \sum_{j=1}^{C_i} (x_{ij}^p - m_i^p)(x_{ij}^p - m_i^p)^T) w, \\ S_b &= \sum_{p=1}^{P} \sum_{i=1}^{L} C_i w^T (m_i^p - m^p)(m_i^p - m^p)^T w \\ &= w^T (\sum_{p=1}^{P} \sum_{i=1}^{L} C_i (m_i^p - m^p)(m_i^p - m^p)^T) w, \end{aligned} \tag{4}$$

where $m_i^p$ and $m^p$ are the mean vectors of patch vectors at position $p$ over the samples in the $i$-th class and the whole sample set, respectively. $P$ is the number of position, $L$ is the number of class and $C_i$ is the number of samples belonging to the $i$-th class. Denoting $\hat{S}_b = \sum_{p=1}^{P} \sum_{i=1}^{L} C_i (m_i^p - m^p)(m_i^p - m^p)^T$ and $\hat{S_w} = \sum_{p=1}^{P} \sum_{i=1}^{L} \sum_{j=1}^{C_i} (x_{ij}^p - m_i^p)(x_{ij}^p - m_i^p)^T$, the optimal filter $w$ can be obtained by solving the generalized eigen-value problem $\hat{S}_b w = \lambda \hat{S_w} w$ corresponding to its leading eigenvalues.

### C. DTA-IDL

In LDA-IDL and PCA-IDL, pixels in the local region are transformed into vector representation, where some structure information may be lost. We introduce Discriminant Tensor Analysis (DTA) [54] into convolutional filter learning to utilize the high-order structure information. For the definitions and operators with tensor representation, please refer to [44]. Considering a local convolution region as a $K$-order tensor $\mathcal{A} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$, the purpose of DTA based convolutional filter learning is to find $K$ projections $w_1, w_2, \cdots, w_K$ so that after mode-$k$ products of tensor $\mathcal{A}$ and $w_1, w_2, \cdots, w_K$, the difference of the between class scatters and the within class scatters is maximized. Mathematically, the objective of DTA-IDL can be formulated as

$$\begin{aligned} &(w_k|_{k=1}^K) = \arg\max_{w_k|_{k=1}^K} \\ &\frac{\sum_{p=1}^{P} \sum_{i=1}^{L} \sum_{j=1}^{C_i} ||(\bar{\mathcal{A}}_i^p - \bar{\mathcal{A}}^p) \times_1 w_1 \cdots \times_K w_K||^2}{\sum_{p=1}^{P} \sum_{i=1}^{L} \sum_{j=1}^{C_i} ||(\mathcal{A}_{ij}^p - \bar{\mathcal{A}}_i^p) \times_1 w_1 \cdots \times_K w_K||^2}, \end{aligned} \tag{5}$$

where $\times_k$ indicates the mode-$k$ product of tensor and matrix. $\bar{\mathcal{A}}_i^p$ and $\bar{\mathcal{A}}^p$ represent the mean tensor at position $p$

over the samples in the $i$-th class and the whole sample set, respectively. As indicated in [54], one can adopt an iterative procedure to solve this problem. In the iteration of solving $w_k$, we fix the other $K-1$ projections $w_1, \ldots, w_{k-1}, w_{k+1}, w_K$ and apply mode-$k$ products of the tensor $\mathcal{A}_{ij}^p$ with $w_1, \ldots, w_{k-1}, w_{k+1}, w_K$ to get the resulted tensor $\mathcal{B}_{ij}^p$. By mode-$k$ flattening the tensor $\mathcal{B}_{ij}^p$ to $X_{ij}^p$, the within and between class scatters can then be computed as

$$\begin{aligned} S_w &= \sum_{p=1}^{P}\sum_{i=1}^{L}\sum_{j=1}^{C_i}(X_{ij}^p - M_i^p)(X_{ij}^p - M_i^p)^T \\ S_b &= \sum_{p=1}^{P}\sum_{i=1}^{L} C_i(M_i^p - M^p)(M_i^p - M^p)^T, \end{aligned} \tag{6}$$

where $M_i^p$ and $M^p$ are the mean matrices of unfolded matrix $X_{ij}^p$ at position $p$ over the samples in the $i$-th class and the whole sample set, respectively. The solution $w_k$ that maximizes the ratio of the between and within class scatters can then be obtained by solving the generalized eigenvalue problem $S_b w = \lambda S_w w$. Algorithm 1 illustrates the optimization of DTA-IDL.

**Algorithm 1** DTA based Image Descriptor Learning

**Input:** A set of samples $\mathcal{A}_{ij}^p, i = 1, \cdots, L, j = 1, \cdots, C_i, p = 1, \cdots, P\}$, where $\mathcal{A}_{ij}^p \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$
**Output:** Image filter projections $w_1, w_2, \cdots, w_K$

**Initialize:** $w_1 = I_{m_1}, w_2 = I_{m_2}, \cdots, w_K = I_{m_K}$.
**for** $t = 1, \ldots, T$ **do**:
  for $k = 1, 2, \cdots, K$ do:
    1) $\mathcal{B}_{ij}^p = \mathcal{A}_{ij}^p \times_1 w_1 \cdots \times_{k-1} w_{k-1} \times_{k+1} w_{k+1} \cdots \times_K w_K$
    2) mode-$k$ flatten the tensor $\mathcal{B}_{ij}^p$ to $X_{ij}^p$.
    3) Compute $S_b, S_w$ using Eq.(6).
    4) Solve the generalized eigenvalue problem
      $S_b w = \lambda S_w w$
      and obtain the eigenvectors $w$ with largest eigenvalues.
    5) $w_k \leftarrow w$.
  end for
**end for**
**Return:** $w_1, w_2, \cdots, w_K$

### D. DFD-IDL

Recently, Lei et al. [22] propose discriminant face descriptor (DFD) by learning image filters and soft sampling matrix from face images. This learning based face descriptor can also be integrated into our framework. In DFD learning, we need first to construct the pixel difference matrix (PDM) $dI$ by comparing the patch vectors in the neighborhood with the central patch vector. At each layer of SID, the responses in local region with multiple channels are transformed into a patch vector $x$. Suppose the central and neighboring patch vectors be $x_c$ and $x_{n_1}, x_{n_2}, \cdots, x_{n_N}$, where $N$ is the number of neighboring patches, the PDM is constructed as $dI = [x_{n_1} - x_c, x_{n_2} - x_c, \cdots, x_{n_N} - x_c]$. The details of PDM construction can be found in [22]. The purpose of DFD is to learn discriminant image filters $w$ and optimal soft sampling matrix $v$ so that the differences of samples from different classes and the same class are maximized. The objective of DFD-IDL is formulated as

$$(w, v) = \arg\max_{w,v} \frac{\sum_{p=1}^{P}\sum_{i=1}^{L} C_i w^T (dm_i^p - dm^p) v v^T (dm_i^p - dm^p)^T w}{\sum_{p=1}^{P}\sum_{i=1}^{L}\sum_{j=1}^{C_i} w^T (dI_{ij}^p - dm_i^p) v v^T (dI_{ij}^p - dm_i^p)^T w}, \tag{7}$$

where $dm_i^p$ and $dm^p$ are mean PDMs over the samples from the $i$-th class and the whole sample set. As shown in [22]. The solution to Eq.(7) can be obtained in an iterative way. First, we fix $v$ and obtain optimal $w$ by solving the generalized eigenvalue problem $\hat{S}_b w = \lambda \hat{S}_w w$ with leading eigenvalues, where

$$\begin{aligned} \hat{S}_w &= \sum_{p=1}^{P}\sum_{i=1}^{L}\sum_{j=1}^{C_i}(dI_{ij}^p - dm_i^p) v v^T (dI_{ij}^p - dm_i^p)^T \\ \hat{S}_b &= \sum_{p=1}^{P}\sum_{i=1}^{L} C_i (dm_i^p - dm^p) v v^T (dm_i^p - dm^p)^T. \end{aligned} \tag{8}$$

Second, we fix $w$ and obtain $v$ by solving the generalized eigenvalue problem $\tilde{S}_b w = \lambda \tilde{S}_w w$ with leading eigenvalues, where

$$\begin{aligned} \tilde{S}_w &= \sum_{p=1}^{P}\sum_{i=1}^{L}\sum_{j=1}^{C_i}(dI_{ij}^p - dm_i^p)^T w w^T (dI_{ij}^p - dm_i^p) \\ \tilde{S}_b &= \sum_{p=1}^{P}\sum_{i=1}^{L} C_i (dm_i^p - dm^p)^T w w^T (dm_i^p - dm^p). \end{aligned} \tag{9}$$

The loop is continued until the stop condition is achieved. The algorithm of DFD-IDL is illustrated in 2. As indicated in [22], in practice, one loop optimization is enough to achieve good result.

**Algorithm 2** DFD based Image Descriptor Learning

**Input:** A set of PDMs $dI_{ij}^p, i = 1, \cdots, L, j = 1, \cdots, C_i, p = 1, \cdots, P\}$, where $dI_{ij}^p \in \mathbb{R}^{d \times N}$. $d$ is the dimension of patch vector and $N$ is the number of neighboring patches.
**Output:** Image filter $w$ and soft sampling matrix $v$

**Initialize:** $w = I, v = I$.
**for** $t = 1, \ldots, T$ **do**:
  1) **Fixing** $v$, compute $\hat{S}_b, \hat{S}_w$ using Eq.(8).
  2) Solve the generalized eigenvalue problem
    $\hat{S}_b w = \lambda \hat{S}_w w$
    and obtain the eigenvectors $w$ with largest eigenvalues.
  3) **Fixing** $w$, compute $\tilde{S}_b, \tilde{S}_w$ using Eq.(9).
  4) Solve the generalized eigenvalue problem
    $\tilde{S}_b v = \lambda \tilde{S}_w v$
    and obtain the eigenvectors $v$ with largest eigenvalues.
**end for**
**Return:** $w$ and $v$

### E. Stacked Image Descriptor Learning

We stack multiple image descriptor layers mentioned above and pooling layers alternatively to form stacked image descriptor (SID). In each image descriptor layer, the image descriptor is learned based on the output responses from the last layer. Finally, we concatenate the responses from the last two layers and apply LDA to derive a compact representation.

Corresponding to PCA-IDL, LDA-IDL, DTA-IDL and DFD-LDA at each single layer, their SIDs are denoted as PCA-SID, LDA-SID, DTA-SID and DFD-SID, respectively. Furthermore, in order to describe the face characteristic more finely, in each image descriptor layer, we partition the face image into $m \times m$ parts and learn the descriptors for each part independently. In this way, we apply different image descriptors to different face regions, so that the useful information can be sufficiently exploited.

In our implementation, The facial partition parameter $m$ is set to 4, 4, 2 and 1 for four image descriptor layers. In each pooling layer, max-pooling operator is applied on a $2 \times 2$ patch. The size of image filters at each image descriptor layer is set to $8 \times 8$, $6 \times 6$, $3 \times 3$ and $3 \times 3$, respectively. For DFD-SID, at each image descriptor layer, we use 8 neighbors with radius 1 to learn the optimal soft sampling matrix. The number of channels of learned filters at each image descriptor layer is preserved to $20, 40, 50, 60$. Finally, we obtain a 160-dimensional face representation for each SID.

## III. Computational Complexity Analysis

In this part, we take LDA-SID and ConvNet [40] as two representative methods of SID and deep convolutional neutral network (DCNN) to analyze the computational complexity. For both methods, the most computational expensive operator is the computation of layer-wise convolution for each image. As described in [14], the complexity of convolution operator is about $O(\sum_{l=1}^{d} n_{l-1} \cdot s_l^2 n_l m_l^2)$, where $l$ is the index of convolutional layer, $d$ is the number of convolutional layers, $n_l$ is the number of learned filters at the $l$-th layer, and $n_{l-1}$ is the number of input channels of the $l$-th layer. $s_l$ and $m_l$ are the spatial size of filter and output feature map. For DCNN, the training time per image is roughly three times of the complexity of convolution operator (one for forward convolution and two for backward propagation) [14]. Since DCNN is optimized in an iterative way, the computational complexity for DCNN is about $3O(TN \sum_{l=1}^{d} n_{l-1} \cdot s_l^2 n_l m_l^2)$, where $T$ is the number of iterations and $N$ is number of training samples. For LDA-SID, besides the convolution operator, the other computational expensive operators are the computation of between and within class scatters and the eigenvalue decomposition operator, whose time complexities are about $O(N \sum_{l=1}^{d} n_{l-1} s_l^2 m_l^2)$ and $O(\sum_{l=1}^{d} (n_{l-1} s_l^2)^3)$, respectively. Therefore, the time complexity of SID is about $O(N \sum_{l=1}^{d} n_{l-1} \cdot s_l^2 n_l m_l^2) + O(N \sum_{l=1}^{d} n_{l-1} s_l^2 m_l^2) + O(\sum_{l=1}^{d} (n_{l-1} s_l^2)^3))$. The complexity of class scatter computation is lower than that of convolution operator. The complexity of eigenvalue decomposition is relatively low because it is irrelative to number of training samples $N$ and the value of $s_l$ is usually small. Therefore, one can see that the computational efficiency of SID compared to traditional DCNN is mainly from two aspects. One is that optimal solution to SID is finally obtained by solving a eigenvalue decomposition problem, which is irrelative to the number of training samples, thus it is efficient for large-scale image training. The other is that the SID is of close-form solution, which avoids the iterative optimization. Table I lists the empirical training time of PCA-SID, LDA-SID, compared to PCANets on a training set of about 160,000 face images from 4,000 subjects. It is reported by using an unoptimized matlab implementation on a PC server with Intel Xeon CPU E5-2670 @2.60GHZ with 128GB memory. The time costs include the convolutional filter learning in each layer and the image convolution operator to generate responses which are the input of convolutional filter learning at the next layer. It is shown that SID learning from 160,000 face images takes about 9 hours, which is comparable to DCNN using GPU implementation.

TABLE I
Empirical training times of PCA-SID, LDA-SID and PCANets.

| Methods | Training time (h) |
|---|---|
| PCANet [5] | 8.84 |
| PCA-SID | 9.15 |
| LDA-SID | 9.40 |

## IV. Relation to previous methods

The SID is an extension of previous shallow IDL methods. For example, DFD-SID can be considered as an extension to DFD [22] by stacking the original DFD layer by layer, so that the representation ability is improved. Moreover, the structure of SID is similar to DCNN. One can also use stochastic gradient descent method to fine-tune the optimal solution of SID.

One recent work, namely PCANet [5], is composed of two layers of convolutional filters learned by PCA, followed by binary encoding and high-dimensional histogram feature extraction (about 204,800 dimensions). The first layer learning of PCANet and PCA-SID is the same; however, the remaining procedures are totally different. First, at the second or deeper layer of PCANet, the responses from different PCA filters (channels) are treated as different samples and a common PCA filter is learned for different channels. In SID, the responses from different channels are concatenated into a novel feature vector and different projective weights are learned for responses from different channels. More discriminant and complementary information can be extracted in this way. Second, the structure of PCANet is similar to traditional local descriptors like LBP [1], LQP [43], bag of words model [21] etc. After image encoding, the histogram features extracted from local overlapping or non-overlapping regions are concatenated to represent faces. Comparatively, SID is motivated from convolutional neural network (CNN), which consists of convolutional layer, pooling layer and fully connected layer. The extracted features are usually low-dimensional embeddings of convolutional responses. Third, besides using PCA or LDA, which involves second-order information for discriminant learning, we also propose DTA-SID, which incorporates high-order information and propose DFD-SID, which learns the discriminant convolutional filter and optimal sampling strategy with deep structure. From the work [5], one can see that following the structure of PCANet, the supervised method like LDANet fails to achieve better recognition performance than PCANet. In contrast, LDA-SID successfully performs better

than PCA-SID, which is reasonable because the supervised information is usually helpful for classification. It indicates that SID is able to extract discriminant information for face recognition.

## V. Experiments

There are two learning parts involved in SID for face recognition. One is the SID learning for face representation and the other is the metric/classifier learning for face recognition. Specifically, we collect a face database from internet, namely webface database (Fig. 2) as a training set to learn the stacked image descriptor. The webface database consists of more than 160,000 face images from about 4,000 subjects. There is no intersection between LFW and webface databases in terms of subjects. In classification phase of all the following experiments, unless explicit explanation, we adopt joint bayesian [6] to learn the discriminative metric on different databases, respectively. We examine the performance of SID on LFW and YTF face databases, which are considered as challenging unconstrained cases including many variations like pose, expression, occlusion simultaneously. We also evaluate the performance of different descriptors on traditional large face databases like FERET, Multi-PIE to examine the generalization of SID regarding to expression, lighting, aging and pose variations, respectively. In all the experiments, for SID related results, the face images are detected and aligned automatically [52] and there is no further pre-processing method adopted.

Fig. 2. Face examples from webface database.

### A. LFW

We first examine the performance of proposed method on LFW face database [19]. There are $13,233$ images from $5,749$ different persons, with large pose, occlusion, expression variations from the internet. Fig. 3 shows some face examples from LFW face database.

For SID learning, we empirically crop 11 face patches with different scales (shown in Fig. 4, the size of the first two patches is $110 \times 94$ and the size of the left 9 patches is $94 \times 94$) according to the detected face landmarks. We learn SID on each face patch, respectively, each of which outputs a 160-dimensional feature. In feature extraction phase, for each face patch, both the original and its mirror are adopted.

Fig. 3. Face examples from LFW database.

Therefore, the extracted feature representation for each face patch is of 320 dimension. In face classification, we follow the unrestricted protocol of LFW. The database of LFW is randomly divided into 10 splits. In each iteration, nine of ten face subsets are used to learn the classification metric (joint bayesian method adopted in this work) and the left one is used to test the classification performance. The final face recognition performance is reported as the mean accuracy of these 10 experiments.

Fig. 4. Examples of 11 face patches cropped from a face image.

*1) Comparison among different SIDs:* We compare the performance of Random-SID, PCA-SID, LDA-SID, DTA-SID and DFD-SID on the 2nd face patch, which is indicated using superscript $^{2nd}$ in Table II. For Random-SID, the values of image filters are uniformly sampled from $[0, 1]$. From the results shown in Table II, one can see that the supervised based SIDs achieves significantly higher face recognition accuracy than unsupervised one. It indicates that it is effective to extract discriminant information by strongly supervised learning in a layer-wise way. The performance of three proposed supervised learning based SIDs is similar. In the following, we report the performance of LDA-SID and DFD-SID as two representative ones of SIDs.

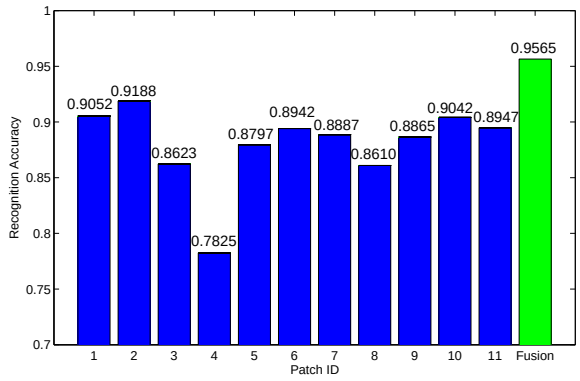
TABLE II
Mean accuracy (%) of different SIDs on LFW database with 320-dimensional feature.

| Methods | Accuracy |
|---|---|
| Rand-SID$^{2nd}$ | 66.05±0.40 |
| PCA-SID$^{2nd}$ | 71.37±0.34 |
| LDA-SID$^{2nd}$ | 91.88±0.58 |
| DTA-SID$^{2nd}$ | 91.78±0.42 |
| DFD-SID$^{2nd}$ | 90.75±0.50 |

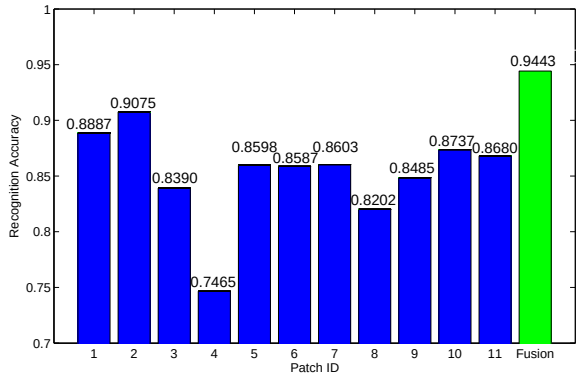
To show the advantage of deep structure of SID, we evaluate performance of the shallow version of LDA-SID where only one image descriptor layer is adopted. The shallow version of LDA-SID on 2nd face patch achieves the accuracy of $0.8097$ versus the accuracy of $0.9188$ by deep LDA-SID,

validating that the deep structure is useful to exploit more discriminant information helpful to face recognition. We also implement PCANet [5] on the same training set. Its accuracy on LFW database is $0.6700$, lower than PCA-SID and LDA-SID, indicating that the proposed SID has the advantage over PCANet to extract discriminant face representation for face recognition.

*2) Comparison among different face patches:* We train 11 SIDs on 11 face patches, respectively. For each face patch, the face image and its mirror are used to extract features, which is of 320 dimension. We also examine the performance by concatenating these features from 11 patches. In classification, the joint bayesian is applied to the SID feature to learn a discriminant metric following unrestricted protocol. Fig. 5 illustrates the face recognition accuracy of each patch and their combination. One can see that the first two face patches achieves the highest face recognition accuracy than other patches. There is complementary information in different face patches and by combining the 11 face patches, it achieves the face recognition accuracy as high as $95.65\%$ and $94.43\%$ for LDA-SID and DFD-SID, respectively. In the following, without explicit explanation, we report the performance of SID by combining the 11 face patches.




(a) LDA-SID




(b) DFD-SID

Fig. 5. Recognition accuracy of different cropped patches on LFW database.

*3) Comparison with state-of-the-art methods:*

*a) Results following unrestricted protocol:* Table III compares the performance of LDA-SID and DFD-SID with state-of-the-art descriptors including deep learning methods. We partially present representative results from the website[1]. All the results are conducted following unrestricted protocol. The proposed method significantly outperforms the conventional hand-crafted (e.g., LBP) or shallow learned face descriptors (e.g., DFD, Fisher vector faces), validating the advantage of deep structure for face representation learning. LDA-SID and DFD-SID achieve close performance to high-dim LBP [2], whose over-complete feature dimension is larger than 100K. Comparatively, the proposed SID feature is only of 3520 dimensions, which is much more compact than high-dim LBP and the storage and computational efficiency in feature matching are improved. Compared with deep learning related methods, SID methods achieve better recognition accuracy than ConvNet-RBM, indicating the effectiveness of SID. The performance of SID is worse than DeepFace (with $4096 \times 4$ DeepFace feature) and DeepID2 (with 4000 DeepID2 feature), which are also using deep structure and fine-tuned on large scale face database. It indicates that there is improvement space for SID in terms of accuracy and optimization. It is worth noting that one can use SID as a good initialization and use optimization method adopted in deep learning to improve the result further.

TABLE III
MEAN ACCURACY (%) OF DIFFERENT METHODS ON LFW DATABASE.

| Methods | Accuracy |
|---|---|
| LBP+PLDA [34] | 87.33±0.55 |
| DFD [22] | 84.02±0.44 |
| Fisher vector faces [36] | 93.03±1.05 |
| high-dim LBP [7] | 93.18±1.07 |
| PCANet [5] | 86.28±1.14 |
| DeepFace [41] | 97.35±0.25 |
| ConvNet-RBM [38] | 92.52±0.38 |
| DeepID2 [39] | 99.15±0.13 |
| LDA-SID | 95.65±0.44 |
| DFD-SID | 94.43±0.47 |

*b) Results following BLUFR:* To evaluate the performance of SID more completely, we also adopted the benchmark of large-scale unconstrained face recognition (BLUFR) proposed by Liao et al. [28] recently. It is indicated that the original protocol of LFW may be of a bias from many real applications of face recognition, where a low false accept rate (FAR) is usually desired. Thereby, a novel face recognition evaluation protocol, namely BLUFR is proposed, focusing on the verification rate and open-set identification rate at low FARs. In BLUFR, 10 trials of training and testing sets are randomly selected from original LFW dataset. The training set of each trial includes $3,524$ face images on average from $1,500$ subjects. The test set of each trial contains the remaining $4,249$ subjects with $9,708$ face images on average. Two face recognition performance measures, i.e., face verification and

[1] http://vis-www.cs.umass.edu/lfw/results.html

[2] To compare the representative ability of SID and high-dim LBP fairly, we report the performance without external training data in metric learning phase.

open-set identification are adopted. In face verification test, all the image pairs from the test set are compared and the computed matching scores are used for evaluation. There are about $156,915$ genuine matching scores and $46,960,863$ imposter matching scores in each trial on average. The open-set identification rate is defined as rank-1 recognition rate whose similarity score is above the threshold. In open-set identification test, in each trial, the test set is further randomly divided into three subsets, the gallery set, the genuine probe set $P_G$ and the imposter probe set $P_N$. We first randomly select $1,000$ subjects from the test set, one image from which is used to form the gallery set and the left forms the genuine probe set. The other images from the test set form the imposter probe set. Table IV lists the constitution of face images in BLUFR evaluation.

After SID extraction, we train the joint bayesian model on training set and evaluate the performance on testing set. The verification rate and open-set identification rate with different FARs are reported. The lower bound ($mean - std.$) of performance over 10 trills is illustrated in Tables V and VI, compared with state-of-the-art descriptors.

From the results, one can see that both LDA-SID and DFD-SID outperform traditional descriptor LBP and learning based descriptor LE. At the threshold of FAR=1%, LDA-SID improves the verification rate and open-set identification rate of LE by 31.2% and 7.8%, respectively, indicating the superiority of deep structure feature than the shallow one. LDA-SID (of 3250 dimension) outperforms the recently proposed high-dim LBP (of more than 100K dimensions) with much lower dimensionality of feature representation, indicating that SID is able to learn effective and compact feature representation for face recognition.

TABLE V
VERIFICATION RATE FOLLOWING BLUFR PROTOCOL. THE RESULTS ARE REPORTED AS THE MEAN ACCURACY (%) SUBTRACTED BY THE CORRESPONDING STANDARD DEVIATION OVER 10 TRIALS.

| Method | FAR=0.1% | FAR=1% |
|---|---|---|
| LBP [28] | 14.18 | 31.39 |
| LE [28] | 23.31 | 46.60 |
| high-dim LBP [28] | 41.66 | 65.84 |
| LDA-SID | 48.99 | 77.84 |
| DFD-SID | 40.70 | 70.18 |

TABLE VI
OPEN-SET IDENTIFICATION RATE AT RANK-1 FOLLOWING BLUFR PROTOCOL. THE RESULTS ARE REPORTED AS THE MEAN ACCURACY (%) SUBTRACTED BY THE CORRESPONDING STANDARD DEVIATION OVER 10 TRIALS.

| Method | FAR=1% | FAR=10% |
|---|---|---|
| LBP [28] | 8.82 | 16.61 |
| LE [28] | 11.26 | 20.73 |
| high-dim LBP [28] | 18.07 | 32.63 |
| LDA-SID | 19.13 | 39.22 |
| DFD-SID | 15.28 | 31.70 |

### B. Youtube Face Database

Youtube face (YTF) database [48] is a video version of LFW. The subjects in YTF is a subset of LFW. There are 3425 videos of 1595 subjects. For each video, we randomly select 10 frames to represent it. In testing phase, given two videos, we generate 100 scores between the two sets of 10 frames and the final score is obtained by computing the mean value of 100 scores. In each fold of evaluation, we apply the joint bayesian to 9 of 10 splits and evaluate the performance on the left split. The mean face recognition accuracy of 10 folds is reported (Table VII). The proposed stacked image descriptor outperforms most existing methods including the recent proposed discriminant deep metric learning (DDML). We can see that even with the direct cosine metric without learning, the extract SID based feature achieves comparable recognition accuracy with other methods, validating that SID is able to exploit strong discriminative features. With joint bayesian metric, the performance of SID is much improved, and close to DeepFace, indicating effectiveness of stacking local image descriptors.

TABLE VII
MEAN ACCURACY (%) OF DIFFERENT METHODS ON YTF DATABASE.

| Methods | Accuracy |
|---|---|
| LBP+MBGS [48] | 76.4±1.8 |
| VSOF+OSS [45] | 79.7±1.8 |
| DeepFace [41] | 91.4±1.1 |
| APEM (fusion) [25] | 79.1±1.5 |
| DDML [17] | 82.3±1.5 |
| LDA-SID+cosine | 82.7±0.7 |
| DFD-SID+cosine | 81.0±0.8 |
| LDA-SID+JB | 87.8±0.4 |
| DFD-SID+JB | 89.1±0.4 |

### C. FERET

The webface, LFW, YTF images are collected from the internet, where the variations of expression, pose, occlusion etc. are not constrained. We further evaluate SID on traditional large face databases in which the face variation is controlled to some extent. That is, following different testing protocols, it usually contains one or two variations from expression, pose, lighting, aging etc. so that the robustness of SID to these variations can be well evaluated separately. By applying the SID feature learned from the webface to traditional face databases directly, we can effectively examine the generalization ability of SID, compared with the state-of-the-art performance on these face databases. Two face databases, FERET [33], and Multi-PIE [12] are adopted.

The FERET database consists of a training set, a gallery set and four probe sets. The training set contains 1002 images. In testing phase, there are one gallery set with 1196 images from 1196 subjects. Four probe sets (fb, fc, dup1 and dup2) includes expression, illumination and aging variations. Fig. 6 show face examples from FERET face database.

Fig. 6. Face examples from FERET face database.

TABLE IV
OVERVIEW OF TRAINING AND TESTING SETS FOLLOWING BLUFR PROTOCOL.

| Image set | | No. Classes | No. Images | No. Genuine matches | No. Imposter matches |
|---|---|---|---|---|---|
| Train | | 1,500 | 3,524 | 85,341 | 6,122,185 |
| Test | All | 4,294 | 9,708 | 156,915 | 46,960,863 |
| | Gallery | 1,000 | 1,000 | - | - |
| | Genuine probe | 3,249 | 4,357 | - | - |

We compare the proposed SID methods with traditional hand-crafted descriptors and learning based descriptors. Table VIII lists the rank-1 recognition performance of different methods. Note that except the SID methods, all the other methods use the aligned face images with manually labeled coordinates. The method without 'PrePro' means there is no illumination pre-processing. From the results, one can see that without illumination pre-processing, most of the descriptors achieve satisfactory in expression subset, while their performance in lighting and aging is usually much worse. With proper pre-processing, the performance on lighting and aging subset is much improved. It is worth noting that WPCA is applied to gallery set, which does not very strictly follow the FERET protocols because the gallery information in testing set is used. The learning based methods usually achieve higher face recognition performance than hand-crafted ones, validating the superiority of learning based descriptors. Although the proposed SID is learned from the webface database, which differs from the FERET database, the performance of SID is competitive with other methods in most cases, even without illumination preprocessing, indicating that the generalization of SID is promising. Meanwhile, the performance of learned SID on lighting variation is not as good as state-of-the-art methods. This may be due to the fact that the lighting patterns are not sufficiently learned from the webface database. Overall speaking, the learned SID from webface database has good performance on expression and aging variations, but the performance on lighting variation still needs to be improved.

TABLE VIII
COMPARISON RESULTS (RANK-1 RECOGNITION RATE (%)) OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON FERET DATABASE.

| Methods | fb | fc | dup I | dup II |
|---|---|---|---|---|
| LBP [1] | 97.0 | 79.0 | 66.0 | 64.0 |
| LGBP [61] | 98.0 | 97.0 | 74.0 | 71.0 |
| HGPP [58] | 98.0 | 99.0 | 78.0 | 76.0 |
| LGXP [51] | 99.0 | 100.0 | 92.0 | 91.0 |
| POEM [46] | 97.6 | 95.0 | 77.6 | 76.2 |
| DT-LBP [30] | 99.0 | 63.0 | 67.0 | 48.0 |
| DLBP [31] | 99.0 | 48.0 | 68.0 | 55.0 |
| G-LQP [43] | 99.5 | 99.5 | 81.2 | 79.9 |
| DT-LBP+PrePro [30] | 99.0 | 100.0 | 84.0 | 80.0 |
| DLBP+PrePro [31] | 99.0 | 99.0 | 86.0 | 85.0 |
| DFD+PrePro [22] | 99.2 | 98.5 | 85.0 | 82.9 |
| DCP+PrePro [10] | 98.2 | 100.0 | 86.3 | 86.8 |
| G-LQP+WPCA [43] | 99.9 | 100.0 | 93.2 | 91.0 |
| DFD+PrePro+WPCA [22] | 99.4 | 100.0 | 91.8 | 92.3 |
| POEM+PrePro+WPCA [46] | 99.6 | 99.5 | 88.8 | 85.0 |
| LDA-SID | 98.5 | 98.5 | 92.4 | 94.9 |
| DFD-SID | 97.8 | 93.8 | 87.5 | 89.7 |

### D. Multi-PIE

Multi-PIE face database contains $754,204$ images from 337 subjects with 15 poses and 20 illuminations, captured in four sessions during different periods. In this part, we first follow the protocols adopted in [60] to compare the performance of SID with sparse representation. All the 249 subjects in session 1 are used. The 14 fontal images with 14 illuminations and neutral expression per subject are used to form the training set. In testing phase, 10 frontal images with illumination id $\{0, 2, 4, 6, 8, 10, 12, 14, 16, 18\}$ with neutral expression from session 2 to session 4 per subject are used. For SID, we adopt joint bayesian to learn the discriminant metric on training set. Table IX lists the recognition performance comparison of SID, compared with sparse representation classifier (SRC) and collaborative representation based classification with regularized least square (CRC-RLS). The results of SRC and CRC-RLS are copied from the paper [60] directly. It is shown that the proposed LDA-SID and DFD-SID outperform SRC and CRC-RLS in all the tests, demonstrating the effectiveness of SID representation.

TABLE IX
RECOGNITION RATE (%) OF SID COMPARED WITH SPARSE REPRESENTATION ON MULTIPIE DATABASE.

| | SRC | CRC-RLS | LDA-SID | DFD-SID |
|---|---|---|---|---|
| Session 2 | 93.9 | 94.1 | 93.9 | 96.8 |
| Session 3 | 90.0 | 89.3 | 96.4 | 95.8 |
| Session 4 | 94.0 | 93.3 | 96.5 | 96.2 |

We further examine the performance of SID following the protocol adopted in [26], [63]. In protocol I (corresponding to Setting I in [63]), we evaluate the robustness of SID to pose variation. The images with illumination 07 from four sessions are adopted. In training phase, we use all the images from the first 200 subjects. In testing phase, one frontal image of each subject is selected to form the gallery set and the remaining images are used to form the probe set with different poses. There are in total 137 images in gallery set and 137 images in each probe set corresponding to different poses. We also adopted protocol II (corresponding to Setting III in [63]) to evaluate the robustness of SID to pose and illumination variations simultaneously. The images from the session one with 7 poses and 20 illuminations are used. In training phase, we use all images from the first 100 subjects. In testing phase, one frontal image with illumination 07 of each subject is selected as the gallery set and the left images with 6 poses and 19 illuminations are used as the probe set. Fig. 7 illustrates an example of face with 7 poses. For SID, the networks learned from webface database are adopted directly to extract face features and the joint bayesian is applied to train the

classification model on Multi-PIE training set.

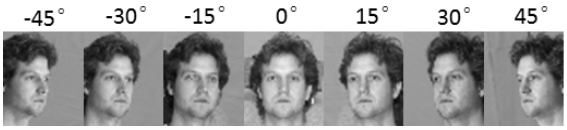



Fig. 7. Face example with different poses from Multi-PIE database.

Table X lists the face recognition performance across poses. In this part, we compare SID with state-of-the-art methods which do not utilize the prior pose information of probe images. It is shown that SID achieves competitive face recognition performance across poses. It significantly outperforms the traditional descriptor like LGBP and achieves higher accuracy than the conventional learning based methods like LE and CRBM. It achieves better performance than FA-EGFC, which utilizes 3D model to extract pose-invariant feature. SID based methods extract pose robust face features without synthesizing frontal face image explicitly. Its performance is competitive to the recently proposed deep learning based methods (FIP [63] and SPAE [20]), which is learned from the Multi-PIE face database, validating the robustness of SID to pose variation and the good generalization across face databases. The deep learning based method (RL [63]) with explicit frontal face synthesis achieve the highest face recognition in the case of large pose variation.

TABLE X
RECOGNITION RATES (%) OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON MULTI-PIE DATABASE ACROSS DIFFERENT POSES.

| Methods | -45° | -30° | -15° | 15° | 30° | 45° | Avg |
|---|---|---|---|---|---|---|---|
| LGBP [61] | 37.7 | 62.5 | 77.0 | 83.0 | 59.2 | 36.1 | 59.3 |
| FA-EGFC [26] | 84.7 | 95.0 | 99.3 | 99.0 | 92.9 | 85.2 | 92.7 |
| CRBM [18]+LDA | 80.3 | 90.5 | 94.9 | 96.4 | 88.3 | 75.2 | 87.6 |
| FIP+LDA [63] | 93.4 | 95.6 | 100.0 | 98.5 | 96.4 | 89.8 | 95.6 |
| RL+LDA [63] | 95.6 | 98.5 | 100.0 | 99.3 | 98.5 | 97.8 | 98.3 |
| SPAE [20] | 84.9 | 92.6 | 96.3 | 95.7 | 94.3 | 84.4 | 91.4 |
| LDA-SID | 92.3 | 96.0 | 98.0 | 96.7 | 94.7 | 91.0 | 94.8 |
| DFD-SID | 91.3 | 95.3 | 97.7 | 96.3 | 94.3 | 90.0 | 94.2 |

Table XI illustrates the face recognition performance across pose and illumination. In this case, the proposed SID methods significantly improve the performance of previous methods. It even enhances the recently proposed deep learning based method [63] by about 19%. On one hand, it validates that SID learned from webface database does have good generalization and robustness across different face databases and scenarios. On the other hand, it indicates that the deep networks trained with limited data in [63] does not exploit the discriminant and robust face representation sufficiently when pose and illumination variations are present simultaneously.

## VI. CONCLUSIONS

This paper proposes a stacked image descriptor (SID). The SID is optimized in a forward layer-wise way. In each image descriptor layer, based on the output of the previous layer, traditional shallow image descriptor learning method is applied to derive the optimal image descriptor. By concatenating image descriptor layer and max-pooling layer, we straightforwardly obtain the stacked image descriptor. Four SID implementations (PCA-SID, LDA-SID, DTA-SID and DFD-SID) are introduced. By applying SID to face recognition, we find that this strongly supervised optimization method at each layer is able to extract discriminant and compact face representation, which achieves good face recognition accuracy in the wild and also has good generalization performance on traditional frontal face recognition. Compared to deep learning, the time complexity of SID is lower when applied to large scale training data. The SID is a good choice to learn discriminative representation from large scale data, especially when GPU device is not available.

TABLE XI
RECOGNITION RATES (%) OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON MULTI-PIE DATABASE WITH POSE AND ILLUMINATION VARIATIONS.

| Recognition Rates on Different Poses | | | | | | | |
|---|---|---|---|---|---|---|---|
| Methods | -45° | -30° | -15° | 15° | 30° | 45° | Avg |
| Li [24] | 63.5 | 69.3 | 79.7 | 75.6 | 71.6 | 54.6 | 69.3 |
| RL+LDA [63] | 67.1 | 74.6 | 86.1 | 83.3 | 75.3 | 61.8 | 74.7 |
| LDA-SID | 86.0 | 95.5 | 98.8 | 98.5 | 95.7 | 87.2 | 93.6 |
| DFD-SID | 83.8 | 93.3 | 97.7 | 96.8 | 94.0 | 83.7 | 91.6 |

| Recognition Rates on Different Illuminations | | | | | | | |
|---|---|---|---|---|---|---|---|
| Methods | 00 | 01 | 02 | 03 | 04 | 05 | 06 |
| Li [24] | 51.5 | 49.2 | 55.7 | 62.7 | 79.5 | 88.3 | 97.5 |
| RL+LDA [63] | 72.8 | 75.8 | 75.8 | 75.7 | 75.7 | 75.7 | 75.7 |
| LDA-SID | 83.5 | 83.0 | 91.5 | 92.8 | 96.3 | 97.3 | 98.3 |
| DFD-SID | 81.5 | 78.9 | 85.1 | 91.6 | 95.8 | 96.4 | 96.7 |
| | 08 | 09 | 10 | 11 | 12 | 13 | 14 |
| Li [24] | 97.7 | 91.0 | 79.0 | 64.8 | 54.3 | 47.7 | 67.3 |
| RL+LDA [63] | 75.7 | 75.7 | 75.7 | 75.7 | 75.7 | 75.7 | 73.4 |
| LDA-SID | 98.6 | 98.1 | 96.8 | 95.2 | 90.8 | 85.1 | 97.2 |
| DFD-SID | 97.7 | 97.5 | 96.2 | 92.5 | 87.1 | 80.2 | 95.0 |
| | 15 | 16 | 17 | 18 | 19 | Avg | |
| Li [24] | 67.7 | 75.5 | 69.5 | 67.3 | 50.8 | 69.3 | |
| RL+LDA [63] | 73.4 | 73.4 | 73.4 | 72.9 | 72.9 | 74.7 | |
| LDA-SID | 97.7 | 98.3 | 97.5 | 97.3 | 83.6 | 93.6 | |
| DFD-SID | 96.6 | 96.6 | 96.3 | 95.2 | 82.3 | 91.6 | |

## ACKNOWLEDGEMENT

This work was supported by the Chinese National Natural Science Foundation Project #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

## REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns:application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006.
[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, July 1997.
[3] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou. Deep nonlinear metric learning with independent subspace analysis for face verification. In *ACM MM*, pages 749–752, 2012.
[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714, 2010.
[5] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *CoRR*, abs/1404.3606, 2014.

[6] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, pages 566–579, 2012.

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

[8] M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. In *ICML*, 2012.

[9] W. Deng, J. Hu, and J. Guo. Extended SRC: undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1864–1870, 2012.

[10] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *CoRR*, abs/1401.5311, 2014.

[11] S. Gao, I. W. Tsang, and L. Chia. Kernel sparse representation for image classification and face recognition. In *ECCV*, pages 1–14, 2010.

[12] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, 2010.

[13] Y. Guo, G. Zhao, M. Pietikäinen, and Z. Xu. Descriptor learning based on fisher separation criterion for texture classification. In *ACCV*, pages 185–198, 2010.

[14] K. He and J. Sun. Convolutional neural networks at constrained time cost. *CoRR*, abs/1412.1710, 2014.

[15] R. He, W. Zheng, and B. Hu. Maximum correntropy criterion for robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1561–1576, 2011.

[16] R. He, W. Zheng, T. Tan, and Z. Sun. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):261–275, 2014.

[17] J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, pages 1875–1882, 2014.

[18] G. B. Huang, H. Lee, and E. G. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012.

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[20] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (SPAE) for face recognition across poses. In *CVPR*, pages 1883–1890, 2014.

[21] Z. Lei, S. Z. Li, R. Chu, and X. Zhu. Face recognition with local gabor textons. In *ICB*, pages 49–57, 2007.

[22] Z. Lei, M. Pietikäinen, and S. Z. Li. Learning discriminant face descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):289–302, 2014.

[23] Z. Lei, D. Yi, and S. Z. Li. Discriminant image filter learning for face recognition with local binary pattern like representation. In *CVPR*, pages 2512–2517, 2012.

[24] A. Li, S. Shan, and W. Gao. Coupled bias-variance tradeoff for cross-pose face recognition. *IEEE Trans. Image Processing*, 21(1):305–315, 2012.

[25] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *CVPR*, pages 3499–3506, 2013.

[26] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *ECCV*, pages 102–115, 2012.

[27] X. Li, D. Dai, X. Zhang, and C. Ren. Structured sparse error coding for face recognition with occlusion. *IEEE Trans. Image Processing*, 22(5):1889–1900, 2013.

[28] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *IJCB*, 2014.

[29] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, 11(4):467–476, 2002.

[30] D. Maturana, D. Mery, and A. Soto. Face recognition with decision tree-based local binary patterns. In *ACCV*, pages 618–629, 2010.

[31] D. Maturana, D. Mery, and A. Soto. Learning discriminative local binary patterns for face recognition. In *FG*, pages 470–475, 2011.

[32] H. S. Mousavi, U. Srinivas, V. Monga, Y. Suo, M. Dao, and T. D. Tran. Multi-task image classification via collaborative, hierarchical spike-and-slab priors. In *ICIP*, 2014.

[33] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000.

[34] S. Prince, P. Li, Y. Fu, U. Mohammed, and J. H. Elder. Probabilistic models for inference about identity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1):144–157, 2012.

[35] M. Shao, M. Ma, and Y. Fu. Sparse manifold subspace learning. *Low-Rank and Sparse Modeling for Visual Analysis*, pages 117–132, 2014.

[36] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013.

[37] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep fisher networks for large-scale image classification. In *NIPS*, pages 163–171, 2013.

[38] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *ICCV*, pages 1489–1496, 2013.

[39] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *CoRR*, abs/1406.4773, 2014.

[40] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014.

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[42] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591, Hawaii, June 1991.

[43] S. ul Hussain, T. Napoleon, and F. Jurie. Face recognition using local quantized patterns. In *BMVC*, 2012.

[44] M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *CVPR*, pages 93–99, 2003.

[45] H. M. Vazquez, Y. Martínez-Díaz, and Z. Chai. Volume structured ordinal features with background similarity measure for video face recognition. In *ICB*, pages 1–6, 2013.

[46] N.-S. Vu and A. Caplier. Enhanced patterns of oriented edge magnitudes for face recognition and image matching. *IEEE Trans. Image Processing*, 21(3):1352–1365, 2012.

[47] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):372–386, 2012.

[48] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.

[49] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[50] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, January 2009.

[51] S. Xie, S. Shan, X. Chen, and J. Chen. Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Trans. Image Processing*, 19(5):1349–1361, 2010.

[52] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[53] Y. Xu, D. Zhang, J. Yang, and J. Yang. A two-phase test sample sparse representation method for use with face recognition. *IEEE Trans. Circuits Syst. Video Techn.*, 21(9):1255–1262, 2011.

[54] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang. Discriminant analysis with tensor representation. In *CVPR*, pages 526–532, 2005.

[55] M. Yang, L. Zhang, X. Feng, and D. Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3):209–232, 2014.

[56] M. Yang, L. Zhang, S. C. K. Shiu, and D. Zhang. Gabor feature based robust representation and classification for face recognition with gabor occlusion dictionary. *Pattern Recognition*, 46(7):1865–1878, 2013.

[57] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *CVPR*, pages 625–632, 2011.

[58] B. Zhang, S. Shan, X. Chen, and W. Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Trans. Image Processing*, 16(1):57–68, 2007.

[59] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang. Joint dynamic sparse representation for multi-view face recognition. *Pattern Recognition*, 45(4):1290–1298, 2012.

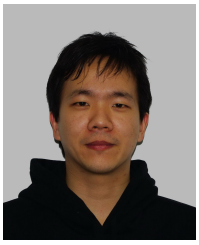
[60] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *ICCV*, pages 471–478, 2011.

[61] W. Zhang, S. Shan, W. Gao, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. In *ICCV*, pages 786–791, 2005.

[62] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *ICCV*, pages 2416–2423, 2013.

[63] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, pages 113–120, 2013.

**Zhen Lei** received the B.S. degree in automation from the University of Science and Technology of China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently an Associate Professor. He has published over 90 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an Area Chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.



**Dong Yi** received the B.S. degree in electronic engineering and the M.S. degree in communication and information system from Wuhan University, Wuhan, China, in 2003 and 2006, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He developed the face recognition modules and systems for the immigration control projects and 2008 Beijing Olympic Games. His research interests include unconstrained face recognition, heterogeneous face recognition, and deep learning.

**Stan Z. Li** received his B.Eng from Hunan University, China, M.Eng from National University of Defense Technology, China, and PhD degree from Surrey University, UK. He is currently a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He worked at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore. He was elevated to IEEE Fellow for his contributions to the fields of face recognition, pattern recognition and computer vision.

His research interest includes pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published over 200 papers in international journals and conferences, and authored and edited 8 books. He was an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and is acting as the editor-in-chief for the Encyclopedia of Biometrics. He served as a program co-chair for the International Conference on Biometrics 2007, 2009 and 2015, and has been involved in organizing other international conferences and workshops in the fields of his research interest.