IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 30, 2019, accepted February 16, 2019, date of publication February 25, 2019, date of current version March 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901376

Dictionary Representation of Deep Features for Occlusion-Robust Face Recognition

FENG CEN[1] AND GUANGHUI WANG[2], (Senior Member, IEEE)
[1]Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[2]Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS 66045, USA

Corresponding author: Feng Cen (feng.cen@tongji.edu.cn)

This work was supported in part by the Shanghai Agriculture Applied Technology Development Program, China, under Grant G20180306, and in part by the China Scholarship Council.

ABSTRACT Deep learning has achieved exciting results in face recognition; however, the accuracy is still unsatisfying for occluded faces. To improve the robustness for occluded faces, this paper proposes a novel deep dictionary representation-based classification scheme, where a convolutional neural network is employed as the feature extractor and followed by a dictionary to linearly code the extracted deep features. The dictionary is composed by a gallery part consisting of the deep features of the training samples and an auxiliary part consisting of the mapping vectors acquired from the subjects either inside or outside the training set and associated with the occlusion patterns of the testing face samples. A squared Euclidean norm is used to regularize the coding coefficients. The proposed scheme is computationally efficient and is robust to large contiguous occlusion. In addition, the proposed scheme is generic for both the occluded and non-occluded face images and works with a single training sample per subject. The extensive experimental evaluations demonstrate the superior performance of the proposed approach over other state-of-the-art algorithms.

INDEX TERMS Face recognition, convolutional neural network, occlusion-robust, deep learning, dictionary representation.

I. INTRODUCTION

Face recognition (FR) is an important and challenging research topic in computer vision and pattern recognition. In recent years, FR has achieved great progress, benefiting from the advancement of convolutional neural networks (CNNs) based methods. Although some exciting results that could approach human vision performance have been reported on challenging face benchmarks [1]–[3], such as Labeled Faces in the Wild (LFW) [4] and YouTube Faces Database (YFD) [5], there is still a long way towards achieving robust FR under challenging situations like partial occlusion and image corruption.

In the CNN-based FR framework, the CNN is generally used as a feature extractor, followed by a classifier [1]–[3]. The state-of-the-art CNNs [6]–[9], usually involving over tens of millions of parameters, require a vast amount of data to mitigate overfitting, when trained from scratch (with random initialization). To collect sufficient task specified training face images is, however, difficult and expensive. Thus, a popular and feasible choice in practice is directly using a network pre-trained on large scale general purpose face datasets (e.g., VGGFace2 [10], CASIA-WebFace [11], and MegaFace [12]) as the off-the-shelf feature extractor for the task of interest or as an initialization to fine-tune the network on the face dataset specified to the application task. Regarding the classifier, a simple classifier, such as softmax, is usually selected due to the powerful capability of CNNs to project the face images into a high-dimensional feature space that is linearly separable.

The CNN-based FR schemes with the CNNs pre-trained on large-scale face datasets are, usually, less sensitive to facial deformation, such as illumination change, expression, and head pose variation. However, if the face is partially occluded, identification of the subject with the simple classifier is error prone. For instance, as shown in Fig.1 (b), for the sunglasses-occluded faces of different subjects in the AR database, the deep feature vectors lie in the subspace Ω_{oj}, as distinct from the subspaces associated with the non-occluded faces. A simple classifier is difficult

The associate editor coordinating the review of this manuscript and approving it for publication was Junchi Yan.

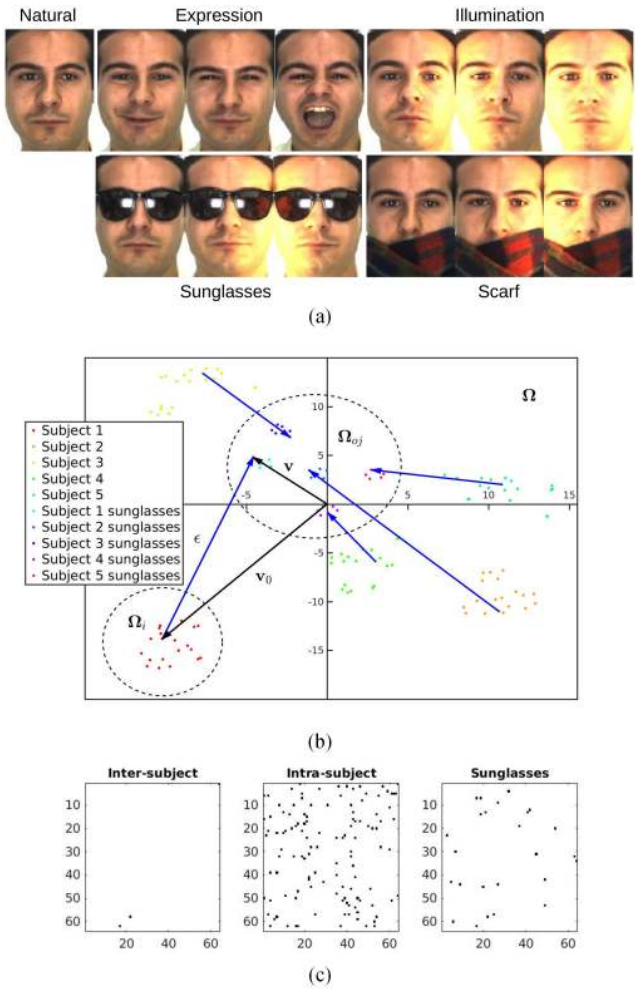

FIGURE 1. Visualization of the deep feature vectors of the face images of 5 subjects randomly selected from the AR database [15]. (a) Example face images of a subject. (b) 2-D visualization of the deep feature vectors. The t-SNE [16] is applied to convert the high dimensional deep features into 2-D vectors for visualization convenience. (c) Visualization of the deep feature components activated commonly for the natural faces associated with different subjects (inter-subject), the non-occluded faces associated with the same subject (intra-subject) and the sunglasses-occluded faces associated with different subjects (sunglasses). The 4096-D feature vector is reshaped to a matrix of 64 × 64 for the convenience of visualization. The black dots represent the components activated commonly for different faces.

to identify the subject for the deep feature vectors lying in Ω_{oj}.

In fact, the subspace Ω_{oj} is associated with the class of sunglasses. Many research works have shown that the features extracted by the CNNs are generic for various visual classification tasks outside the training domain [13], [14]. This means that the CNN trained over the face images can also effectively projects the images involving sunglasses with sufficiently large size into adjacent locations to form a subspace in the deep feature space. By examining the images of the faces wearing sunglasses in Fig.1(a), we can observe that the sunglasses occupy pretty large area of the face images. From Fig.1(c), we can observe that the deep feature vectors lying in different subspaces (the inter-subject case) and the same subspace (the intra-subject case) have very few and many common activations, respectively, and for the sunglasses case, the number of the common activations is in-between. Even less than that in the intra-subject case, the quantity of the common activations in the sunglasses case is sufficient to support a subspace with loose boundaries.

On the other hand, within the subspace Ω_{oj}, the faces belonging to the same subject are very close to each other, as shown in in Fig.1 (b), since the deep feature vectors not only describe the features of the sunglasses but also reflect the features of the faces. This phenomenon indicates that it is possible to employ an advanced classifier rather than a simple classifier to identify the subject correctly for the occluded face images.

Let $\mathbf{v}$ and $\mathbf{v_0}$ be the deep feature vectors of the occluded face image $\mathbf{y}$ and non-occluded face image $\mathbf{y_0}$ associated with the same subject, respectively. As shown in Fig.1 (b), the relationship between $\mathbf{v}$ and $\mathbf{v_0}$ can be modeled by

$$\mathbf{v} = \mathbf{v_0} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}$ is the mapping vector from $\mathbf{v_0}$ to $\mathbf{v}$ (or the difference vector between $\mathbf{v}$ and $\mathbf{v_0}$). If $\mathbf{v_0}$ is recovered from $\mathbf{v}$,the FR can be easily achieved.

Recovering $\mathbf{v_0}$ from $\mathbf{v}$ is quite similar to the intensively studied problem arising from traditional FR for corrupted face images in image space, i.e., recovering $\mathbf{y_0}$ from $\mathbf{y}$. In the image space, the relationship between $\mathbf{y}$ and $\mathbf{y_0}$ can be modeled by

$$\mathbf{y} = \mathbf{y_0} + \mathbf{e}, \tag{2}$$

where $\mathbf{e}$ is the corruption or occlusion error. The most prominent traditional framework to deal with this problem is the sparse representation classification (SRC) proposed by Wright *et al.* [17]. To effectively recover $\mathbf{y_0}$, the SRC-based schemes require that the error $\mathbf{e}$ can be shared across (or highly correlated between) different face images and subjects. Apparently, this requirement is met in the image space (*e.g.*the image variations caused by the same sunglasses for different subjects are highly correlated). Furthermore, in the linear feature space, the linear transformation of the image space, such as Eigenfaces, Fisherfaces and Gabor feature space, the conformance to this requirement is retained since the linear transformation can be represented as a multiplication of a matrix $\mathbf{R}$.[1] Hence, the SRC and its improvements, such as Extended SRC (ESRC) [18] and Superposed SRC (SSRC) [19], can be easily applied to the linear feature space. In contrast, the CNN-based feature extraction is, however, a nonlinear projection from the image space to the deep feature space, since it involves nonlinear operations, such as ReLU and pooling. As a result, the deep feature is a kind of holistic feature and the mapping vector $\boldsymbol{\epsilon}$ in Eq.(1) is not a projection of $\mathbf{e}$ to the deep feature space, i.e., $\boldsymbol{\epsilon} \neq f(\mathbf{e})$, where $f(\bullet)$ denotes the nonlinear projection of the CNN, but related

[1] Applying $\mathbf{R}$ to both sides of (2) yields $\mathbf{Ry} = \mathbf{Ry_0} + \mathbf{Re}$. Therefore, if $\mathbf{e}$ is shared across different face images, then $\mathbf{Re}$ will also be shared.

to both the physical occlusion error $\mathbf{e}$ and the face image $\mathbf{y_0}$. Therefore, the extension of the SRC and its improvements to the deep feature space is not so apparent.

In this paper, we observe that for occluded face images, the mapping vector $\boldsymbol{\epsilon}$ can be approximated by a linear combination of the mapping vectors of other face images associated with the same occlusion pattern. Motivated by this observation, we propose a deep dictionary representation based classification (DDRC) method to perform FR through the recovery of $\mathbf{v_0}$. Let Ω_i be the subspace associated with the non-occluded faces of the ith subject and Ω_{oj} the subspace associated with the faces occluded with the jth occlusion pattern. Suppose that the faces are occluded with the jth occlusion pattern. We observed that in the deep feature space, the mapping function from the subspace Ω_i to the subspace Ω_{oj} can be approximated by a linear combination of the mapping vectors associated with the jth occlusion pattern, i.e., the vectors drawn with blue arrows in Fig. 1 (b). Based on this observation, we propose to use a dictionary consisting of a gallery part and an auxiliary part to code the deep feature vector of the testing sample. The deep feature vectors of the training samples are used as the atoms of the gallery part. The auxiliary part is constructed by concatenating the mapping vectors associated with the occlusion pattern presented in the testing samples. With this dictionary, the deep feature vector of the occluded testing sample is coded by a squared l_2-norm minimization to recover the deep feature vector from occlusion. The classification of the proposed DDRC is achieved by searching for the subject having the closest deep feature subspace to the recovered deep feature vector of the occluded testing sample.

The research is originally motivated by cloud-based face applications. With the development of deep neural network accelerator chips [20], the CNN-based feature extraction will gradually shift to the terminal or mobile end. To reduce the bandwidth of transmission, only the deep features, rather than the images, of the faces will be transmitted to the cloud for further processing, such as classification and verification. The transmission of deep features requires less bandwidth and is much safer in terms of protection of private information than using images. In such situation, however, we have to cope with the occlusion in the deep feature space at the cloud end. Therefore, instead of removing occlusion in image space, we focus on alleviating the negative impact of occlusion on FR in the deep feature space.

Our main contributions are summarized as follows.

1) We observe that for occluded face images, the mapping vector $\boldsymbol{\epsilon}$ can be approximated by a linear combination of the mapping vectors of other face images associated with the same occlusion pattern.
2) We propose to apply the dictionary representation coding with squared l_2-norm minimization in the deep feature space for FR. To alleviate the negative effect of occlusion in the deep feature space, based on the above observation, we propose a novel auxiliary dictionary, which is generated with the mapping vectors associated with the same occlusion pattern as the testing face images.
3) We present extensive experiments on publicly available databases and show that the proposed DDRC is a real-time occlusion-robust scheme and can simultaneously handle multiple complex FR challenges even with a single training sample per subject. We also demonstrate the significant improvement of the proposed DDRC over the state-of-the-art approaches.
4) Although end-to-end learning is desired in many applications, the proposed approach, by making optimal use of the available CNN and the classical learning approaches, provides a new perspective to other similar problems.

The rest of the paper is organized as follows. Section II briefly reviews the related work. Section III describes the proposed DDRC. Section IV shows experiments, and the paper is concluded in Section V.

II. RELATED WORK

The defining characteristic of these approaches is the use of the CNNs as a feature extractor. A representative system is DeepFace, where Taigman *et al.* [1] derived a face representation from a nine-layer deep neural network by training it on the largest facial dataset and reached a FR accuracy closely approaching human-level performance on the LFW dataset. The DeepFace work was extended by the DeepID series of papers by Sun *et al.* [2], [21]–[23]. Multiple CNNs were introduced in [2]; multi-task learning over classification and verification was proposed in [21]; different CNNs architectures which branch a fully connected layer after each convolution layer were presented in [22] and very deep networks are used in [23]. These series of works steadily improved the performance on LFW dataset. Researchers from Google [3] employed a massive dataset and a triplet-based loss to train the CNNs. Wen *et al.* [24] proposed a latent factor guided CNNs to address the age-invariant face recognition task, and then, introduced a center loss function to help the CNNs to learn the discriminative features for FR [25]. However, all of these works primarily focus on the network construction and network learning, while none of them are concerned about the occlusion problem in FR.

The pioneer SRC FR system was proposed by Wright *et al.* [17]. In their work, the occluded face image was first coded via l_1-norm minimization as a sparse linear combination of the expanded dictionary which consists of the training samples and the occlusion dictionary. Then, the classification was conducted by searching for which class of training samples could result in the minimum reconstruction error with the sparse coding coefficients.

Following Wright *et al.*'s work, many researchers worked towards improving the SRC accuracy under various conditions. Deng *et al.* [18] employed an auxiliary intraclass variant dictionary to improve the generalization ability for undersampled FR under variable expressions, illuminations

and disguise, and then, they proposed SSRC to learn the prototype images as the gallery dictionary to obtain a better gallery dictionary to deal with non-linear variations [19]. Yang and Zhang [26] and Yang *et al.* [27] introduced the Gabor feature into the SRC framework to compress the occlusion dictionary and reduce the computational cost in coding occluded face images. Ou *et al.* [28] proposed to learn the occlusion dictionary from the data and incorporate a mutual incoherence regularization term into the dictionary learning objective function, which aims at weakening the correlation between the occlusion dictionary and the training sample dictionary. Wang *et al.* [29] improved the robustness to illumination variations and occlusions by using a manifold regularized local sparse representation model. Zhang *et al.* [30] proposed a mixed norm sparse representation classification method to exploit the correlation among the variance face images in the query for multi-view recognition. Zhuang *et al.* [31] introduced a sparse illumination learning and transfer technique to increase the robustness of FR with a single face training sample under the condition of misalignment, illumination variation and random corruption. Zhao *et al.* [32] combined the forward and backward sparse representation together for robust FR. Gao *et al.* [33] improved the performance under the condition of the small number of labeled examples by using a semi-supervised sparse representation-based classification approach.

However, due to the use of l_1-norm minimization, the SRC is computationally expensive as the dictionary size grows. Shi *et al.* [34] and Zhang *et al.* [35] independently argued that the l_1-norm does not play a critical role in robust FR. Based on such observations, Zhang *et al.* [35] put forward to use the squared l_2-norm minimization to gain significant advantage on the computational cost over the traditional SRC algorithm. Cai *et al.* [36] introduced a probabilistic collaborative representation based classifier (ProCRC), which jointly maximizes the likelihood that a test sample belongs to each of the multiple classes. Liu *et al.* [37] employed an iterative relaxed collaborative representation model with adaptive weights learning to enhance the resolution of face images corrupted by noise.

Nevertheless, all of these SRC based works focus on the image space or linear feature space and none of them address the occlusion problem in the deep feature space.

Recently, with the rapid development of generative adversarial networks (GAN), some efforts have been made to apply the GAN to recover occluded face image. Li *et al.* [38] use a deep generative model to generate the corrupted part of the face. The deep generative model consists of an encoding-decoding generator and two adversarial discriminators to synthesize the missing contents from random noise. However, this model could not well handle some unaligned faces, and moreover, this method needs to know the occlusion position in advance. Zhao *et al.* [39] introduced a robust LSTM-autoencoders model, consisting of a multi-scale spatial LSTM encoder to produce an occlusion-robust representation of the face and a dual-channel LSTM decoder to recurrently remove the occlusion in the image space. These methods, acted as preprocessing procedures for FR, are usually complex and, depending on the problem scales, at times require very large scale occluded training datasets.

III. THE PROPOSED DDRC

A. FEATURE EXTRACTION

In the proposed DDRC, the CNNs are first applied to extract the face features. Suppose that the size of the input patch and the length of the output feature vector of the CNNs are $h \times w$ and l, respectively. The CNNs actually perform a nonlinear mapping from the image space to the deep feature space, i.e.,

$$f : \mathbb{R}^{h\times w} \longmapsto \mathbb{R}^{l} . \tag{3}$$

The construction and training of the CNNs are out of the scope of this paper. Here, we should note that in spite of the fact that the VGG-Face networks [40] are the only CNNs evaluated in following experiments, any CNNs well pre-trained with numerous extra face images can be applied to the proposed DDRC.

B. DICTIONARY REPRESENTATION

Suppose there are K subjects and the ith subject has k_i training samples. Let $\mathbf{u}_{i,j} \in \mathbb{R}^l, j = 1, 2, \cdots, k_i$ denote the deep feature vector of the jth training sample of the ith subject and $\mathbf{D}_i = [\mathbf{u}_{i,1}, \mathbf{u}_{i,2}, \cdots, \mathbf{u}_{i,k_i}] \in \mathbb{R}^{l\times k_i}$ the matrix formed by stacking all the $\mathbf{u}_{i,j}$'s of the ith subject. Assume that the deep feature vectors of the non-occluded face samples associated with the same subject lie in a low-dimensional subspace. If there are sufficient training samples for the ith subject, the subspace of the ith subject can be approximated with the linear span of $\mathbf{D}_i$. For a test sample $\mathbf{y}$, if it belongs to the ith subject, its deep feature vector $\mathbf{v}$ will be able to be represented with a very small error as a linear combination of the column vectors of $\mathbf{D}_i$.

However, the membership of the test sample is unknown beforehand. It turns out that we need to figure out the subject that has the best approximation of $\mathbf{v}$ within its subspace. By concatenating the $\mathbf{D}_i$ of all subjects we get a gallery

$$\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K] \in \mathbb{R}^{l\times k}, \tag{4}$$

where $k = \sum_{i=1}^{K} k_i$. Then, the linear representation of $\mathbf{v}$ on the space spanned by the entire training set can be written as

$$\mathbf{v} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{z}, \tag{5}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_i^T, \ldots, \boldsymbol{\alpha}_K^T]^T$ is the coding coefficient vector and $\mathbf{z}$ denotes the noise term. Here, $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,k_i}]^T \in \mathbb{R}^{k_i}$ is the coefficient vector associated with the ith training subject. Using only the coefficients associated with the ith training subject, we can approximate $\mathbf{v}$ of the given test sample as $\hat{\mathbf{v}}_i = \mathbf{D}_i\boldsymbol{\alpha}_i$. Then, we classify $\mathbf{y}$ by assigning it to the subject that minimizes the residual between $\mathbf{v}$ and $\hat{\mathbf{v}}_i$,

$$identity(\mathbf{y}) = \underset{i}{\operatorname{argmin}}\, r_i(\mathbf{v}), \tag{6}$$

where $r_i(\mathbf{v}) = ||\mathbf{v} - \hat{\mathbf{v}}_i||_2$.

Obviously, if the size of the gallery is greater than the feature dimension, i.e., $k > l$, Eq.(5) is underdetermined and the number of possible representations is infinite. This difficulty can be resolved by imposing a regularization constraint to $\boldsymbol{\alpha}$, and then, the solution is given by

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\{||\mathbf{v} - \mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda g(\boldsymbol{\alpha})\}, \tag{7}$$

where λ is a positive regularization parameter and $g(\boldsymbol{\alpha})$ is the regularization function.

Conventionally, l_1-norm, $g(\cdot) = ||\cdot||_1$, and squared l_2-norm, $g(\cdot) = ||\cdot||_2^2$, are two frequently used regularization functions. The l_1-norm minimization leads to a sparse solution and is robust to the outlier [41]. However, it is computationally expensive, even with fast implementation, such as the interior-point method [42] and the dual augmented Lagrangian method (DALM) [43]. The squared l_2-norm minimization has the advantage of lower computational cost and it can achieve competitive classification accuracy compared with the l_1-norm for FR in the image space or the linear feature space [27], [35]. When the squared l_2-norm minimization is used, the solution of Eq.(7) can be analytically derived as

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I}\right)^{-1}\mathbf{D}^T\mathbf{v}\,. \tag{8}$$

Let $\mathbf{P} = \left(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I}\right)^{-1}\mathbf{D}^T$. Clearly, $\mathbf{P}$ is independent of $\mathbf{v}$ and can be calculated in advance. Consequently, the computational complexity will only be proportional to the number of rows of $\mathbf{P}$ for a given deep feature space. This means that the squared l_2-norm minimization has a complexity of $O(n)$. The linear computational complexity is a preferred property for realistic FR applications. Hence, we adopt the squared l_2-norm in the proposed DDRC. In addition, as recommended in [35] and verified by us, the l_2-norm "sparsity" $||\boldsymbol{\alpha}_i||_2$ can also bring some discrimination information for classification. Therefore, we use $||\boldsymbol{\alpha}_i||_2$ to normalize the residual $r_i(\mathbf{v})$ in Eq.(6).

C. OCCLUSION

In this section, we extend the above proposed DDRC to deal with occlusion. Suppose the occluded test sample $\mathbf{y}$ belongs to the ith subject. If the impaction of the occlusion on the deep feature $\mathbf{v}$ is small, i.e., the magnitude of $\boldsymbol{\epsilon}$ in Eq. (1) is small, we can treat $\mathbf{v}$ approximately as if it stays in $\boldsymbol{\Omega}_i$. On the contrary, if $\mathbf{v}$ is heavily affected by the occlusion, i.e., the magnitude of $\boldsymbol{\epsilon}$ is large, $\mathbf{v}$ will move out of $\boldsymbol{\Omega}_i$.

For the latter case, we assume that the deep feature vectors of the face images associated with the same occlusion pattern lie in the same deep feature subspace denoted by $\boldsymbol{\Omega}_{oj}$. Here, the subscript j is the index of the occlusion pattern. The subspace associated with the sunglasses occlusion pattern in Fig. 1 is an example to show the rationality of such assumption (see the explanation in section I).

The occlusion can be regarded as a mapping from the subspace $\boldsymbol{\Omega}$ of non-occluded faces to the subspace $\boldsymbol{\Omega}_{oj}$. Suppose

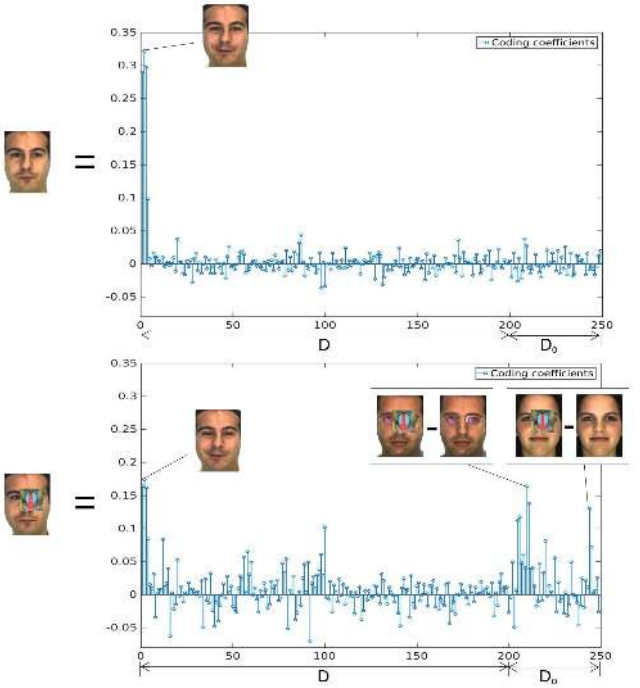

FIGURE 2. Illustration of the coding coefficients for a non-occluded face sample (top) and an occluded face sample (bottom). The VGG-Face network is employed as the deep feature extractor. Instead of the feature vectors, the corresponding face images are presented for easy understanding. For the occluded face sample, some atoms of the auxiliary part have quite large coefficients even comparable to the coefficients associated with the correct subject in the gallery part. This indicates that the linear combination of auxiliary atoms effectively represents the mapping vector introduced by the occlusion.

that the mapping is bijective, then, the relationship between $\mathbf{v}_0$ and $\mathbf{v}$ can be written as

$$\mathbf{v} = \mathbf{v}_0 + \Phi_j(\mathbf{v}_0), \tag{9}$$

where $\Phi_j(\cdot)$ is a function related to the mapping of the jth occlusion pattern.

Let $\boldsymbol{\epsilon}_{j,i} \in \mathbb{R}^l$ be the ith mapping vector associated with the jth occlusion pattern. Actually, $\boldsymbol{\epsilon}_{j,i}$ is the difference between the feature vectors of the non-occluded and the occluded faces associated with the same subject. We observe that $\Phi_j(\cdot)$ can be approximately modeled by a linear combination of sufficient mapping vectors associated with the jth occlusion pattern, i.e.,

$$\Phi_j(\mathbf{v}_0) \approx \mathbf{D}_{oj}\boldsymbol{\beta}_j, \tag{10}$$

where $\mathbf{D}_{oj} = [\boldsymbol{\epsilon}_{j,1}, \ldots, \boldsymbol{\epsilon}_{j,q_j}] \in \mathbb{R}^{l\times q_j}$ is the auxiliary dictionary associated with the jth occlusion pattern, which is formed by stacking all the training mapping vectors, q_j is the number of training mapping vectors, and $\boldsymbol{\beta}_j \in \mathbb{R}^{q_j}$ the coding coefficient vector. An example to illustrate the observation is shown in Fig. 2. We further assume that the approximation of Eq. (10) is held generically for occluded face images.

Suppose there are Q types of occlusion patterns. Since the occlusion pattern of the test sample is unknown at the beginning, we concatenate the auxiliary dictionaries of Q

occlusion patterns to form the auxiliary dictionary for the proposed DDRC, i.e.,

$$\mathbf{D}_o = [\mathbf{D}_{o1}, \mathbf{D}_{o2}, \ldots, \mathbf{D}_{oQ}] \in \mathbb{R}^{l \times k_o}, \tag{11}$$

where $k_o = \sum_{j=1}^{Q} q_j$. Then, $\Phi_j(\mathbf{v}_0)$ can be coded as $\Phi_j(\mathbf{v}_0) \approx \mathbf{D}_o\boldsymbol{\beta}$ with $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_Q] \in \mathbb{R}^{k_o}$. Considering Eq. (5) the dictionary representation of the feature vector of the non-occluded face, we can model $\mathbf{v}$ by

$$\mathbf{v} = \mathbf{D}_a\boldsymbol{\omega} + \mathbf{z}, \tag{12}$$

where $\mathbf{D}_a = [\mathbf{D}, \mathbf{D}_o]$, and $\boldsymbol{\omega} = [\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T]^T$

Generally speaking, $\mathbf{D}_a$ is overcomplete and some kind of regularization restriction must be imposed on $\boldsymbol{\omega}$ to obtain the unique solution. As discussed in section III-B, the squared l_2-norm is a proper trade-off between the accuracy and the computational cost. By using the squared l_2-norm minimization, similar to Eq.(8), $\boldsymbol{\omega}$ can be analytically estimated by

$$\hat{\boldsymbol{\omega}} = \mathbf{P}_a\mathbf{v}, \tag{13}$$

where $\hat{\boldsymbol{\omega}} = [\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T]^T$ and $\mathbf{P}_a = \left(\mathbf{D}_a^T\mathbf{D}_a + \lambda\mathbf{I}\right)^{-1}\mathbf{D}_a^T$.

After obtaining the estimated coding coefficients $\hat{\boldsymbol{\omega}}$, we can approximately recover the deep feature vector in the subspace $\boldsymbol{\Omega}$ for the occluded face as

$$\mathbf{v}_0 = \mathbf{v} - \mathbf{D}_o\hat{\boldsymbol{\beta}} \tag{14}$$

Then, the classification is done by comparing the similarity of $\mathbf{v}_0$ to the feature subspace of the non-occluded faces associated with each subject. This can be achieved by replacing the $\mathbf{v}$ in Eq. (6) with the $\mathbf{v}_0$ in Eq.(14).

In many situations, if we cannot collect sufficient training samples and mapping vectors, Eq. (13) might lead to an overfitting solution. To avoid such problem, we first use the principal component analysis (PCA) to reduce the dimensionality of $\mathbf{D}_a$ from $l \times (k + k_o)$ to $m \times (k + k_o)$, where $m < l$. Accordingly, the deep feature $\mathbf{v}$ of the test sample should be first projected onto this m-dimension space.

The implementation details of the proposed DDRC algorithm is summarized in Algorithm 1.

To illustrate the working principle of the proposed DDRC, Fig. 2 shows the coding coefficients for a non-occluded face sample and an occluded face sample.

D. GENERATION OF AUXILIARY DICTIONARY

In the proposed DDRC approach, we only assume that $\Phi_j(\mathbf{v}_0)$ is related to the occlusion pattern. This assumption implies that it is unnecessary to associate the atoms of $\mathbf{D}_o$ with the training subjects in the gallery part. Therefore, the auxiliary dictionary can be generated with the subjects from either inside or outside the training set.

In general, $\boldsymbol{\epsilon}_{j,i}$ can be obtained in various ways as long as they can reflect the mapping from $\boldsymbol{\Omega}$ to $\boldsymbol{\Omega}_{oj}$. For instance, let $\mathbf{x}_{oj,i}$ be the occluded face sample associated with the jth occlusion pattern, and $\mathbf{x}_i$ the non-occluded counterpart of $\mathbf{x}_{oj,i}$ associated with the same subject, which is acquired under the same or similar facial expression, head pose, and environment condition as $\mathbf{x}_{oj,i}$. Then, $\boldsymbol{\epsilon}_{j,i}$ can be calculated by

$$\boldsymbol{\epsilon}_{j,i} = \mathbf{u}_{oj,i} - \mathbf{u}_i, \tag{15}$$

where $\mathbf{u}_{oj,i}$ is the deep feature vector of $\mathbf{x}_{oj,i}$ and $\mathbf{u}_i$ the deep feature vector of $\mathbf{x}_i$. This method is named *pair-matching*.

However, in many situations, the non-occluded counterpart may not be collected. Instead, we can use a 'natural' face image of the same subject as an alternative to $\mathbf{x}_i$. We name this method *natural-matching*. The $\boldsymbol{\epsilon}_{j,i}$ generated with the *natural-matching* method may contain some information related to other kinds of variations, such as illumination changes and facial expressions. This irrelevant information, as shown in section IV-A1, may undermine the ability of the auxiliary dictionary to represent the occlusion mapping and eventually degrade the classification accuracy.

Algorithm 1 The DDRC Algorithm

1) For each training and testing sample, the CNN is used to extract the features.
2) Construct the dictionary $\mathbf{D}_a$ with the deep features from the training samples.
3) Employ PCA to reduce the number of rows of $\mathbf{D}_a$ to m and project the deep feature $\mathbf{v}$ of the test sample onto the m-dimension space.
4) Normalize the columns of $\mathbf{D}_a$ to have unit l_2-norm.
5) For each test sample $\mathbf{y}$:
   a) Normalize $\mathbf{v}$ to have unit l_2-norm.
   b) According to Eq.(12), code $\mathbf{v}$ over $\mathbf{D}_a$ by $\hat{\omega} = \left(\mathbf{D}_a^T\mathbf{D}_a + \lambda\mathbf{I}\right)^{-1}\mathbf{D}_a^T\mathbf{v}$.
   c) Compute the residuals with respect to each training subject by using $r_i(\mathbf{v}) = \frac{\|\mathbf{v} - \mathbf{D}_i\hat{\boldsymbol{\alpha}}_i - \mathbf{D}_o\hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\alpha}}_i\|_2}$.
   d) Output the identity of $\mathbf{y}$ as identity$(\mathbf{y}) = \arg\min_i\{r_i(\mathbf{v})\}$.

IV. EXPERIMENTS

In this section, we present experiments on publicly available databases to demonstrate the performance of the proposed DDRC. In the experiments, the VGG-Face network with configuration A [40] is adopted as the deep feature extractor for the proposed DDRC, because it has a publicly available pre-trained model and a demonstrated face verification accuracy close to human vision on the LFW dataset. The VGG-Face network consists of 16 convolutional layers, 5 max-pooling layers, and 3 fully-connected layers. It takes color image patch of size 224 × 224 as the input. To satisfy the input requirement of the VGG-Face network, the face images are first resized to 224×224, and the grayscale images are transformed to three-channel pseudo-color images. The deep feature vector, which has a dimension of 4096, is taken from the second last fully-connected hidden layer neuron activations of the VGG-Face network. The implementation of VGG-Face network used in the experiments is based on

the MatConvNet [44]. The experiments were conducted on a computer with an Intel i7 CPU and without GPU acceleration.

A. EVALUATION ON AR DATABASE

As adopted in [17], a subset face crops of the AR database [15], [45] that contains 50 male and 50 female subjects are used for evaluation. For each subject, 26 images are recorded in two different sessions separated by two weeks. Each session consists of 13 images with different facial variations, including illumination changes, expressions, and facial disguises, as shown in Fig. 1.

1) PERFORMANCE EVALUATION

In this experiment, we divided the 100 subjects into two subsets, a testing subset consisting of 50 subjects randomly selected for testing and an occlusion subset consisting of the remaining 50 subjects (no overlapping with the testing subset) for generating the auxiliary dictionary. For each subject in the testing subset, the seven non-occluded images with illumination changes and expressions from session 1 were used as training samples, and the other images from session 1 and all the images from session 2 were used for testing.

In accordance to the experiment setting in [27], the original 165 × 120 color images were converted to 83 × 60 gray scale images. Four algorithms were considered for comparison in the experiment: VGG-Face network followed by a softmax classifier (VGG+softmax) [40], sparse representation based classification (SRC) [41], Gabor feature based sparse representation based classification with l_1 norm regularization (GRRC_l_1) [27], and the proposed DDRC. The VGG+softmax method can be regarded as the original VGG-Face network with the last two layers (the last fully-connected layer and softmax prob layer) fine-tuned on the task-specified dataset.

Before comparing different algorithms, we first studied the impact of parameter selection on the performance of the proposed DDRC. The experiments were conducted on the face images disguised with sunglasses from session 1.

The result is shown in Fig. 3, where the auxiliary dictionary generated by using all 50 subjects with *pair-matching* method includes both sunglasses and scarf occlusion patterns. From Fig. 3, we can observe that a wide range of λ results in quite close FR accuracy. This observation indicates that the proposed DDRC is less sensitive to the variation of λ. We can also learn that the excessive reduction of the feature vector dimension by using PCA can dramatically increase the FR error rate, *e.g.*around 20% loss in FR rate for the case of 50 comparing to the other cases. This can be attributed to the loss of important discriminative information for the excessive reduction.

In Fig.4, four types of auxiliary dictionaries were studied: *pair-matching* with sunglasses and scarf occlusion patterns (pair sunglasses+scarf), *pair-matching* with sunglasses occlusion pattern (pair sunglasses), *natural-matching* with sunglasses and scarf occlusion patterns (natural sunglasses+scarf), and *natural-matching* with sunglasses

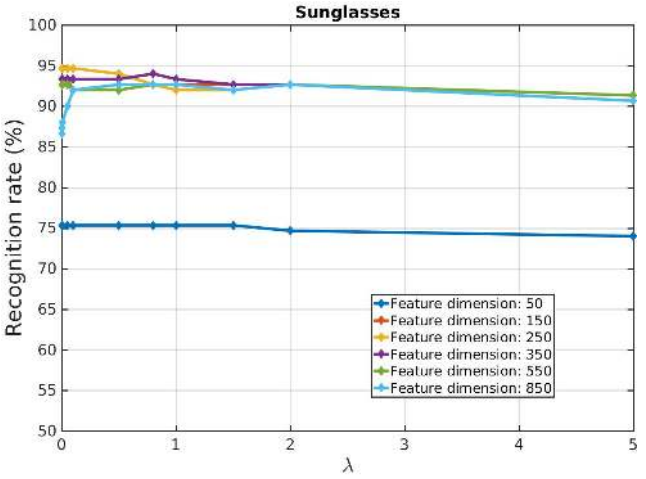

FIGURE 3. Recognition rate of DDRC with respect to λ for various feature dimension reductions with PCA for the sunglasses from session 1 on a subset of the AR database.

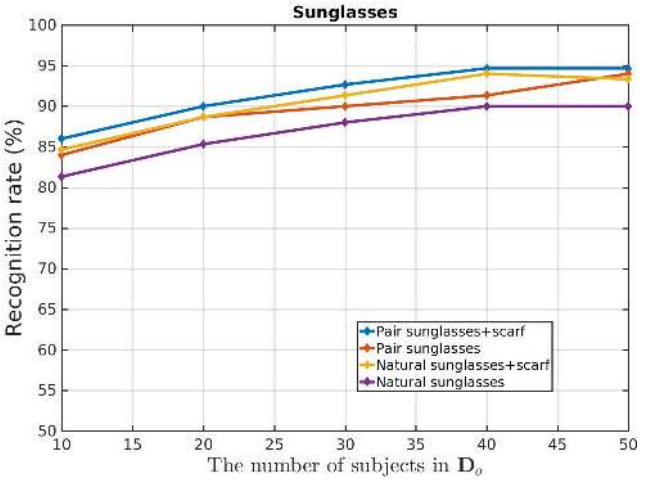

FIGURE 4. Recognition rates of the proposed DDRC for the sunglasses from session 1 on a subset of the AR database versus the number of subjects in the occlusion dictionary for different generation methods of the occlusion dictionary.

occlusion pattern (natural sunglasses). For each subject in the occlusion subset, there are 6 mapping vectors associated with the sunglasses occlusion pattern and 6 mapping vectors associated with the scarf occlusion pattern. Therefore, the number of atoms in an auxiliary dictionary is proportional to the number of subjects in the auxiliary. For instance, if the auxiliary dictionary involving both sunglasses and scarf occlusion patterns is generated by using 10 subjects, it will contain 120 atoms. The experiments were conducted with the parameters $m = 250$ and $\lambda = 0.05$.

From Fig. 4, we can see that the more subjects that are used to generate the auxiliary dictionary, the higher the recognition rate that can be achieved. Furthermore, we can observe that the scarf occlusion pattern, if it is included in the auxiliary dictionary, can slightly improve the FR accuracy for the face images occluded with sunglasses. That may be because some responses in the deep feature space are shared by these two

TABLE 1. Recognition rates(%) on the AR database with disguise.

	Sunglasses session 1	Sunglasses session 2	Scarf session 1	Scarf session 2	Non-occlusion session 2
VGG+softmax	68.0	66.0	98.0	96.0	**100**
SRC	86.7	52.7	54.7	36.2	93.7
GRRC_l_1	94.0	66.0	96.7	94.0	96.9
DDRC	**94.7**	**85.3**	**99.3**	**98.7**	**100**

TABLE 2. Single-sample FR rates(%) on the AR database.

	Expression	Illumination	Disguise	Disguise + Illumination	Average
VGG+softmax	97.1	98.0	81.3	69.4	85.4
SRC	85.8	80.8	43.7	22.8	56.6
ESRC	92.5	98.7	85.6	77.8	88.0
GRRC_l_1	89.2	98.3	82.5	67.8	83.2
DDRC_SS	**97.5**	**99.6**	**95.6**	**90.0**	**95.2**

occlusion patterns and the more information related to the occlusion pattern the auxiliary dictionary involves, the better representation of the occlusion pattern the auxiliary dictionary can achieve. On the contrary, if the information unrelated to the occlusion is involved in the auxiliary dictionary, the representation ability of the auxiliary dictionary can be hurt. For instance, the *natural-matching* method, which introduces more information about the illumination change, facial expressions etc. into the auxiliary dictionary, shows worse performance than the corresponding *pair-matching* method. Therefore, in the following experiments on the AR database, the *pair-matching* method is used to generate the auxiliary dictionary and both sunglasses and scarf occlusion patterns are included in the auxiliary dictionary.

The performance comparison of different algorithms is presented in Table 1. For the proposed DDRC, the auxiliary dictionary, consisting of 600 atoms, was generated with all 50 subjects in the occlusion subset and m and λ were set to 250 and 0.05, respectively. For the GRRC_l_1, the source codes were provided by the authors and the default setting for the AR database was used. For the SRC, we implement the error constrained model with the same error tolerance as in their original paper [41], i.e., $\epsilon = 0.05$. To better understand the performance improvement of the proposed DDRC with respect to facial variations, we separately list the FR rate results for sunglasses disguise in session 1, sunglasses disguise in session 2, scarf disguise in session 1, scarf disguise in session2, and non-occlusion case in session 2.

In Table 1, it is evident that the proposed DDRC achieves the highest FR rates in all five testing scenarios. In comparison with the VGG+softmax method for the ''Sunglasses'' scenarios, the proposed DDRC has at least 19.3% increase in FR rate. We can also find that the VGG+softmax achieves the second highest recognition rates (98% and 96%) for the two cases of scarf occlusion. The reason might be that the discriminative features captured by the VGG-Face network are less related to the jaw and mouth than the other parts of the face. We should note that even for such high FR rates, the proposed DDRC can effectively increase the FR accuracy as well. These results demonstrate that the proposed DDRC can effectively handle the occlusion and significantly improve the FR accuracy of the original network under the occlusion scenarios.

In addition, we can observe that the same perfect FR rates on the non-occlusion testing subset from session 2 are achieved by both the DDRC and VGG+softmax. This is because the dictionary representation classifier can effectively discriminate the linear subspace in deep feature space. This result shows that the proposed DDRC is a generic approach for both occlusion and non-occlusion scenarios.

2) SINGLE-SAMPLE FR

There is a situation frequently confronted in practical applications that only one training sample per subject can be provided, such as recognizing individuals based on their ID card photos. To address this problem, we evaluated the robustness of the DDRC with one training sample per subject in this section.

In order to make a fair comparison, we adopted the experiment setting in [18], i.e., the testing subset consists of 80 randomly selected subjects from the AR database, the occlusion subset consists of the remaining 20 subjects, the natural image for each subject from session 1 was used as the training sample, and the other 12 images for each subject from session 1 were selected for testing. The *pair-matching* method with all the 20 subjects in the occlusion subset is used to generate the occlusion dictionary for the DDRC. Both sunglasses and scarf occlusion patterns were included in the auxiliary dictionary such that the auxiliary dictionary has the same number of atoms as that in [18], i.e., 240 atoms. The feature dimension m and the regularization parameter λ were set to 320 and 0.35 for the proposed DDRC, respectively. The proposed DDRC with a single training sample per subject is named as DDRC_SS in the following experiments. In Table 2, the results of the SRC and ESRC are reported from the original paper [18]. Except for the SRC and ESRC, all other methods listed in Table 2 adopt the same converted images as in section IV-A1.

From Table 2, we can observe that, even with insufficient occlusion samples, DDRC_SS achieves the best FR accuracy (over 90%) in all the testing scenarios. For the scenarios with occlusion (''Disguise'' and ''Disguise+Illumination''), at least 10% increase in FR rate is achieved when compared to all the other methods. In comparison with the VGG+softmax method, the proposed DDRC achieves 14% and 20.6% increase in FR rate for the ''Disguise'' and ''Disguise+Illumination'' scenarios, respectively.

The VGG+softmax and the proposed DDRC both have nearly perfect FR accuracy for the testing scenarios of ''expression'' and ''illumination''. This indicates that the VGG-Face deep feature space is more discriminative than the

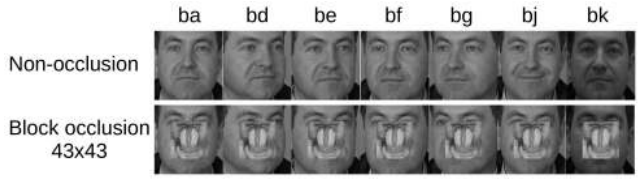

FIGURE 5. Examples of a subject of the pose subset of the FERET database (top row) and block occlusion with block size of 43 × 43 on the center of each sample (bottom row).

image space. Because of this property, it is unnecessary for the DDRC to handle the facial variation that is not a result of the occlusion or corruption. Consequently, less work is needed by the DDRC_SS to generate the auxiliary dictionary than by the ESRC, because only the occlusion images need to be collected. The results again demonstrate that the proposed DDRC is a generic approach for both occlusion and non-occlusion scenarios.

B. EVALUATION ON FERET POSE DATABASE

In addition to illumination change and expressions, another situation frequently encountered in real face images is the head pose variation. To evaluate the performance on head pose variation, the pose subset of the FERET database [46] was selected, which includes 194 subjects (about 7 images each). This subset is composed of the images marked with 'ba','bd','be','bf','bg','bj', and 'bk' for natural, pose variations of $+25°$, $+15°$, $-15°$, and $-25°$, expression and illumination, respectively. These images were cropped to the size of 80 × 80. Some example images of a subject are shown in Fig.5. Since the original pose subset of the FERET database does not have the variation caused by occlusion, we simulated the block occlusion by replacing a square block centered on each resized original image with an unrelated image. Four different sizes of the block occlusion were considered in the simulation, 25×25, 35×35, 43×43, and 50×50 corresponding to occlusion ratios of 9.7%, 19.1%, 28.9% and 39.1%, respectively. The block occlusion examples with the block size of 43 × 43 are shown in Fig.5.

In the experiments, 150 subjects randomly drawn from all 194 subjects were used for testing and constructing the gallery, and the other 44 subjects were used to generate the auxiliary dictionary for the proposed methods. The performance on different ratios of block occlusion are tested separately and the results are listed in Table 3. The 'ba', 'bk' and 'bj' images of each subject from the original non-occlusion subset are chosen as the training samples (except for the DDRC_SS, which only employs the 'ba' image as the training sample). The GRRC_l_1 used the same setting as in [27]. The parameters of the proposed DDRC and DDRC_SS are set to $m = 600$ and $\lambda = 0.5$. The auxiliary dictionary is generated with the *pair-matching* method from the 44 subjects.

The evaluation results are listed in Table 3, from which we can see that, for the occlusion ratio less than around 30%, the proposed DDRC and DDRC_SS achieves the best performance in FR rate. Even with the block occlusion of

TABLE 3. Recognition rates(%) on the FERET database with block occlusion.

Block size	25 × 25	35 × 35	43 × 43	50 × 50
Occlusion ratio	9.7%	19.1%	28.9%	39.1%
VGG+softmax	95.8	86.4	71.2	28.4
SRC	64.5	60.0	53.5	49.0
GRRC_l_1	74.4	70.8	68.2	**62.2**
DDRC_SS	96.1	91.0	83.1	48.7
DDRC	**98.6**	**94.5**	**88.2**	54.5

28.9%, the proposed DDRC can reach an FR rate close to 90%. However, for the extremely large size of occlusion, *e.g.*50×50, the occlusion cannot be recovered effectively by the proposed DDRC and the FR rate dropped dramatically. We can also note that for the block occlusion of small size, *e.g.*25×25, the VGG+softmax achieves similar high FR rate to the proposed DDRC. This indicates that the block occlusion of small size may not map the deep feature vector out of the subspace associated with non-occlusion faces of the subject. The recognition performance on the non-occlusion images is also evaluated for the proposed DDRC and a 100% accuracy was observed. So we do not list the result in Table 5.

C. EVALUATION ON CelebA DATABASE

This experiment is designed to evaluate the performance of the proposed DDRC on a large-scale database. The CelebA database [47] is adopted in the experiment, since it is a large-scale face database containing large variations of pose, expression, and illumination etc. Because the training dataset of CelebA is originally used to train the CNNs, we only use the test dataset and part of the validation dataset of CelebA for evaluation. In the experiment, due to few occluded faces in the original CelebA dataset, the synthesized occluded faces are used for evaluation.

The test dataset of CelebA consists of 1, 000 subjects with 19, 963 images. We partitioned the test dataset of CelebA near equally into two parts with randomly drawn method: the training set and testing set. Each part consists of around 10, 000 images. Since three subjects have only 1 image each, we put them into the training set to make sure that each identity has at least one training sample. In total, the training set and testing set cover 1, 000 and 997 subjects, respectively. The training set is used to build the gallery dictionary and the testing set is employed to synthesize the occluded test face images. From the validation dataset of CelebA, we randomly selected 20 identities each with randomly drawn 20 images to construct the auxiliary dictionary. Two occlusion patches shown in Fig.6 are used to contaminate the face images on the center.

Only the VGG+softmax and proposed DDRC were evaluated, because the methods based on the image space or linear transform of image space and single-sample based method do not suit to handle a large-scale dataset with large variation. In the experiment, we crop the face in each image

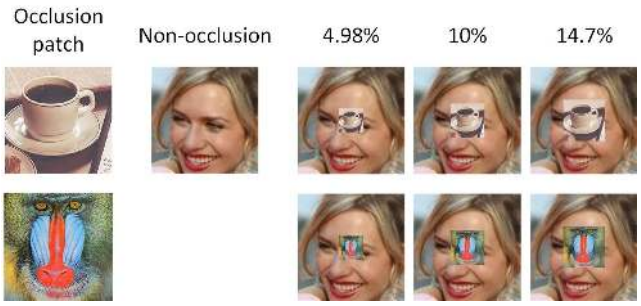

FIGURE 6. Occlusion patches and example with images various occlusion ratios for the test dataset of the CelebA database.

TABLE 4. Recognition rates(%) on the test dataset of CelebA database with block occlusion.

Occlusion ratio	0%	4.98%	10%	14.7%
VGG+softmax	**92.3**	71.8	51.9	33.5
DDRC	91	**76.1**	**62.4**	**46.8**

with the size of 128 × 128 first and for the occluded face image, the occlusion patch is superimposed at the center of the face image. Four occlusion ratios were tested, 0% (non-occlusion), 4.98%, 10%, and 14.7%. The examples of occluded face crops with various occlusion ratios are shown in Fig.6.

For the proposed DDRC, the gallery dictionary is constructed with all the face images of the training set. The auxiliary dictionary includes all 4 occlusion ratios and are constructed with the *pair-matching*. The parameters of the proposed DDRC are set as $m = 1000$ and $\lambda = 0.05$. For the VGG+softmax, the softmax classifier is trained using the training set.

The experimental results are listed in Table 4. From the results, we can observe that the proposed DDRC significantly improves the FR rate for the occluded face images (13.3% increase in FR rate for 14.7% occlusion). We also note that the proposed DDRC is slightly worse than the VGG+softmax for the non-occlusion case. We reevaluate the proposed DDRC without PCA for the non-occlusion case and observe that the FR rate reaches 92.3%. Therefore, the slightly low FR rate at non-occlusion case is caused by the PCA dimension reduction.

D. COMPUTATIONAL COMPLEXITY

The computational cost of the proposed DDRC mainly lies in the calculation of the deep features. The maximum time cost for the feature extraction is less than 0.4 seconds per image.

The step 5b of Algorithm 1 has a computational complexity of $O(n)$ with respect to the number of the rows of $\mathbf{P}_a$ (see the analysis in section III-B). The step 5c of Algorithm 1 is proportional to the number of the columns of $\mathbf{D}$ for a given deep feature space. Since the number of the rows of $\mathbf{P}_a$ is larger than the number of the columns of $\mathbf{D}$, given the dimensionality of the deep feature, the 5-*th* step of Algorithm 1 has a computational complexity less than $O(2n)$. Therefore, the proposed DDRC is a computationally efficient approach and suitable for real-time large-scale applications.

In the above experiments, for the largest size of $\mathbf{P}_a$, i.e., the $\mathbf{P}_a$ used for the proposed DDRC method in section IV-C, the computational time of the 5-*th* step of Algorithm 1 is around 16ms per test image and is irrelevant to the test image.

V. CONCLUSION

In this paper, we have proposed an effective method to alleviate the occlusion effect in deep feature space for FR. The proposed DDRC is generic to the FR of both occluded and non-occluded face images. In addition, the experiments show that the proposed DDRC can well handle the FR with a single training sample per subject for simple application scenarios. From the computational complexity analysis and the experiments, we demonstrate that the proposed DDRC is a promising real-time algorithm for practical FR applications.

The weakness of the proposed method lies in the assumption that the occlusion pattern of the testing face is included in the auxiliary dictionary. However, in many face recognition applications, the major types of the occlusion patterns are limited. As a result, the proposed method can effectively boost the performance of the FR in practical applications. In the future, we will improve the design of the auxiliary dictionary to handle the occlusion patterns outside the auxiliary dictionary.

REFERENCES

[1] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.
[2] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.
[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.
[4] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
[5] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 529–534.
[6] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[8] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[9] K. Zhang, M. Sun, X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1303–1314, Jun. 2018.
[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. (2017). "VGGFace2: A dataset for recognising faces across pose and age." [Online]. Available: https://arxiv.org/abs/1710.08092
[11] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (2014). "Learning face representation from scratch." [Online]. Available: https://arxiv.org/abs/1411.7923
[12] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7044–7053.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.
[14] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[15] A. M. Martinez, "The AR face database," CVC, New Delhi, India, Tech. Rep. #24, 1998.
[16] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[17] 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.
[18] 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.
[19] W. Deng, J. Hu, and J. Guo, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 399–406.
[20] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 262–263.
[21] 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.
[22] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2892–2900.
[23] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). "DeepID3: Face recognition with very deep neural networks." [Online]. Available: https://arxiv.org/abs/1502.00873
[24] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4893–4901.
[25] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.
[26] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 448–461.
[27] 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 Recognit.*, vol. 46, no. 7, pp. 1865–1878, Jul. 2013.
[28] 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.
[29] L. Wang, H. Wu, and C. Pan, "Manifold regularized local sparse representation for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 651–659, Apr. 2015.
[30] X. Zhang, D.-S. Pham, S. Venkatesh, W. Liu, and D. Phung, "Mixed-norm sparse representation for multi view face recognition," *Pattern Recognit.*, vol. 48, no. 9, pp. 2935–2946, Sep. 2015.
[31] L. Zhuang, T.-H. Chan, A. Y. Yang, S. S. Sastry, and Y. Ma, "Sparse illumination learning and transfer for single-sample face recognition with image corruption and misalignment," *Int. J. Comput. Vis.*, vol. 114, no. 2, pp. 272–287, 2014.
[32] Z.-Q. Zhao, Y.-M. Cheung, H. Hu, and X. Wu, "Corrupted and occluded face recognition via cooperative sparse representation," *Pattern Recognit.*, vol. 56, pp. 77–87, Aug. 2016.
[33] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.
[34] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 553–560.
[35] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 471–478.
[36] S. Cai, L. Zhang, W. Zuo, and X. Feng, "A probabilistic collaborative representation based approach for pattern classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2950–2959.
[37] L. Liu, S. Li, and C. L. P. Chen, "Iterative relaxed collaborative representation with adaptive weights learning for noise robust face hallucination," *IEEE Trans. Circuits Syst. Video Technol.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8345667. doi: 10.1109/TCSVT.2018.2829758.
[38] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3911–3919.
[39] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 778–790, Feb. 2018.
[40] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, vol. 1. no. 3, p. 6.
[41] J. Wright, A. Ganesh, A. Yang, Z. Zhou, and Y. Ma. (2011). "Sparsity and robustness in face recognition." [Online]. Available: https://arxiv.org/abs/1111.1014
[42] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.
[43] 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.
[44] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[45] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[46] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.
[47] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.

FENG CEN received the Ph.D. degree in computer application technology from Shanghai Jiao Tong University, Shanghai, China, in 2003.

He is currently an Associate Professor with the Department of Control Science and Engineering, Tongji University, Shanghai. His research interests include computer vision, image processing, and pattern recognition.

GUANGHUI WANG (M'10–SM'17) was a Research Fellow and a Visiting Scholar with the Department of Electronic Engineering, The Chinese University of Hong Kong, from 2003 to 2005. From 2006 to 2010, he was a Research Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Canada. He is currently an Assistant Professor with The University of Kansas, USA. He has authored one book, *Guide to Three Dimensional Structure and Motion Factorization* (Springer-Verlag). He has published over 100 papers in peer-reviewed journals and conferences. His research interests include computer vision, structure from motion, object detection and tracking, machine learning, and robot localization and navigation.

Dr. Wang has served as an Associate Editor and on the Editorial Board of two journals, as an Area Chair or TPC member for over 20 conferences, and as a Reviewer for over 20 journals.