Learning Kernel Extended Dictionary for Face Recognition

Ke-Kun Huang, *Member, IEEE*, Dao-Qing Dai, *Member, IEEE*, Chuan-Xian Ren, *Member, IEEE*, and Zhao-Rong Lai, *Student Member, IEEE*

***Abstract*— A sparse representation classifier (SRC) and a kernel discriminant analysis (KDA) are two successful methods for face recognition. An SRC is good at dealing with occlusion, while a KDA does well in suppressing intraclass variations. In this paper, we propose kernel extended dictionary (KED) for face recognition, which provides an efficient way for combining KDA and SRC. We first learn several kernel principal components of occlusion variations as an occlusion model, which can represent the possible occlusion variations efficiently. Then, the occlusion model is projected by KDA to get the KED, which can be computed via the same kernel trick as new testing samples. Finally, we use structured SRC for classification, which is fast as only a small number of atoms are appended to the basic dictionary, and the feature dimension is low. We also extend KED to multikernel space to fuse different types of features at kernel level. Experiments are done on several large-scale data sets, demonstrating that not only does KED get impressive results for nonoccluded samples, but it also handles the occlusion well without overfitting, even with a single gallery sample per subject.**

***Index Terms*— Face occlusion, face recognition, kernel discriminant analysis (KDA), sparse representation classifier (SRC).**

I. Introduction

FACE recognition has attracted much attention in image processing, pattern recognition, and computer vision because of its wide range of applications, such as access control and video surveillance [1]. After many years of investigation, face recognition is still very challenging due to the dramatic intraclass variations, such as expression, viewing angle, lighting conditions, and occlusions. For different communities to verify their methods, many large-scale face data sets, such as Face Recognition Technology (FERET) [2], Pose, Expression, Accessories, and Lighting (PEAL) of Chinese Academy of Sciences (CAS-PEAL) [3], and Labeled Face in the Wild (LFW) [4], have been established for evaluation.

Partial face occlusion, such as wearing sunglasses, hat, or scarf, is one of the most challenging problems in real world. Occluded sample is outlier, thus the traditional methods, such as principal component analysis (PCA) [5], linear discriminant analysis (LDA) [6], locality preserving projection [7], and marginal fisher analysis [8], cannot deal with face occlusion well.

Wright *et al.* [9] proposed a general classification algorithm for face recognition called the sparse representation classifier (SRC), where an input testing image is coded as a sparse linear combination of training images via l_1 minimization. The SRC leads to higher classification accuracy compared with many well-known face recognition methods. In addition, it handles the problem of random pixel corruption and small-scale face occlusion well by using identity matrix as the extended dictionary.

However, the original SRC is not robust to large contiguous occlusion [9]. The main reason is that the contiguous occlusion violates the assumption that the occlusion has sparse representation with respect to the identity matrix dictionary. There are many works to extend SRC, such as correntropy-based sparse representation [10], structured sparse error coding (SSEC) [11], regularized robust coding (RRC) [12], and robust kernel representation with statistical local features (SLF-RKR) [13]. The above methods achieve better performance for occlusion, but they may overfit probe samples, i.e., they are likely to regard nonoccluded samples as occluded samples. Thus, they will decline the performance for nonoccluded samples at the same time.

Deng *et al.* [14] proposed the extended sparse representation-based classifier (ESRC), which applies an auxiliary intraclass variant dictionary to represent the possible variation between training images and testing ones. ESRC improves the performance for occluded samples and that for nonoccluded samples simultaneously. In [15], superposed SRC (SSRC) is proposed. Similar to ESRC, SSRC adopts the sample-to-centroid differences as the auxiliary dictionary. Different from ESRC, SSRC constructs the basic dictionary by the class centroids instead of all the training samples. Though the number of atoms of the dictionary of SSRC is less than that of ESRC, it is still large. Thus, the classification of SSRC and ESRC is not fast enough.

Manuscript received February 3, 2015; revised January 14, 2016; accepted January 22, 2016. This work was supported in part by the National Science Foundation of China under Grant 61203248, Grant 11171354, Grant 90920007, Grant 61375033, Grant 61403164, and Grant 61572536, in part by the Ministry of Education of China under Grant SRFDP-20120171120007 and Grant 20120171110016, in part by the Natural Science Foundation of Guangdong Province under Grant S2013020012796, in part by the Fundamental Research Funds for the Central Universities under Grant 13lgpy26, and in part by the Foundation for Distinguished Young Talents in Higher Education of Guangdong, China, under Grant 2013LYM 0085. *(Corresponding author: Dao-Qing Dai.)*

K.-K. Huang is with the Intelligent Data Center, Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China, and also with the Department of Mathematics, Jiaying University, Meizhou 514015, China (e-mail: kkcocoon@163.com).

D.-Q. Dai, C.-X. Ren, and Z.-R. Lai are with the Intelligent Data Center, Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China (e-mail: stsddq@mail.sysu.edu.cn; rchuanx@mail.sysu.edu.cn; laizhr@mail2.sysu.edu.cn).

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/TNNLS.2016.2522431

On the other hand, the kernel discriminant analysis (KDA) [16] is another successful method for face recognition. KDA projects the data onto a nonlinear discriminant subspace to suppress intraclass variations and maximize the gap between the images from different persons. Different from LDA or SRC, where only l_2-norm distance is used, an advantage of the kernel method is that it can exploit different distances for different features. By selecting appropriate kernel, the performance of KDA is better than LDA and SRC. Because of its good performance, there are many methods extending KDA in recent years, such as multiple kernel learning for dimensionality reduction [17], regularized KDA [18], multiscale local phase quantization using the kernel fusion of multiple descriptors [19], and band-reweighed Gabor kernel embedding using multiple orientation and scale transforms (MOST) [20]. The above KDA-based methods achieve the state-of-the-art performances for nonoccluded samples, but they are not good at dealing with face occlusion, because the occluded sample is the outlier.

Although SRC and KDA have shown powerful abilities for face recognition, few works have been proposed to integrate them together for better performance. In this paper, we propose kernel extended dictionary (KED) for face recognition, which provides an efficient way for combining KDA and SRC. It should be mentioned that KED is not the simple combination of KDA and SRC. The key of KED is how to construct and compute the extended dictionary, so that it can represent the possible occlusion variations efficiently in the kernel space.

It is worth mentioning that the proposed KED is different from the dictionary learning methods [21]–[24]. Given a set of training samples, these methods seek the dictionary that leads to the best sparse representation for each subject in the training set, as well as make the representation more discriminative. Our method is different from these methods as follows. First, they do not learn discriminant subspace, while our method is based on a nonlinear discriminant subspace. Second, the dictionary they learn is only for the subjects in the training set. They cannot handle the situation when new subjects enrolled, while our method can, even with a single gallery sample per subject.

Recently, deep learning, in particular convolutional neural network (CNN), achieves very promising results for face recognition, such as DeepFace [25], WebFace [26], DeepID [27], and its extensions [28]–[30]. CNN is a feed-forward architecture, involving multiple computational layers that alternate linear operations, such as convolutions and nonlinear operations, such as max pooling. Unlike the traditional hand-crafted features, the CNN learning-based features are more robust to complex intrapersonal variations. Despite their promising performance, deep architectures come with some challenges. First, it remains unclear how to design a good CNN architecture to adapt to a specific classification task due to the lack of theoretical guidance [31]. Turning parameters again and again is the main job to design the architecture. Second, the training data are very important for the performance of face recognition. To achieve better results, we need to add more faces, which are collected in the same situation as the evaluation data set [32]. Third, because of the requirement for large amounts of training data, it needs high computational cost during the training process [33]. Finally, as a feature extraction method, CNN still needs to use some classifiers, such as joint Bayesian (JB) [34] or metric-learning methods [35], to learn a more efficient low-dimensional representation to distinguish faces of different identities [32]. So, finding some classifiers that are typically less costly to train and evaluate is still competitive for face recognition.

The main contributions of this paper are listed as follows.

1) Propose an occlusion model. The occlusion model is several kernel principal components of occlusion variations, which can represent the possible occlusion variations efficiently.
2) Propose KED. KED provides an efficient way for combining KDA and SRC, which not only extracts the nonlinear discriminant feature to deal with nonoccluded samples, but also exploits sparse representation to deal with occlusion. The classification procedure is fast as only a small number of dictionary atoms are appended, and the feature dimension is low.
3) Extend KED to multikernel space to fuse different types of features at kernel level.

The remainder of this paper is organized as follows. In Section II, we briefly introduce SRC and ESRC. In Section III, we describe the proposed methods in detail. The experimental results are given in Section IV. Finally, we provide the conclusion in Section V.

II. Related Works

Because our method is based on SRC and ESRC, to facilitate describing our method, we briefly introduce the two related methods first.

A. Sparse Representation Classifier

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ be the n samples in the gallery set, where d is the feature dimension. The class label of $\mathbf{x}_i$ is assumed to be c_i. Given a probe sample $\mathbf{y} \in \mathbb{R}^{d\times 1}$, SRC computes its sparse representation coefficient $\boldsymbol{\beta} \in \mathbb{R}^{n\times 1}$ via

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{1}$$

Let $\delta_c(\boldsymbol{\beta}) \in \mathbb{R}^{n\times 1}$ be a vector whose entries are zero except those associated with class c. SRC assigns $\mathbf{y}$ to the class that minimizes the residual

$$r_c(\mathbf{y}) \doteq \|\mathbf{y} - \mathbf{X}\delta_c(\boldsymbol{\beta})\|_2. \tag{2}$$

SRC represents $\mathbf{y}$ collaboratively by the samples of all classes subject to the condition that the coefficient is sparse. If $\mathbf{y}$ is from class c, it is more likely that we can use only a few samples in the class to represent $\mathbf{y}$ with a good accuracy.

In many practical face recognition scenarios, the probe sample $\mathbf{y}$ may be partially corrupted or occluded. In this case, SRC computes its sparse representation via

$$\min_{\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}} \left\|\mathbf{y} - [\mathbf{X}, \mathbf{I}]\begin{bmatrix}\boldsymbol{\beta}\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right\|_2^2 + \lambda\left\|\begin{bmatrix}\boldsymbol{\beta}\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right\|_1 \tag{3}$$

where $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix.

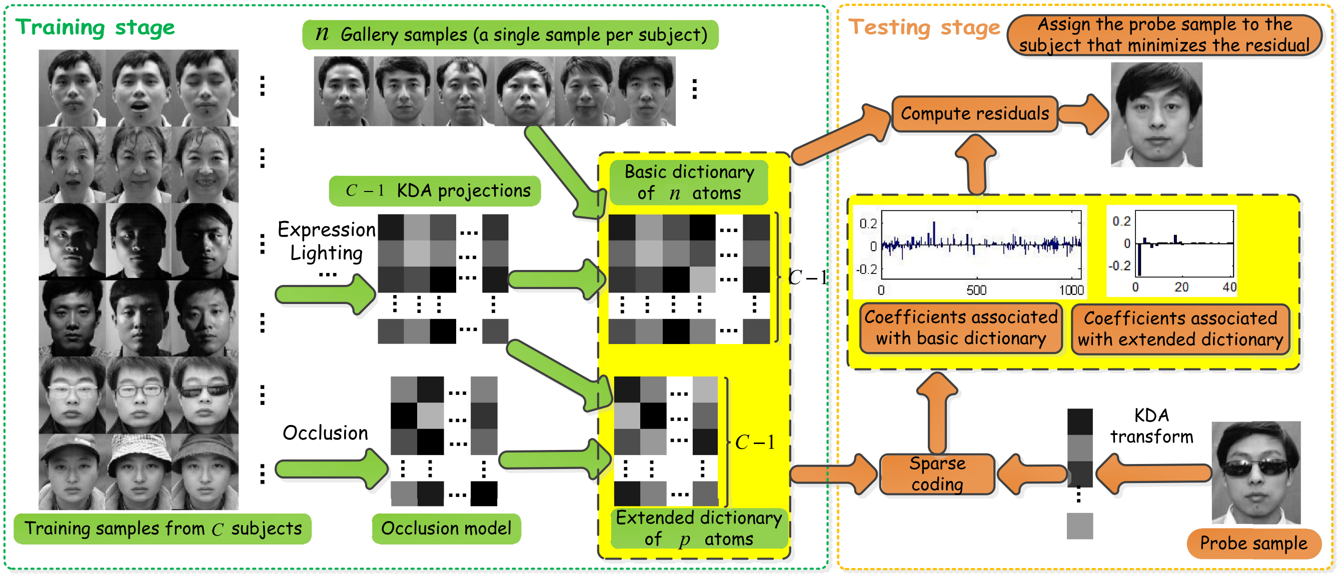

Training stage
n Gallery samples (a single sample per subject)
$C-1$ KDA projections
Expression
Lighting
...
Occlusion
Training samples from C subjects
Occlusion model
Basic dictionary of n atoms
$C-1$
$C-1$
Extended dictionary of p atoms
Testing stage
Assign the probe sample to the subject that minimizes the residual
Compute residuals
0.2
0
-0.2
0
500
1000
0.2
0
-0.2
0
20
40
Coefficients associated with basic dictionary
Coefficients associated with extended dictionary
Sparse coding
KDA transform
Probe sample

Fig. 1. Pipeline of KED. In the training stage, we first train the $C-1$ projections of KDA from the training samples of C subjects and learn p kernel principal components of occlusion variations as an occlusion model. Then, the gallery samples are projected by KDA to get the basic dictionary, and the occlusion model is also projected by the same kernel trick to get the extended dictionary. In the testing stage, we first project a probe sample by KDA, and then find its sparse representation in terms of the basic dictionary and the extended dictionary. Finally, we assign the probe sample to the subject that minimizes the reconstruction residual.

B. Extended Sparse Representation Classifier

ESRC assumes that the intraclass variations of different subjects are sharable, and the variant bases can be acquired either from the gallery samples themselves or from the training samples outside of the gallery. Let $\tilde{\mathbf{x}}_i$ be the ith samples in the training set to learn the intraclass variant bases, and $\tilde{\mu}_i$ be the natural sample, corresponding to $\tilde{\mathbf{x}}_i$ in the training set. If the corresponding natural sample is not available, we can use the class mean instead. Then, the intraclass variant bases $\tilde{\mathbf{E}}$ can be computed from the samples by subtracting the natural sample or class mean

$$\tilde{\mathbf{E}} = [\tilde{\mathbf{x}}_1 - \tilde{\mu}_1, \tilde{\mathbf{x}}_2 - \tilde{\mu}_2, \ldots, \tilde{\mathbf{x}}_{\tilde{n}} - \tilde{\mu}_{\tilde{n}}] \in \mathbb{R}^{d\times\tilde{n}}. \quad (4)$$

ESRC finds a sparse representation of a probe sample in terms of the gallery set as well as the intraclass variant bases

$$\min_{\beta,\tilde{\beta}} \left\| \mathbf{y} - [\mathbf{X}, \tilde{\mathbf{E}}] \begin{bmatrix} \beta \\ \tilde{\beta} \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \beta \\ \tilde{\beta} \end{bmatrix} \right\|_1. \quad (5)$$

The nonzero coefficients are expected to concentrate on the gallery samples with the same identity as the probe sample and on the intraclass variant bases. ESRC has better generalization ability than SRC for face recognition, even with a single image per class in gallery. But, the size of intraclass variant bases using (5) is equal to that of the training set. The classification procedure of ESRC is much slower than that of the original SRC, because the l_1-minimization problem is time consuming, and greatly affected by the number of atoms of dictionary.

III. Proposed Method

In this section, we describe the proposed method in detail. In Section III-A, we propose the KED first. Then, we discuss some properties of KED in Section III-B.

A. Kernel Extended Dictionary

KDA learns a nonlinear discriminant subspace to suppress intraclass variation and maximize the gap between different subjects. By selecting appropriate kernel, the performance of KDA can be better than LDA and SRC. But, KDA is not good at dealing with face occlusion, because the occluded sample is the outlier. On the other hand, SRC does well in occlusion. But, SRC is just a classification method, which does not learn a discriminant subspace. In order to provide an efficient way for combining KDA and SRC, we propose the KED, which extracts nonlinear discriminant features and exploits the sparsity in face recognition.

Fig. 1 shows the pipeline of the proposed KED. First, we train KDA from the training samples of C subjects. Second, we learn p principal components from the differences between occlusion samples and corresponding normal samples in the kernel space. Third, we get the basic dictionary from the gallery samples projected by KDA, and the extended dictionary from the p principal components projected by KDA. Finally, we use the structured sparse representation to classify a probe sample after KDA projection. We will prove as follows that the extended dictionary can be attained by the same kernel trick as new probe samples, which is the key of the proposed method.

1) Kernel Discriminant Analysis: The proposed method first uses KDA to suppress intraclass variations and maximize the gap between different subjects, especially for nonoccluded samples. Furthermore, KED is also based on KDA. So, we first give the resulting formula of KDA, and the detail process of KDA can be found in [16].

We consider the problem in a feature space F induced by a nonlinear mapping: $\phi : \mathbb{R}^d \rightarrow F$. For a properly chosen ϕ, an inner product $\langle \cdot, \cdot \rangle$ can be defined on F

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = k(\mathbf{x}, \mathbf{x}')$$

which makes a so-called reproducing kernel hilbert space.

Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{\hat{n}}] \in \mathbb{R}^{d \times \hat{n}}$ be the $\hat{n}$ samples to learn the projection vectors of KDA. According to [16], we can learn a projection vector of KDA: $\mathbf{v}^{\phi} = \sum_{i=1}^{\hat{n}} \alpha_i \phi(\tilde{\mathbf{x}}_i)$.

Then, we can project a sample $\mathbf{x}$ in the kernel space to the vector, where the inner product in the kernel space can be computed by a kernel function without knowing the nonlinear mapping explicitly

$$\langle \mathbf{v}^{\phi}, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{\hat{n}} \alpha_i \langle \phi(\tilde{\mathbf{x}}_i), \phi(\mathbf{x}) \rangle = \boldsymbol{\alpha}^T \mathbf{K}(:, \mathbf{x}) \tag{6}$$

where $\mathbf{K}(:, \mathbf{x}) = [k(\tilde{\mathbf{x}}_1, \mathbf{x}), k(\tilde{\mathbf{x}}_2, \mathbf{x}), \ldots, k(\tilde{\mathbf{x}}_{\hat{n}}, \mathbf{x})]^T$.

2) Learning Occlusion Model in Kernel Space: Most of the intraclass variations can be suppressed by KDA. However, because occluded sample is outlier, KDA cannot suppress the occlusion variation well. To deal with this problem, we propose an occlusion model. It consists of several kernel principal components of occlusion variations, which can represent the possible variations efficiently by only a few number of principal components. Without loss of generality, suppose the first $\tilde{n}$ samples in $\tilde{\mathbf{X}}$ are used to learn the occlusion model.

The covariance of occlusion variations in F is

$$\begin{aligned} \tilde{\mathbf{S}}^{\phi} &= \sum_{i=1}^{\tilde{n}} (\phi(\tilde{\mathbf{x}}_i) - \phi(\tilde{\mu}_i))(\phi(\tilde{\mathbf{x}}_i) - \phi(\tilde{\mu}_i))^T \\ &= \sum_{i=1}^{\tilde{n}} \Phi(i)\Phi^T(i) \end{aligned} \tag{7}$$

where $\Phi(i) = \phi(\tilde{\mathbf{x}}_i) - \phi(\tilde{\mu}_i)$, $\tilde{\mathbf{x}}_i$ is the occluded sample, and $\tilde{\mu}_i$ is the corresponding natural sample or class mean.

We seek the optimal projection $\tilde{\mathbf{v}}^{\phi}$ that maximizes the variance after projection

$$\begin{aligned} &\max_{\tilde{\mathbf{v}}^{\phi}} \ (\tilde{\mathbf{v}}^{\phi})^T \tilde{\mathbf{S}}^{\phi} \tilde{\mathbf{v}}^{\phi} \\ &\text{s.t. } \langle \tilde{\mathbf{v}}^{\phi}, \tilde{\mathbf{v}}^{\phi} \rangle = 1 \end{aligned} \tag{8}$$

where $\tilde{\mathbf{v}}^{\phi}$ is the principal component of $\tilde{\mathbf{S}}^{\phi}$, that is

$$\tilde{\mathbf{S}}^{\phi} \tilde{\mathbf{v}}^{\phi} = \lambda \tilde{\mathbf{v}}^{\phi}. \tag{9}$$

We can find the solution $\tilde{\mathbf{v}}^{\phi} = \sum_{i=1}^{\tilde{n}} \tilde{\alpha}_i \Phi(i)$, where $\tilde{\boldsymbol{\alpha}} = [\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_{\tilde{n}}]^T$ satisfies the following equation:

$$\tilde{\mathbf{K}} \tilde{\boldsymbol{\alpha}} = \lambda \tilde{\boldsymbol{\alpha}} \tag{10}$$

where

$$\begin{aligned} \tilde{\mathbf{K}}_{i,j} &= \langle \Phi(i), \Phi(j) \rangle \\ &= \langle \phi(\tilde{\mathbf{x}}_i) - \phi(\tilde{\mu}_i), \phi(\tilde{\mathbf{x}}_j) - \phi(\tilde{\mu}_j) \rangle \\ &= \langle \phi(\tilde{\mathbf{x}}_i), \phi(\tilde{\mathbf{x}}_j) \rangle - \langle \phi(\tilde{\mathbf{x}}_i), \phi(\tilde{\mu}_j) \rangle \\ &\quad - \langle \phi(\tilde{\mu}_i), \phi(\tilde{\mathbf{x}}_j) \rangle + \langle \phi(\tilde{\mu}_i), \phi(\tilde{\mu}_j) \rangle. \end{aligned} \tag{11}$$

Then, the top p kernel principal components $\tilde{\mathbf{V}}^{\phi} = [\tilde{\mathbf{v}}_1^{\phi}, \tilde{\mathbf{v}}_2^{\phi}, \ldots, \tilde{\mathbf{v}}_p^{\phi}]$ constitute the occlusion model.

3) Constructing Kernel Extended Dictionary via Kernel Trick: We can project a sample $\mathbf{x}$ by KDA without knowing the nonlinear mapping explicitly via (6). Similarly, we can also use the same kernel trick to project the kernel principal component $\tilde{\mathbf{v}}^{\phi} = \sum_{i=1}^{\tilde{n}} \tilde{\alpha}_i \Phi(i)$ by KDA with the following equation:

$$\begin{aligned} (\mathbf{v}^{\phi})^T \tilde{\mathbf{v}}^{\phi} &= \langle \mathbf{v}^{\phi}, \tilde{\mathbf{v}}^{\phi} \rangle = \left\langle \sum_{i=1}^{\hat{n}} \alpha_i \phi(\tilde{\mathbf{x}}_i), \sum_{j=1}^{\tilde{n}} \tilde{\alpha}_j \Phi(j) \right\rangle \\ &= \left\langle \sum_{i=1}^{\hat{n}} \alpha_i \phi(\tilde{\mathbf{x}}_i), \sum_{j=1}^{\tilde{n}} \tilde{\alpha}_j (\phi(\tilde{\mathbf{x}}_j) - \phi(\tilde{\mu}_j)) \right\rangle \\ &= \boldsymbol{\alpha}^T \mathbf{K} \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^T \mathbf{K}^{\mu} \tilde{\boldsymbol{\alpha}} \end{aligned} \tag{12}$$

where $\mathbf{K}_{i,j} = \langle \phi(\tilde{\mathbf{x}}_i), \phi(\tilde{\mathbf{x}}_j) \rangle$, $\mathbf{K}_{i,j}^{\mu} = \langle \phi(\tilde{\mathbf{x}}_i), \phi(\tilde{\mu}_j) \rangle$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{\hat{n}}]^T$ is a coefficient vector of KDA, and $\tilde{\boldsymbol{\alpha}} = [\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_{\tilde{n}}]^T$ is a coefficient vector of the occlusion model.

Then, we project the occlusion model by multiple directions of KDA to form the KED.

4) Main Steps of KED: The main steps of KED can be summarized as follows.

1) Train the $C - 1$ projections of KDA, i.e., $\mathbf{V}^{\phi} = [\mathbf{v}_1^{\phi}, \mathbf{v}_2^{\phi}, \ldots, \mathbf{v}_{C-1}^{\phi}]$, from the training set $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{\hat{n}}]$, where C is the number of class. The training set can be outside the gallery set.
2) Train the p principal components of occlusion variations in the kernel space, i.e., $\tilde{\mathbf{V}}^{\phi} = [\tilde{\mathbf{v}}_1^{\phi}, \tilde{\mathbf{v}}_2^{\phi}, \ldots, \tilde{\mathbf{v}}_p^{\phi}]$, from a subset of $\tilde{\mathbf{X}}$, as the occlusion model.
3) Project the gallery set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ to construct the basic dictionary by KDA via (6), that is
$$\mathbf{D} = (\mathbf{V}^{\phi})^T \phi(\mathbf{X}). \tag{13}$$
Note that the subjects in the gallery set can be never seen in the training set.
4) Project $\tilde{\mathbf{V}}^{\phi}$ by KDA to construct the KED via (12), that is
$$\tilde{\mathbf{D}} = (\mathbf{V}^{\phi})^T \tilde{\mathbf{V}}^{\phi}. \tag{14}$$
5) Project a probe sample $\mathbf{y}$ by KDA
$$\mathbf{y}_{\text{KDA}} = (\mathbf{V}^{\phi})^T \phi(\mathbf{y}). \tag{15}$$
6) Find a sparse representation for $\mathbf{y}_{\text{KDA}}$ in terms of the basic dictionary $\mathbf{D}$ as well as the extended dictionary $\tilde{\mathbf{D}}$
$$\min_{\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}} \left\| \mathbf{y}_{\text{KDA}} - [\mathbf{D}, \tilde{\mathbf{D}}] \begin{bmatrix} \boldsymbol{\beta} \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{\beta} \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} \right\|_1. \tag{16}$$
7) Assign $\mathbf{y}$ to the class that minimizes the residual
$$r_c(\mathbf{y}_{\text{KDA}}) = \left\| \mathbf{y}_{\text{KDA}} - [\mathbf{D}, \tilde{\mathbf{D}}] \begin{bmatrix} \delta_c(\boldsymbol{\beta}) \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} \right\|_2. \tag{17}$$

Note that according to (7), we first map the samples into F before computing the difference between the normal sample $\tilde{\mu}_i$ and the occluded sample $\tilde{\mathbf{x}}_i$, instead of $\sum_{i=1}^{\tilde{n}} (\phi(\tilde{\mathbf{x}}_i - \tilde{\mu}_i))(\phi(\tilde{\mathbf{x}}_i - \tilde{\mu}_i))^T$. If $\phi(\mathbf{y})$ has an occlusion variation similar to $\tilde{\mathbf{v}}^{\phi}$, then $\phi(\mathbf{y})$ can be decomposed as

$$\phi(\mathbf{y}) \approx \phi(\mathbf{x}) - \tilde{\mathbf{v}}^{\phi} \tag{18}$$

where $\phi(\mathbf{x})$ is the gallery sample belonging to the same class as $\mathbf{y}$ in F. Thus, $\mathbf{y}_{\text{KDA}} = (\mathbf{V}^{\phi})^T \phi(\mathbf{y})$ can be decomposed as

$$(\mathbf{V}^{\phi})^T \phi(\mathbf{y}) \approx (\mathbf{V}^{\phi})^T \phi(\mathbf{x}) - (\mathbf{V}^{\phi})^T \tilde{\mathbf{v}}^{\phi}. \tag{19}$$

According to our experiments, the occlusion model can represent the possible occlusion variations efficiently. As shown in Fig. 1, there is one significant coefficient associated with the extended dictionary for the sparse representation of a probe occluded sample. So, the relationship of occlusion variation is maintained in F by the nonlinear mapping ϕ, and we can successfully classify the occluded sample by neglecting the occlusion.

It should be mentioned that though the formation of (16) is the same as (5), KED and ESRC are different as follows. First, KED is based on KDA, i.e., both the dictionary and the probe sample are projected by KDA, thus the intraclass variations are suppressed, while ESRC is based on the original feature. Second, the extended dictionary of the proposed method is attained by the proposed occlusion model, which is composed by only a few number of atoms, while that of ESRC is the samples by subtracting the corresponding natural sample, where the number of atoms is much larger than that of the proposed method.

In real world, though there are different kinds of occlusion, the number of common large-scale occlusion types is limited, such as occluded by sunglasses, hat, or scarf. So, as many papers, such as ESRC [14] and SSRC [15], we can build the occlusion model beforehand.

Furthermore, our model can handle many occlusion types. When there is a new type of occlusion, we can append only a small number of atoms to the dictionary for sparse representation, because the occlusion model is the top several principal components of occlusion variations in the kernel space, other than ESRC, where the extended dictionary is all the training samples by subtracting the corresponding natural sample. So, the occlusion model can deal with many occlusion types simultaneously, without increasing too much computational cost.

5) Constructing the Kernel Function: KED is a kernel method, which is important to construct an appropriate kernel function. Different from LDA or SRC, where only l_2-norm distance is used, an advantage of kernel method is that it can exploit different distances for different features. For example, the cosine of the difference of gradient directions can get better performance with gradient feature [36], the χ^2 distance does well with Local Binary Patterns (LBP) feature [37], and the radial basis function (RBF) kernel is good with the Gabor feature [20]. Because of the good performance of the LBP feature [1], [38], [39], if not specified, we use the following RBF kernel with χ^2 distance based on the LBP feature:

$$k_L(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\chi^2(\mathbf{x}, \mathbf{x}')}{\sigma^2}\right) \tag{20}$$

where $\chi^2(\mathbf{x}, \mathbf{x}') = \sum_i ((x_i - x_i')^2 / x_i + x_i' + \epsilon)$.

The LBP feature is similarly derived from [37], except that we extract the feature from the images under multiple scales. We first build an image pyramid of three scales, with a scaling factor $\sqrt{2}$. Then, we divide the image of each scale into a grid of cells and histogram, the uniform LBPs within each cell. Finally, the cell-level histograms are concatenated to produce the final LBP feature. The LBP codes are computed using eight sampling points on a circle of radius 2. We set the standard deviation $\sigma = 3\mu$, where μ is the mean distance of all pairs of training samples.

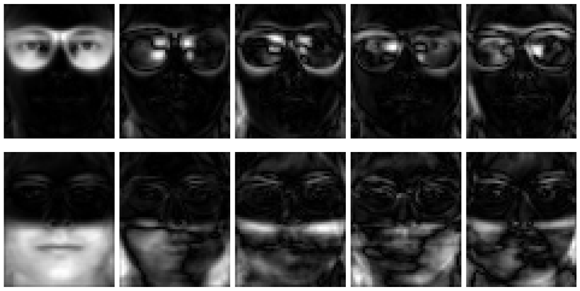


Fig. 2. Occlusion model with linear kernel for wearing sunglasses (first row) and wearing scarf (second row) on the AR data set. The first principal component (first column) represents most of the differences between the occlusion samples and the corresponding natural samples.

B. Extension and Analysis

In this section, we analyze some properties of the proposed method. First, we present a linear version of the occlusion model to give an intuitional interpretation. Second, we extend KED to multikernel space to fuse different types of features. Finally, we analyze the computational complexity of KED.

1) Linear Visualization for Occlusion Model: When we set the nonlinear mapping as a linear one: $\phi(\mathbf{x}) = \mathbf{x}$, i.e., using linear kernel, the occlusion model can be visualized. Fig. 2 shows the occlusion model with linear kernel for wearing sunglasses and wearing scarf on the AR data set. We can find that the energy is concentrated in the first principal component. So, we can append only a small number of atoms to the basic dictionary to represent the occlusion variations. Compared with ESRC, which appends a large number of atoms to the basic dictionary, the classification procedure of the proposed method is much faster. It will be demonstrated in the experiment as Table II.

2) Multikernel Extended Dictionary: To capture rich enough information available in the face images, especially in an uncontrolled data set, we need to fuse different types of features for face recognition. So, we extend KED to multikernel space, where we can fuse different types of features at kernel level.

Suppose there is an ensemble kernel k generated by linearly combining the base kernels

$$k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} \gamma_m k_m(\mathbf{x}, \mathbf{x}') \tag{21}$$

where $\gamma_i \geq 0$. Similar to KDA, the objective function of multi-KDA (MKDA) is

$$\max_{\mathbf{A}, \boldsymbol{\gamma}} \frac{\mathrm{Tr}(\mathbf{A}^T \mathbf{K}\mathbf{W}\mathbf{K}\mathbf{A})}{\mathrm{Tr}(\mathbf{A}^T \mathbf{K}\mathbf{K}\mathbf{A})} \tag{22}$$

where $\mathbf{K}_{i,j} = \sum_{m=1}^{M} \gamma_m k_m(\mathbf{x}_i, \mathbf{x}_j)$, $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_M]^T$, and $\mathbf{W}$ is defined as

$$\mathbf{W}_{i,j} = \begin{cases} 1/n_c, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the } c\text{-th class} \\ 0, & \text{otherwise} \end{cases}$$

where n_c is the number of samples of class c.

The projection matrix $\mathbf{A}$ and the fusing weights $\boldsymbol{\gamma}$ can be alternatively optimized until some predefined iteration precision is achieved [17].

Then, we can replace $k(\mathbf{x}, \mathbf{x}')$ with $\sum_{m=1}^{M} \gamma_m k_m(\mathbf{x}, \mathbf{x}')$ to extend KED to multikernel space. For example, we can rewrite (11) into multikernel space

$$\begin{aligned} \tilde{\mathbf{K}}_{i,j} &= \langle \phi(\tilde{\mathbf{x}}_i), \phi(\tilde{\mathbf{x}}_j) \rangle - \langle \phi(\tilde{\mathbf{x}}_i), \phi(\tilde{\mu}_j) \rangle \\ &\quad - \langle \phi(\tilde{\mu}_i), \phi(\tilde{\mathbf{x}}_j) \rangle + \langle \phi(\tilde{\mu}_i), \phi(\tilde{\mu}_j) \rangle \\ &= \sum_{m=1}^{M} \gamma_m k_m(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) - \sum_{m=1}^{M} \gamma_m k_m(\tilde{\mathbf{x}}_i, \tilde{\mu}_j) \\ &\quad - \sum_{m=1}^{M} \gamma_m k_m(\tilde{\mu}_i, \tilde{\mathbf{x}}_j) - \sum_{m=1}^{M} \gamma_m k_m(\tilde{\mu}_i, \tilde{\mu}_j). \end{aligned} \tag{23}$$

Similar to KED, we can get the p principal components of the occlusion variations in the multikernel space, i.e., multiple KED (MKED), and use the same classification method as (16).

3) Computational Complexity Analysis: In the training stage, constructing the $\tilde{n} \times \tilde{n}$ kernel matrix $\tilde{\mathbf{K}}$ in (11) requires $O(\tilde{n}^2 d)$, where $\tilde{n}$ is the number of samples to train KED and d is the number of features. The eigendecomposition of $\tilde{\mathbf{K}}$ in (10) requires $O(\tilde{n}^3)$ [40]. So, training the p principal components of the intraclass difference samples in the kernel space requires $O(\tilde{n}^2 d) + O(\tilde{n}^3)$.

In addition, KED is based on KDA, and it also needs to compute a $\hat{n} \times \hat{n}$ kernel matrix and an eigen-decomposition problem, which requires $O(\hat{n}^2 d) + O(\hat{n}^3)$, where $\hat{n}$ is the number of samples to train KDA.

Projecting the gallery set $\mathbf{X} \in \mathbb{R}^{d \times n}$ by KDA to get the basic dictionary requires $O(\hat{n} n d)$. Projecting p kernel principal components by KDA to get the KED requires $O(\tilde{n}^2 \hat{n} p)$.

In the testing stage, most of the computational cost lies in solving the sparse representation model (16). There are many methods to solve the l_1-minimization problem, such as orthogonal matching pursuit [41], least angle regression [42], fast iterative soft-thresholding algorithm [43], and augmented Lagrange multiplier (ALM) [44]. Though different methods are of different complexities, they are all based on iterative approximation, and each iteration requires about between $O(n^2)$ and $O(n^3)$ in the worst case [44], where n is the number of atoms of the dictionary. Because the number of the atoms of the dictionary for the proposed KED is much less than that for ESRC, KED is much faster than ESRC in the testing stage, as will be demonstrated in Table II.

IV. Experimental Results

In this section, we perform experiments on several large-scale data sets, such as CAS-PEAL [3], AR [45], FERET [2], and LFW [4], to demonstrate the performance of the proposed methods. In Section IV-A, we present the results on the CAS-PEAL data set. We first compare the proposed methods with some state-of-the-art methods. Then, some related methods are compared with the proposed method based on the same feature. At the end of the subsection IV-A, we give the comparison of classification speed. In Section IV-B, we compare KED with state-of-the-art methods on the AR data set, with both multiple gallery samples per subject and a single gallery sample per subject. In Sections IV-C and IV-D, we give the results on the FERET data set and the LFW data set, respectively.

To evaluate the performance of face recognition methods, many papers [9], [10], [13], [20] consider only two kinds of subsets: a training set and a testing set. There may be multiple samples per subject in the training set, and all subjects in the testing set can be found in the training set. But, for a practical situation, the face recognition system may enroll new subjects, and usually only a single sample per subject is attained. We would not retrain the system at this time. So, face recognition with a single sample per subject is an important issue. To address this problem, Lu *et al.* [46] proposed the discriminative multimanifold analysis by learning discriminative features from image patches. Deng *et al.* [47], [48] proposed the equidistant prototypes embedding for single sample-based face recognition, which maps gallery samples to the equally distant locations, rather than preserving the global or local structure of the training data. Here, we consider three kinds of subsets, i.e., a training set, a gallery set, and a probe set, to evaluate the performance [3]. Note that the subjects in the gallery set may never be seen in the training set. We can use the training set to build a recognition model. Then, the samples in the gallery set and the probe set are transformed by the model and then matched each other.

Before feature extraction and classification, we must align and crop face images first. There are many methods to align images, such as alignment by sparse representation [49], transform-invariant PCA [50], and explicit shape regression [51]. Here, we align face images based on facial landmarks as many papers do. For each image, we first run the Viola–Jones face detector [52] to locate the face. Based on the face location, using the method in [53], we detect nine facial landmark positions. Then, we apply similarity transformation using these landmark points to transform a face to a canonical frame, and crop a center region of size 120 × 100 for further classification.

A. Results on CAS-PEAL Data Set

The CAS-PEAL face database has been constructed by the CAS [3], which provides the worldwide researchers of face recognition community, a large-scale Chinese face database for training and evaluating their algorithms, and facilitates the development by providing large-scale face images with different sources of variations, especially PEAL. The CAS-PEAL face database contains 99 594 images of 1040 individuals (595 males and 445 females). We use the available frontal subset containing 9031 images of the 1040 subjects for experiment. Some example images on the CAS-PEAL data set are shown in Fig. 1 (left).

TABLE I

COMPARISON OF RELATED METHODS BASED ON THE SAME FEATURE ON THE CAS-PEAL DATA SET

Method	Mean acc. ± std. dev.			
	Accessory	Hat	Lighting	Expression
SRC	72.87 ± 0.6	51.17 ± 1.3	17.31 ± 0.7	98.21 ± 0.4
ESRC	87.05 ± 1.0	75.66 ± 2.2	82.08 ± 0.4	99.70 ± 0.1
KDA+SRC	80.81 ± 1.9	65.85 ± 3.2	82.73 ± 0.6	99.69 ± 0.1
KDA+ESRC	80.85 ± 1.9	65.93 ± 3.2	83.03 ± 0.6	99.69 ± 0.1
LED	84.78 ± 0.4	72.26 ± 0.8	66.83 ± 0.9	99.12 ± 0.1
KED	$\mathbf{91.03 \pm 0.6}$	$\mathbf{83.56 \pm 1.0}$	$\mathbf{83.09 \pm 0.5}$	$\mathbf{99.70 \pm 0.1}$

To create the training set, we randomly select some subjects in three kinds of subsets, i.e., 200 subjects from the lighting subset, 100 from the expression subset, and 20 from the accessory subset, with four samples for each subject. Besides the 1280 samples, we also add the normal sample of a subject if any sample of the subject appears in the training set. To form the gallery set, we use all the normal images, i.e., 1040 images of the 1040 subjects, with each subject having only one image. Then, we create six probe sets correspond to the six subsets: expression, lighting, accessory, background, distance, and time. All the images that appear in the training set are excluded from these probe sets. Note that most of the subjects in the gallery set and the probe set are never seen in the training set. To ensure that our results will not depend on any special choice of the training set, we repeat the experiments ten times, and report the mean recognition accuracy and standard deviation.

1) Comparison of Related Methods: To demonstrate the effectiveness of KED, we compare KED with SRC, ESRC, KDA + SRC, and KDA + ESRC. The linear version of KED, which adopts the linear kernel, called LED, is also compared to show the role of a nonlinear discriminant analysis in KED. In LED, the occlusion model is learned with linear kernel, then it is projected by LDA to form the extended dictionary, and use structured sparse representation for classification. For the purpose of fair comparison, all methods are based on the same LBP feature. Because of the high dimension of LBP feature, for SRC and ESRC, we first project the original feature to a PCA subspace with 600 dimensions for the sake of computation efficiency without losing too much information. The dual ALM [44] method is used to solve all the l_1-minimization problems,[1] where the sparsity coefficient $\lambda = 0.001$. For KED,[2] the extended dictionary is learned by the top ten components for each type of accessory variation, i.e., wearing glasses and wearing hat, in the kernel space. We use RBF kernel with χ^2 distance as presented in Section III-A. It should be mentioned that the proposed method uses the same training information as the other methods, including the samples occluded by accessory.

Table I shows the results of each method for four subsets, where subset Hat denotes the samples wearing hat. According to the results, we can draw the following conclusions.

1) All methods get good results for expression variation based on the LBP feature.
2) ESRC performs much better than SRC, which demonstrates the superiority of ESRC by its intraclass variant dictionary.
3) KDA + SRC improves the performance compared with SRC, especially for lighting variation, which proves that KDA can learn a nonlinear discriminant subspace to well suppress the intraclass variation. But, the recognition accuracy of KDA + SRC is lower than ESRC for accessory variation, because KDA is not good enough for occlusion.
4) KDA + ESRC gets the similar results as KDA + SRC, because after KDA transform, most of intraclass variations are suppressed and the relationship of occlusion variations is broken.
5) LED achieves good results for accessory variation, but fails for lighting variation, because the linear transform cannot suppress the dramatic intraclass variations well, and LDA is likely to lose some useful energy for SRC.
6) KED attains the best results for all subsets. KED significantly improves the performance with accessory variation compared with SRC and KDA + SRC. KED is also much better than ESRC with hat variation. It demonstrates that KED can well model the occlusion variation by its extended dictionary in the kernel space, and KED is not the simple combination of KDA and SRC.
7) With lighting variation, KDA + SRC and KED get the similar results, which proves that KED also does well in the nonoccluded samples without overfitting.

TABLE II

COMPARISON OF CLASSIFICATION TIME AND CORRESPONDING ACCURACY FOR RELATED METHODS BASED ON THE SAME FEATURE ON THE CAS-PEAL DATA SET

	SRC	ESRC	**KED**
# Dictionary atoms	1040	1040+1580	1040+20
# Dictionary rows	600	600	290
Classification time (s)	1281	5101	890
Recognition accuracy	72.87	87.05	**91.03**

Table II shows the comparison of classification time and corresponding recognition accuracy for SRC, ESRC, and KED. The classification time is the total time for all the probe images, and the recognition accuracy is for accessory variation. All methods use the same LBP feature for the purpose of fair comparison. For SRC, we use the 1040 gallery samples as the dictionary. For ESRC, we use the 1580 training samples to construct the intraclass variation dictionary by (4). So, there are 2620 atoms in the dictionary of ESRC besides the 1040 gallery samples. For SRC and ESRC, 600 principal components of the data are used in building the dictionaries. The extended dictionary of KED is learned by the top ten principal components for each type of accessory variation, so there are 1060 atoms in the dictionary. There are 291 subjects

[1] MATLAB code: http://www.eecs.berkeley.edu/~yang/software/l1benchmark/

[2] MATLAB code: http://www.mathworks.com/matlabcentral/profile/authors/5133554-ke-kun-huang

TABLE III

COMPARISON OF STATE-OF-THE-ART METHODS ON THE CAS-PEAL DATA SET

Method	Mean recognition accuracy (%)					
	Accessory	Lighting	Expression	Time	Background	Distance
1NN	38.75	2.874	80.31	38.79	39.69	76.91
SSEC	66.64	17.39	74.51	51.94	66.83	84.23
RRC	84.19	29.32	93.98	96.67	95.64	97.90
SLF-RKR	90.88	28.81	99.64	98.48	99.88	99.69
MOST	80.35	82.39	98.15	97.88	99.01	99.75
KED	**91.03**	**83.09**	**99.70**	**99.70**	**99.88**	**99.88**

in the training set, so the number of dictionary rows of KED is 290 after KDA projection.

According to the results, we can draw the following conclusions.

1) Though ESRC significantly improves the performance compared with SRC, the classification procedure is much slower than SRC for the large extended dictionary. As the number of dictionary atoms increases, the classification time of SRC grows fast.
2) Though KED is a kernel method, which needs to compute the kernel function value between the probe sample and each training sample, the classification is still faster than SRC, in particular, much faster than ESRC, because KED appends only a small number of atoms to the basic dictionary and is based on low feature dimension. Most important of all, the recognition accuracy of KED is the highest.

2) Comparison of State-of-the-Art Methods: We compare the proposed method with some state-of-the-art methods, such as SSEC [11], RRC [12], SLF-RKR [13], and MOST [20]. To give a baseline for comparison, we also present the results of the nearest neighbor classifier. For SSEC, we set $\lambda_E = 2$, $\lambda_V = 0$, $\kappa = 0.3$, and $T = 5$ as proposed in [11]. For RRC, we set $\mu = (\varsigma/\delta)$, $\varsigma = 8$, and $\tau = 0.8$ as proposed in [12]. For SLF-RKR, we set $S = 0$, $P_0 = 5$, and $Q_0 = 4$ as proposed in [13]. Because SSEC, RRC, and SLF-RKR do not require any training process, we only use the gallery samples as the dictionary for recognition. Because most of the subjects in the gallery set and the probe set are never seen in the training set, and the training set contains the same type of occlusion as some probe samples, it will decline the performance if we add the training samples to the dictionary. For instance, RRC only gets an accuracy of 21.27% for accessory, when adding the training samples to the dictionary. For MOST and KED, we use the training set to train the models, and only use the gallery set for recognition. All the methods use the same images of size 120 × 100. The MATLAB codes of RRC and SLF-RKR are available.[3]

Table III lists mean recognition accuracies for state-of-the-art methods on the CAS-PEAL data set with six types of intraclass variations. According to the results, we can draw the following conclusions.

1) The accessory and lighting variations are more challenging than other variations on the data set.
2) Though SSEC improves the recognition accuracy for occlusion variation, it declines the performances for the other subsets, because SSEC overfits the occlusion, i.e., it may regard a nonoccluded sample as an occluded sample.
3) RRC and SLF-RKR achieve good results for occlusion, but they fail for lighting variation, because they do not use any training process to suppress the dramatic intraclass variation.
4) As a KDA-based method, MOST gets impressive results for six subsets, but is still not good enough for occlusion.
5) The proposed KED achieves the best results for all the six subsets. Especially, it simultaneously attains the best recognition accuracies for occlusion and lighting variations. KED increases an average of 10.7% for occlusion compared with MOST, and raises an average of 54.3% for lighting variation compared with SLF-RKR. It demonstrates that not only does KED extract the nonlinear discriminant feature to get impressive results for nonoccluded samples, but it also exploits the sparsity to handle the occlusion problem without overfitting.

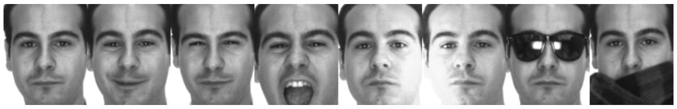


Fig. 3. Example images of one subject on the AR data set.

B. Results on AR Data Set

The AR face database, created by Martinez and Benavente [45], consists of over 3315 facial images from 136 subjects (76 men and 60 women). These images suffer from different facial variations, including various facial expressions, illumination variations, and occlusion by sunglasses or scarf. Fig. 3 shows the face images of the first subject in the AR database. There are 119 subjects who completely participate in two sessions, separated by two weeks, with 13 images in each session for each subject. There are less than 26 images for each of the remaining 17 subjects.

To create a training set, we use the 833 images without occlusion from the 119 subjects in Section I and all the images from the remaining 17 subjects for training. Two types of gallery set are considered here. The first gallery set is the same as the training set where there are multiple samples per subject, while the second gallery set is the collection of the neutral sample of each subject in session 1 where there is only a single sample per subject. All the images from the 119 subjects in session 2 are used as probe set, which is divided into four subsets: sunglasses, scarf, lighting, and expression. Note that the training set and the first gallery set contain the occluded images from the 17 subjects who are not in the probe set. For the proposed method, the KED is learned from the two types of occlusion variation from the 17 subjects, i.e., wearing sunglasses and wearing scarf.

Table IV shows the recognition accuracies and classification time for state-of-the-art methods on the AR data set with

[3] MATLAB code: http://www4.comp.polyu.edu.hk/~cslzhang/code.htm

TABLE IV
RECOGNITION ACCURACIES AND TESTING TIME FOR STATE-OF-THE-ART METHODS ON AR DATA SET WITH MULTIPLE GALLERY SAMPLES PER SUBJECT

Method	Recognition accuracy (%)				Classification time (s)
	Sunglasses	Scarf	Expression	Lighting	
SRC [9]	43.70	48.18	87.82	91.60	102
ESRC [14]	59.38	58.82	91.18	93.00	348
SSEC [11]	45.30	50.22	70.84	83.59	2165
RRC [12]	64.71	57.98	92.65	95.80	845
SLF-RKR [13]	79.55	93.56	97.48	98.32	6110
MOST [20]	73.11	90.20	97.69	97.48	**20**
LBP+SRC	89.92	80.95	97.69	98.04	95
LBP+ESRC	93.00	87.11	97.69	98.04	326
KED	**95.52**	**94.96**	**99.79**	**99.72**	41

TABLE V
RECOGNITION ACCURACIES FOR STATE-OF-THE-ART METHODS ON AR DATA SET WITH A SINGLE GALLERY SAMPLE PER SUBJECT

Method	Recognition accuracy (%)				Classification time (s)
	Sunglasses	Scarf	Expression	Lighting	
SRC [9]	22.13	19.05	64.29	52.66	34
ESRC [14]	69.19	67.51	88.87	92.72	117
SSEC [11]	24.93	28.85	61.18	49.89	234
RRC [12]	50.42	35.29	74.37	67.51	255
SLF-RKR [13]	19.05	23.81	34.45	28.29	1378
MOST [20]	84.87	91.60	97.69	97.48	**8**
LBP+SRC	78.15	65.83	87.39	84.31	32
LBP+ESRC	94.96	91.60	97.69	98.04	105
KED	**95.24**	**93.84**	**99.79**	**99.44**	11

multiple gallery samples per subject. The parameters of the methods are the same as those of the previous subsection IV-A. For SRC and ESRC, we also use the same LBP feature as the proposed method for fair comparison, which are denoted by LBP + SRC and LBP + ESRC, respectively. According to the results, we can find that the following holds.

1) The performances of LBP-based or Gabor-based methods, such as SLF-RKR, MOST, LBP + SRC, LBP + ESRC, and KED, are much better than pixel-based methods, such as SRC, ESRC, SSEC, and RRC.
2) MOST and SLF-RKR get good performance for the samples wearing scarf, but are not good for the samples wearing sunglasses.
3) KED attains the best performances for all subsets. It demonstrates that KED not only does well in nonoccluded samples, but also handles the occlusion well.
4) MOST is not a sparse-based method, so the classification is the fastest. KED is only a little slower than MOST and much faster than ESRC, because it appends only a small number of atoms constructed by the proposed occlusion model to the basic dictionary, rather than the difference images as ESRC.

Table V shows the recognition accuracies and classificatioin time for state-of-the-art methods on the AR data set with a single gallery sample per subject. According to the results, we can find that the following holds.

1) Compared with Table IV, the performance of KED is almost not changed, while the classification is much faster, which demonstrates the robustness of KED.

TABLE VI
RECOGNITION ACCURACIES FOR STATE-OF-THE-ART METHODS ON THE FERET DATA SET

Method	Recognition accuracy (%)			
	fb	fc	dup1	dup2
SRC [9]	80.59	84.54	48.20	35.90
ESRC [14]	90.46	92.27	59.70	46.15
SSEC [11]	71.80	40.62	47.56	31.97
RRC [12]	89.04	89.69	69.94	63.68
SLF-RKR [13]	98.41	82.99	76.18	72.22
MOST [20]	96.49	98.45	70.64	64.96
LBP+SRC	98.66	92.27	74.93	61.11
LBP+ESRC	99.33	93.30	81.02	64.10
KED	**99.58**	**100.0**	**91.55**	**86.32**

2) ESRC and MOST even get better performance by a single gallery sample per subject, especially with sunglasses occlusion, because there is not any gallery sample with occlusion that may introduce mistake.
3) SRC, RRC, SSEC, and SLF-RKR attain worse performance for occlusion, because the sparse representation-based methods cannot eliminate occlusion without occluded gallery samples.

C. Results on FERET Data Set

The FERET database [2] is assembled to support the government monitored testing and evaluation of face recognition algorithms using standardized tests and procedures. The database consists of 14 051 images of human heads with views ranging from frontal to left and right profiles. We use the standard procedure for experiment. The gallery set consists of 1196 images from fa subset, with only single image per subject. The probe subsets fb and fc are captured with expression and illumination variations (the images in fc are captured by a different camera). Especially, the probe subsets dup1 and dup2 consist of images that are taken at different times. For some people, more than two years elapsed between the gallery set and dup1 or dup2 set. Fig. 4 shows some example images of one subject on the FERET data set. We use the gallery set and the remaining front images outside the probe sets for training. KED is learned from the training samples who wearing glasses in the CAS-PEAL database.

Table VI shows the recognition accuracies for state-of-the-art methods on the FERET data set. We can find that the performances of SLF-RKR, MOST, LBP + SRC, LBP + ESRC, and KED are excellent for fb subset. MOST also gets good performance for fc subset, but fails for dup1 and dup2 subsets. On the whole, KED attains the best performances for all subsets. For the dup1 subset, KED increases the accuracy by 43.3% than SRC, 31.8% than ESRC, 44% than SSEC, 21.6% than RRC, 15.4% than SLF-RKR, 20.9% than MOST, 16.6% than LBP + SRC, and 10.5% than LBP + ESRC. For the dup2 subset, the improvement of KED is much more significant.

D. Results on LFW Data Set

The LFW data set [4] contains 13 233 images of 5749 people downloaded from the Web, which is

Fig. 4. Example images of one subject on the FERET data set.

Fig. 5. Some cropped examples of one person in the LFW data set.

designed for unconstrained face recognition with dramatic variations of pose, illumination, expression, misalignment, occlusion, and so on. Some cropped examples of one person are shown in Fig. 5.

Since the LFW data set was established, many face verification methods have been proposed on the uncontrolled data set. Face verification is to determine whether a pair of face images is from the same person or not, which is a binary-class problem. But in traditional face identification, we must identify which person a face image belongs to, which is a multiclass problem. Face identification and face verification are two subproblems in face recognition [34]. Though some methods can achieve impressive results in a subproblem, they may get poor performance in another subproblem. So, we perform two kinds of experiments on the LFW data set, i.e., face identification and face verification.

1) Face Identification on LFW Data Set: As [13], a subset of LFW is used in the identification experiments, which consists of 5425 images of 311 subjects with no less than six samples per subject. We randomly select five samples per subject for training, and the rest for testing. Once we select a random training sample set, they are used for all methods. We repeat the experiments ten times.

The parameters of SRC, ESRC, SLF-RKR, and MOST are the same as previous. Because JB [34] classifier and high-dimensional LBP (HDLBP) [38] feature have demonstrated their efficiency for uncontrolled face recognition [54], we also compare HDLBP + JB, HDLBP + SRC, and HDLBP + ESRC. For the proposed method, to capture rich enough information available in the uncontrolled face images, similar to [38], we use the multiscale LBP features sampling around some landmarks. Besides LBP feature, we also use Gabor feature sampling around some landmarks with RBF kernel. Then, we use MKED to fuse the two types of features for classification. To learn KED and MKED, we use some of the images of the subjects with no less than two samples per subject but less than six samples per subject in LFW, which are occluded by sunglasses, hat, or some other objects.

Fig. 6 shows ten random face identification results for state-of-the-art methods and the proposed methods on the LFW data set. According to the results, we can find that the following holds.

1) SRC and ESRC get poor performance on the uncontrolled database, because they are based on original pixel feature.

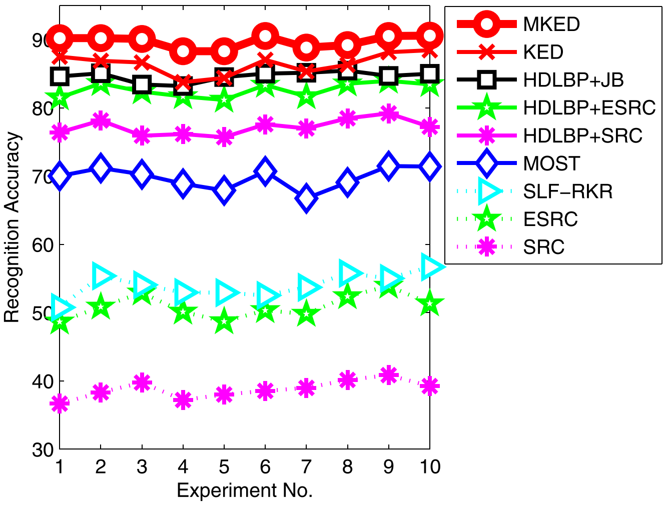


Fig. 6. Face identification results for state-of-the-art methods and the proposed methods on the LFW data set.

2) Though SLF-RKR is an LBP-based method, the performance is not good enough on the uncontrolled database, because its LBP feature is extracted at fixed grids, which may fail on the misaligned condition for the dramatic expression and pose variations on the LFW database.
3) Though the Gabor feature of MOST is also extracted at fixed grids, it gets better performance than SLF-RKR on the database, which shows that the KDA-based method can well suppress the dramatic intraclass variations.
4) The methods based on the HDLBP feature significantly improve the performance compared with the methods based on the fixed-grid feature, which demonstrates that the landmark-based feature can handle the pose or expression variation better than the grid-based feature. The performance of HDLBP + JB is better than HDLBP + SRC and HDLBP + ESRC, which shows the power of JB for uncontrolled face recognition.
5) As a KDA-based method, KED achieves better performance than the above methods by landmark-based feature. Combining landmark-based Gabor feature and LBP feature by MKED, it achieves the best performance. MKED increases an average of 3.2%, 5.1%, 7.1%, 12.5%, 19.9%, and 35.7% compared with KED, HDLBP + JB, HDLBP + ESRC, HDLBP + SRC, MOST, and SLF-RKR, respectively. Note that HDLBP + JB achieves good results for face verification, but our method significantly outperforms HDLBP + JB for face identification. It clearly shows the effectiveness of MKED in dealing with uncontrolled face identification.

2) Face Verification on LFW Data Set: LFW specifies a number of evaluation protocols for face verification. We follow the unrestricted protocol and use only the training data provided by LFW. The data set is divided into ten disjoint splits, which contain different identities. The performance is measured by performing a tenfold cross validation, training the model on nine splits, and testing it on the remaining split. In the unrestricted protocol, the training information is provided as simply the names of the people in each split. The test-

TABLE VII

MEAN VERIFICATION ACCURACIES AND STANDARD DEVIATIONS FOR STATE-OF-THE-ART METHODS ON LFW DATA SET UNDER UNRESTRICTED PROTOCOL

	Method	Mean acc. ± std. dev.
without outside training data	Combined multishot [55]	89.50 ± 0.51
	LBP multishot [55]	85.17 ± 0.61
	LBP PLDA [56]	87.33 ± 0.55
	Joint Bayesian [34]	90.90 ± 1.48
	HDLBP+JB [38]	93.18 ± 1.07
	Fisher vector [57]	93.03 ± 1.05
	ConvNet-RBM [58]	91.75 ± 0.48
	SLBP [59]	90.00 ± 1.33
	VisionLabs [60]	92.90 ± 0.31
	Aurora [61]	93.24 ± 0.44
	HPEN + HDLBP + JB [54]	94.87 ± 0.38
	MKED	93.60 ± 1.21
with outside training data	ConvNet-RBM [58]	92.52 ± 0.38
	DeepFace [25]	97.35 ± 0.25
	DeepID [27]	97.45 ± 0.26
	DeepID2 [28]	99.15 ± 0.13
	DeepID3 [29]	99.53 ± 0.10

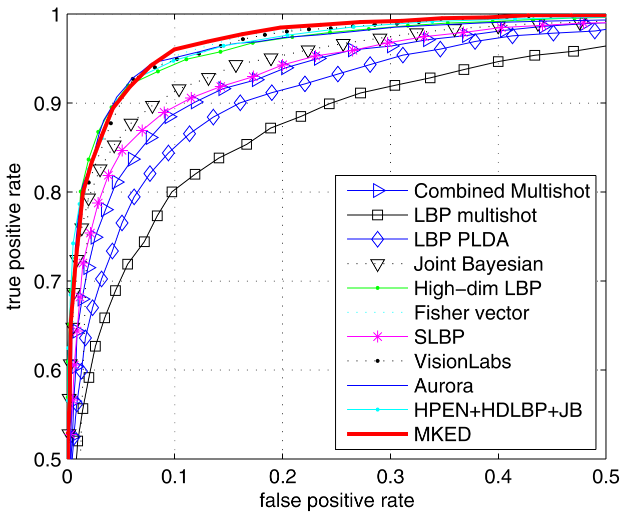


Fig. 7. ROC curves averaged over tenfold cross validation under LFW unrestricted protocol without outside training data.

ing set is the 600 predefined image pairs in the remaining split.

For the proposed method, we use MKED to fuse the landmark-based LBP and Gabor feature, and select the top 300 persons, which have the most number of samples to train the MKDA projection and the fusing weights. To apply MKED for face verification, judging whether a given pair of samples, i.e., I and J, belong to the same subject or not, as [62], we use a set of background samples together with I and its horizontally mirrored sample as the basic dictionary, and the other sample J as the probe. The verification score is defined as $1 - r_c$, where r_c is the residual defined in (17). In order to get better performance for face verification, we also combine the Mahalanobis distance in the kernel space learned by information theory [63].

We compare against the best results listed in the LFW official Web site, including the methods without outside training data, such as combined multishot [55], LBP multishot [55], LBP Probabilistic Linear Discriminant Analysis [56], JB [34], HDLBP + JB [38], Fisher vector [57], ConvNet-Restricted Boltzmann Machine [58], Soft Local Binary Patterns [59], VisionLabs [60], Aurora [61], and High-fidelity Pose and Expression Normalization + HDLBP + JB [54], and the deep learning-based methods with outside training data, such as ConvNet-RBM [58], DeepFace [25], DeepID [27], DeepID2 [28], and DeepID3 [29].

Table VII lists mean verification accuracies and standard deviations for state-of-the-art methods on the LFW data set under unrestricted protocol. Our method achieves 93.60% face verification accuracy, which is only lower than HPEN + HDLBP + JB without outside training data. Though the performances of the deep learning-based methods are better than the proposed method, they need large amounts of outside training data, which are collected in the same situation as the LFW data set.

Fig. 7 shows the Receiver Operating Characteristic curves averaged over tenfold cross validation for the methods without outside training data. For more explicit, we only show the figure where false positive rate is from 0 to 0.5, and true positive rate is from 0.5 to 1. It can be found that the performance of our method is as good as the best results, which demonstrates that our method can not only get good identification performance but also do well in verification task.

V. CONCLUSION

The experiments suggest a number of conclusions as follows.

1) Compared with sparse representation-based classifiers, such as SRC and ESRC, not only does KED achieve better performance for occlusion, but it also gets much higher recognition accuracies for dramatic intraclass variations, such as LFW data set and dup2 in FERET data set. Furthermore, KED is much faster than SRC and ESRC because only a small number of atoms are appended to the basic dictionary and the feature dimension is low.
2) Compared with kernel-based methods, such as MOST, KED yields much better recognition accuracies for occlusion, and also improves the performance for nonoccluded samples.
3) KED significantly improves the performance for occlusion compared with KDA + SRC. KED is not the simple combination of KDA and SRC. The key of KED is how to construct and compute the extended dictionary, so that it can represent the possible occlusion variations efficiently in the kernel space.
4) MKED can fuse different types of features at kernel level to capture rich enough information available in the face images. On the LFW data set, MKED achieves much higher identification accuracy than some state-of-the-art methods, and gets competitive results for verification task.

References

[1] 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.

[2] 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.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[3] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[4] 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," Janelia Farm Res. Campus, HHMI, Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[6] 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.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[8] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[10] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[11] X.-X. Li, D.-Q. Dai, X.-F. Zhang, and C.-X. Ren, "Structured sparse error coding for face recognition with occlusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1889–1900, May 2013.

[12] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[13] M. Yang, L. Zhang, S. C.-K. Shiu, and D. Zhang, "Robust kernel representation with statistical local features for face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 900–912, Jun. 2013.

[14] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[15] W. Deng, J. Hu, and J. Guo, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 399–406.

[16] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, 2000.

[17] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.

[18] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.

[19] C. H. Chan, M. A. Tahir, J. Kittler, and M. Pietikäinen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1164–1177, May 2013.

[20] C.-X. Ren, D.-Q. Dai, X.-X. Li, and Z.-R. Lai, "Band-reweighed Gabor kernel embedding for face image representation and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 725–740, Feb. 2014.

[21] 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.

[22] M. Yang, Z. Feng, S. C. K. Shiu, and L. Zhang, "Fast and robust face recognition via coding residual map learning based adaptive masking," *Pattern Recognit.*, vol. 47, no. 2, pp. 535–543, 2014.

[23] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, and Z. Zhu, "Robust face recognition via occlusion dictionary learning," *Pattern Recognit.*, vol. 47, no. 4, pp. 1559–1572, Apr. 2014.

[24] H.-D. Liu, M. Yang, Y. Gao, Y. Yin, and L. Chen, "Bilinear discriminative dictionary learning for face recognition," *Pattern Recognit.*, vol. 47, no. 5, pp. 1835–1845, 2014.

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1701–1708.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (2014). "Learning face representation from scratch." [Online]. Available: http://arxiv.org/abs/1411.7923

[27] Y. Sun, X. G. Wang, and X. O. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1891–1898.

[28] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[29] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). "DeepID3: Face recognition with very deep neural networks." [Online]. Available: http://arxiv.org/abs/1502.00873

[30] W. Ouyang *et al.*, "DeepID-Net: Deformable deep convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2403–2412.

[31] G. Hu *et al.* (2015). "When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition." [Online]. Available: http://arxiv.org/abs/1504.02351

[32] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. (2015). "Targeting ultimate accuracy: Face recognition via deep embedding." [Online]. Available: http://arxiv.org/abs/1506.07310

[33] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3743–3752.

[34] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 566–579.

[35] F. Schroff, D. Kalenichenko, and J. Philbin. (2015). "FaceNet: A unified embedding for face recognition and clustering." [Online]. Available: http://arxiv.org/abs/1503.03832

[36] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599–2606, Nov. 2009.

[37] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[38] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3025–3032.

[39] 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.

[40] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *VLDB J.*, vol. 20, no. 1, pp. 21–33, 2011.

[41] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Construct. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.

[42] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[43] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[44] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, "Fast ℓ_1-minimization algorithms for robust face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3234–3246, Aug. 2013.

[45] A. Martínez and R. Benavente, "The AR face database," Comput. Vis. Center, Barcelona, Spain, Tech. Rep. #24, Jun. 1998.

[46] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, Jan. 2013.

[47] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng, "Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach," *Pattern Recognit.*, vol. 43, no. 5, pp. 1748–1762, 2010.

[48] W. Deng, J. Hu, X. Zhou, and J. Guo, "Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning," *Pattern Recognit.*, vol. 47, no. 12, pp. 3738–3749, 2014.

[49] 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.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[50] W. Deng, J. Hu, J. Lu, and J. Guo, "Transform-invariant PCA: A unified approach to fully automatic facealignment, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1275–1284, Jun. 2014.

[51] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.

[52] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2001.

[53] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Image Vis. Comput.*, vol. 27, no. 5, pp. 545–559, 2009.

[54] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.

[55] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1978–1990, Oct. 2011.

[56] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144–157, Jan. 2012.

[57] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2013, pp. 8.1–8.12.

[58] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1489–1496.

[59] C. Huang, S. Zhu, and K. Yu, "Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval," NEC Lab. Amer., Princeton, NJ, USA, Tech. Rep. 2011-TR115, Oct. 2011.

[60] VisionLabs. *VisionLabs Version 1.0*. [Online]. Available: http://www.visionlabs.ru/face-recognition, accessed Feb. 5, 2016.

[61] T. Heseltine, P. Szeptycki, J. Gomes, M. C. Ruiz, and P. Li, "Evaluation of algorithm 'Aurora-c-2014-1' on labeled faces in the wild," Aurora Comput. Services Ltd., Higham Ferrers, U.K., Tech. Rep., Jan. 2014.

[62] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

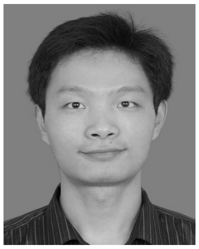
[63] J. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 209–216.

Ke-Kun Huang (M'14) received the B.Sc. and M.Sc. degrees in applied mathematics from Sun Yat-sen University, Guangzhou, China, in 2002 and 2005, respectively, where he is currently pursuing the Ph.D. degree.

He is an Associate Professor with the Department of Mathematics, Jiaying University, Meizhou, China. His current research interests include image processing and face recognition.

Dao-Qing Dai (M'09) received the B.Sc. degree from Hunan Normal University, Changsha, China, in 1983, the M.Sc. degree from Sun Yat-sen University, Guangzhou, China, in 1986, and the Ph.D. degree from Wuhan University, Wuhan, China, in 1990, all in mathematics.

He was an Alexander von Humboldt Research Fellow with Free University, Berlin, Germany, from 1998 to 1999. He is currently a Professor with the Faculty of Mathematics and Computing, Sun Yat-sen University. He has authored or co-authored over 100 refereed technical papers. His current research interests include image processing, wavelet analysis, face recognition, and bioinformatics.



Chuan-Xian Ren (M'14) received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2010.

He was with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, as a Senior Research Associate from 2010 to 2011. He is currently an Associate Professor with the Faculty of Mathematics and Computational Science, Sun Yat-sen University. His current research interests include image processing, face recognition, and machine learning.

Dr. Ren is an Invited Reviewer of American Mathematical Reviews. He was elected a candidate for the Thousand-Hundred-Ten Talents Program of Guangdong Province in 2014.

Zhao-Rong Lai (S'15) received the B.Sc. degree in mathematics, the M.Sc. degree in computational science, and the Ph.D. degree in statistics from Sun Yat-sen University, Guangzhou, China, in 2010, 2012, and 2015, respectively.

His current research interests include face recognition, statistical machine learning, and statistical optimization.