Undersampled Face Recognition via Robust Auxiliary Dictionary Learning

Chia-Po Wei and Yu-Chiang Frank Wang, *Member, IEEE*

***Abstract*— In this paper, we address the problem of robust face recognition with undersampled training data. Given only one or few training images available per subject, we present a novel recognition approach, which not only handles test images with large intraclass variations such as illumination and expression. The proposed method is also to handle the corrupted ones due to occlusion or disguise, which is not present during training. This is achieved by the learning of a robust auxiliary dictionary from the subjects not of interest. Together with the undersampled training data, both intra and interclass variations can thus be successfully handled, while the unseen occlusions can be automatically disregarded for improved recognition. Our experiments on four face image datasets confirm the effectiveness and robustness of our approach, which is shown to outperform state-of-the-art sparse representation-based methods.**

***Index Terms*— Dictionary learning, sparse representation, face recognition.**

I. Introduction

FACE recognition has been an active research topic, since it is challenging to recognize face images with illumination and expression variations as well as corruptions due to occlusion or disguise. A typical solution is to collect a sufficient amount of training data in advance, so that the above intra-class variations can be properly handled. However, in practice, there is no guarantee that such data collection is applicable, nor the collected data would exhibit satisfactory generalization. Moreover, for real-world applications, e.g. e-passport, driving license, or ID card identification, only one or very few face images of the subject of interest might be captured during the data acquisition stage. As a result, one would encounter the challenging task of *undersampled* face recognition [1].

Existing solutions to undersampled face recognition can be typically divided into two categories: patch-based methods and generic learning from external data. For patch-based methods, one can either extract discriminative information from patches collected by different images, or utilize/integrate the corresponding classification results for achieving recognition. For example, the former has considered local binary pattern (LBP) [2], Gabor features [3], or manifold learning [4], while the latter advanced weighted plurality voting [5] or margin distribution optimization [6]. Nevertheless, the major concern of patch-based methods comes from the fact that local patches extracted from undersampled training data only contain limited information, especially for the scenario of single-sample face recognition (i.e., one training image per person). As a result, the classification results would degrade significantly when there exists large variations between the query and the gallery ones. Moreover, patch-based methods often assume that the image patches are free from occlusion; this would limit their uses in practical scenarios.

In contrast to patch-based approaches for undersampled face recognition, the second type of methods advocate the use of external data which contain the subjects *not* of interest. These approaches aim at learning the classifiers with improved recognition abilities (see [7], [8]), or modeling the intra-class variations (see [9]–[11]). For example, based on the assumption that the face images of different subjects are independent, adaptive generic learning (AGL) [7] utilized external data for estimating the within-class scatter matrix for each subject to be recognized. Different from AGL which requires the above assumption, Kan *et al.* [8] further proposed a nonlinear estimation model to calculate the within-class scatter matrix.

Different from [7] and [8], recent works like [9]–[11] employed external data for describing possible intra-class variations when performing recognition. Although promising result have been shown in [9] and [10], these approaches require the query image and the external data to exhibit the same type of occlusion, which might not be practical. Since we typically do not have the prior knowledge on the occlusion of concern, how to select external data for learning intra-class variations would become a problem for methods like [9], [10]. Recently, [11] considered the modeling of intra-class variations without using the prior knowledge of occlusion, and it characterized occlusion as sparse errors when performing recognition. As noted in [12], such characterization might not be accurate and would be insufficient to describe the occlusion presented in real-world face images.

In this paper, we advocate the extraction of representative information from external data via dictionary learning without assuming the prior knowledge of occlusion in query images. This framework is considered as *robust auxiliary dictionary learning* (RADL). With the same setting as [9]–[11], we consider the scenario that only *one* or *few* non-occluded training images are available for each subject of interest.

Manuscript received April 29, 2014; revised August 27, 2014 and January 9, 2015; accepted February 25, 2015. Date of publication March 6, 2015; date of current version March 23, 2015. This work was supported in part by the Ministry of Science and Technology under Grant MOST103-2221-E-001-021-MY2 and in part by the National Science Council under Grant NSC102-2221-E-001-005-MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shiguang Shan.

The authors are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11574, Taiwan (e-mail: cpwei@citi.sinica.edu.tw; ycwang@citi.sinica.edu.tw).

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

Digital Object Identifier 10.1109/TIP.2015.2409738

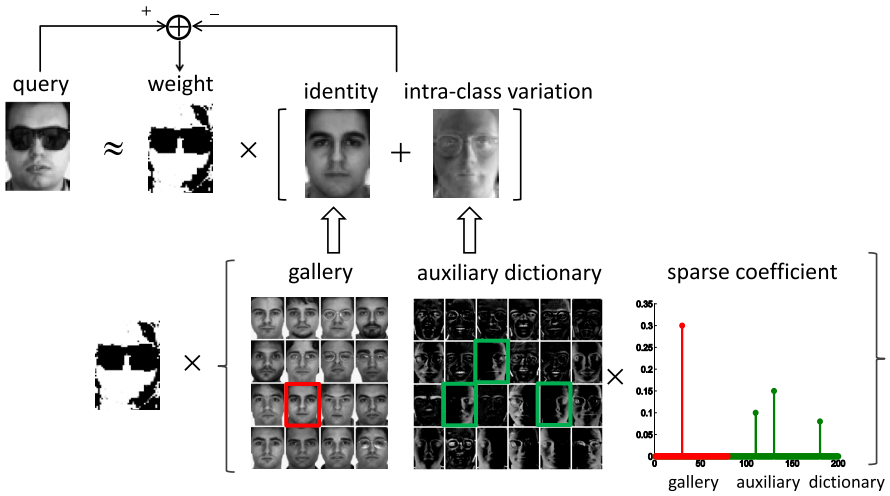


Fig. 1. Illustration of our proposed method for undersampled face recognition, in which the gallery set only contains one or few face images per subject of interest, while the auxiliary dictionary is learned from external data for observing possible image variants. Note that the corrupted image regions of the query input can be automatically disregarded using our proposed method.

Unlike [9], [10], which require the prior knowledge of the occlusion, our approach eliminates such assumptions by introducing a novel classification method based on robust sparse coding. It is worth noting that existing dictionary learning algorithms like KSVD [13] can also be used to learn dictionaries for images from external datasets. However, these learned dictionaries cannot guarantee the recognition performance for the subjects of interest, since KSVD only considers the representation ability of dictionaries. In our work, we jointly solve the tasks of auxiliary dictionary learning and robust sparse coding in a *unified* optimization framework (detailed in Section III). This makes our approach able to improve the performance for robust face recognition under the scenario of undersampled training data.

Fig. 1 illustrates our idea of the proposed method. By learning an auxiliary dictionary from an external dataset together with robust sparse coding, the benefits of our approach are threefold. Firstly, we are able to address undersampled face recognition problem, since only one or few training images of the subjects to be recognized are required for training. Therefore, there is no need to collect a large training dataset for covering image variants for *all* subjects of interest. Secondly, our approach provides a new tool for recognizing occluded face images by means of robust sparse coding and the auxiliary dictionary, while no assumptions are made about the information on occlusion. Finally, our algorithm for auxiliary dictionary learning allows one to model intra-class variations including illumination and expression changes from external data. By solving both auxiliary dictionary learning and robust face recognition in a unified framework, improved recognition performance can be expected.

The remaining of this paper is organized as follows. Section II reviews related works on sparse representation based approaches for face recognition and dictionary learning. In Section III, we present our proposed algorithm for auxiliary dictionary learning and undersampled face recognition, including the optimization details. Experimental results on three face image databases are presented in Section IV. Finally, Section V concludes this paper.

II. Related Work

A. SRC and Extended SRC

Recently, Wright *et al.* [14] proposed sparse representation based classification (SRC) for face recognition. Since our proposed method is extended from SRC, we briefly review this classification technique for the completeness of this paper.

Given a test image $\mathbf{y}$, SRC represents $\mathbf{y}$ as a sparse linear combination of a codebook $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_L]$, where $\mathbf{D}_i$ denotes the training images associated with class i. Precisely, SRC derives the sparse coefficient $\mathbf{x}$ of $\mathbf{y}$ by solving the following L1-minimization problem:

$$\min_{\mathbf{x}} \left\|\mathbf{y} - \mathbf{D}\mathbf{x}\right\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{1}$$

After the sparse coefficient $\mathbf{x}$ is obtained, the test input $\mathbf{y}$ is recognized as class ℓ^* if it satisfies

$$\ell^* = \arg\min_{\ell} \left\|\mathbf{y} - \mathbf{D}\delta_\ell(\mathbf{x})\right\|_2, \tag{2}$$

where $\delta_\ell(\mathbf{x})$ is a vector whose only nonzero entries are the entries in $\mathbf{x}$ that are associated with class ℓ. That is, the test image $\mathbf{y}$ will be assigned to the class with the minimum class-wise reconstruction error. The idea of SRC is that the test image $\mathbf{y}$ can be best linearly reconstructed by the columns of $\mathbf{D}_{\ell^*}$ if it belongs to class ℓ^*. As a result, most non-zero elements of $\mathbf{x}$ will be associated with class ℓ^*, and $\left\|\mathbf{y} - \mathbf{D}\delta_{\ell^*}(\mathbf{x})\right\|_2$ gives the minimum reconstruction error.

A major assumption of SRC is that it requires the collection of a large amount of training data as the over-complete dictionary $\mathbf{D}$. Therefore, directly applying SRC to tackle undersampled face recognition will lead to degraded performance. To address this issue, Deng *et al.* [9] proposed Extended SRC (ESRC), which solves the following minimization problem:

$$\min_{\mathbf{x}} \left\|\mathbf{y} - [\mathbf{D},\ \mathbf{A}]\begin{bmatrix}\mathbf{x_d}\\ \mathbf{x_a}\end{bmatrix}\right\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{3}$$

where $\mathbf{x} = [\mathbf{x_d};\ \mathbf{x_a}]$. In the scenario of undersampled face recognition, each subject in $\mathbf{D}$ only has one or few images.

To be able to model all possible variations of interests, ESRC introduces the intra-class variant dictionary $\mathbf{A}$, which consists of image data collected from an external dataset (e.g., subjects not of interest). In a similar spirit of SRC, ESRC proposed the following classification criterion:

$$\ell^* = \arg\min_{\ell} \left\| \mathbf{y} - [\mathbf{D},\ \mathbf{A}] \begin{bmatrix} \delta_\ell(\mathbf{x_d}) \\ \mathbf{x_a} \end{bmatrix} \right\|_2. \quad (4)$$

We note that, compared to (2), the operator $\delta_\ell(\cdot)$ in (4) is only applied to $\mathbf{x_d}$ instead of the entire coefficient vector $\mathbf{x}$. This is because that $\mathbf{x_a}$ is not associated with any class label information. Although ESRC has shown promising results on undersampled face recognition, there are three concerns with ESRC. Firstly, ESRC directly apply external data as $\mathbf{A}$, which might be noisy or contain undesirable artifacts. Secondly, the computation of (3) would be very expensive due to the large size of $\mathbf{A}$. This is due to the fact that ESRC needs the matrix $\mathbf{A}$ for covering all intra-class variations of interest. Finally, ESRC regards occlusion as intra-class variations during the collection of $\mathbf{A}$ from external data. In other words, ESRC assumes the type of occlusion to be known when collecting external data, which might not be practical.

B. Dictionary Learning for Sparse Coding

Recent research on computer vision and image processing has shown that the learning of data or application-driven dictionaries outperforms approaches using predefined ones [16]. In general, the optimization algorithms for dictionary learning can be designed in an *unsupervised* or *supervised* manner. Unsupervised dictionary learning such as MOD [17] or KSVD [13] focuses on data representation, and is suitable for image synthesis tasks like image denoising.

Nevertheless, for addressing recognition tasks, one requires supervised dictionary learning strategies which aim at introducing improved discriminative capability for the observed learning model. Several approaches have been proposed by introducing different classification criteria to the objective function. For example, Ramirez *et al.* incorporated an incoherent term on dictionaries from different classes into the sparse representation based formulation [18]. Yang *et al.* added the Fisher discrimination term to the objective function such that the learned dictionaries would favor data classification [19]. Another common approach integrated classifier design into the sparse representation framework, so that both classifiers and dictionaries will be jointly learned for improved recognition [20]–[23].

The above dictionary learning approaches all require a sufficient amount of training data, and thus they will not generalize well for undersampled face recognition. Recent works [10], [11] address this issue via the learning of intra-class variations from external data. However, as ESRC discussed in the previous subsection, [10] also views occlusion as the intra-class variation. Consequently, [10] demands the information on occlusion of test images for learning intra-class dictionaries, which largely limits their applications in practice. In this paper, we also consider the learning of an auxiliary dictionary for modeling intra-class variations using external data, but unlike [10], our approach treat occlusion as the pixels that have large reconstruction errors. As a result, our learned intra-class dictionary does not depend on the occlusion information in test images. Another recent work [11] characterized occlusion as sparse errors when performing recognition. This approach does not require the prior knowledge of occlusion, but as noted in [12], such characterization might be imprecise and would be insufficient to represent the occlusion presented in real-world face images.

TABLE I
COMPARISONS OF RECENT SRC-BASED APPROACHES FOR FACE RECOGNITION

	Undersampled Gallery Set	Dictionary Learning	Robustness to Occlusion
SRC [14]	×	×	×
RSC [12]	×	×	√
LRSI [15]	×	×	√
ESRC [9]	√	×	×
ADL [10]	√	√	×
SVDL [11]	√	√	×
Ours	√	√	√

C. Remarks on SRC-Based Approaches for Face Recognition

We highlight and compare the properties of recent sparse representation based face recognition methods in Table I, in which SRC [14], ESRC [9], ADL [10], SVDL [11] have been discussed in previous two subsections. It is worth mentioning that Yang *et al.* [12] have proposed an iteratively reweighted sparse coding algorithm to improve SRC for better dealing with outliers such as occlusion or corruption. Another recent work [15] utilized low-rank matrix decomposition with structural incoherence to address the scenario where both training and test data can have occluded images. Both [12], [15] do not require the knowledge of occlusion in test images, but they need a sufficient amount of training data to cover image variants for all subjects of interest. Directly applying the methods of [12] and [15] to undersampled face recognition can lead to degraded recognition performance. Later in the experiments, we will confirm that our approach outperforms state-of-the-art SRC based methods.

III. OUR PROPOSED METHOD

A. Face Recognition via Robust Auxiliary Dictionary Learning

1) Our Classification Formulation: We now present our classification algorithm for undersampled face recognition via robust auxiliary dictionary learning, as shown in the upper part of Fig. 2. Let $\mathbf{y} \in \mathbb{R}^d$ be the query image and $\mathbf{D} \in \mathbb{R}^{d \times n}$ be the gallery matrix. The gallery matrix $\mathbf{D}$ is composed of data matrices from L classes, i.e. $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_L]$. The auxiliary dictionary $\mathbf{A} \in \mathbb{R}^{d \times m}$ is learned from external data, and the detailed algorithms for learning $\mathbf{A}$ will be discussed in Section III-B. Our goal is to determine the identity of the query input $\mathbf{y}$.

Although ESRC in (3) can be employed to classify $\mathbf{y}$, it assumes that the types of occlusion (or corruption) of the test

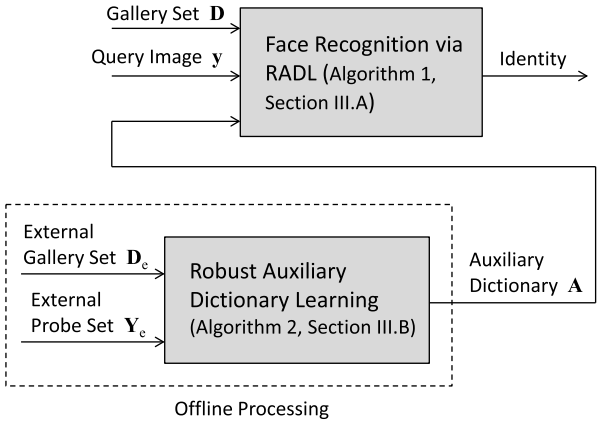


Fig. 2. Flowchart for our proposed framework for undersampled face recognition.

image $\mathbf{y}$ must be known and present in their pre-collected dictionary $\mathbf{A}$. It is obvious that this assumption might not be practical in real-world scenarios. To address this issue, instead of solving (3), we consider the following minimization problem:

$$\min_{\mathbf{x}} \boldsymbol{\rho}\left(\mathbf{y}-[\mathbf{D},\ \mathbf{A}]\begin{bmatrix}\mathbf{x_d}\\ \mathbf{x_a}\end{bmatrix}\right)+\lambda\|\mathbf{x}\|_1, \tag{5}$$

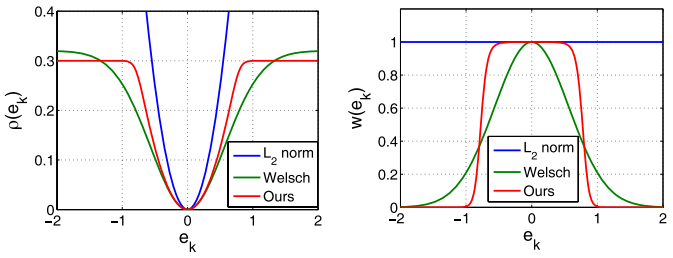
where $\mathbf{x}=[\mathbf{x_d};\ \mathbf{x_a}]$ is the sparse coefficient of $\mathbf{y}$, and the residual function $\boldsymbol{\rho}(\cdot)\colon \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{aligned}\boldsymbol{\rho}(\mathbf{e}) &= \sum_{k=1}^{d}\rho(e_k),\\ \rho(e_k) &= -\frac{1}{2\mu}\left(\ln\left(1+\exp(-\mu e_k^2+\mu\delta)\right)\right.\\ &\quad \left.-\ln(1+\exp\mu\delta)\right),\end{aligned} \tag{6}$$

where e_k is the kth entry of $\mathbf{e}=\mathbf{y}-[\mathbf{D},\mathbf{A}]\mathbf{x}$, and the parameters μ and δ will be detailed at the end of this subsection. In the theory of robust M-estimators [24], the residual function $\rho(\cdot)$ in (5) is designed to minimize the influence of outliers. Standard residual functions used in robust M-estimators include Huber, Cauchy, and the Welsch functions. We consider the residual function $\rho(\cdot)$ defined in (6), because this type of residual functions has shown promising results in recent literatures of robust face recognition [12], [25].

We note that, similar to the least-squares approach, ESRC utilizes the L2-norm in (3) as the residual function, which is known to be sensitive to outliers. This is because that the L2-norm grows quadratically as the absolute value of its input increases (see the blue curve in Fig. 3(a)). The red and green curves in Fig. 3(a) plot our residual function and the Welsch function, respectively. After deriving the solution of (5), we will discuss how the three residual functions in Fig. 3(a) affect face recognition.

2) Remarks on Robust Sparse Coding: It is worth mentioning that both robust sparse coding [12] and our formulation (5) aim at solving a non-convex optimization problem with L1-norm regularization. However, our approach to obtain the optimal solution is very different from the one used in [12].



(a) (b)

Fig. 3. (a) The residual function $\rho(\cdot)$ and (b) the corresponding weight function $w(\cdot)$.

In particular, [12] assumes that the objective function can be approximated by a first order Taylor expansion with a quadratic residual term. As a result, what RSC minimizes is an approximated version of the original objective function. On the other hand, our approach directly solves the optimization problem by the technique of variable substitution and the chain rule for calculating the derivatives (see Section III-A3 for detailed derivations). We note that the derivations of RSC and ours lead to similar algorithms that both iteratively solve a weighted sparse coding problem and update the weight matrix accordingly. However, our derivation guarantees the optimal solution, while the derivation of RSC might result in an approximated one. We note that RSC is extended from SRC, which requires a sufficient amount of training data (i.e., an over-complete dictionary) and thus is not able to handle undersampled recognition problems. Later in our experiments, the superiority of our approach over RSC can be verified.

3) Optimization: Next, we show how to derive the solution of (5). Taking the derivative of the objective function in (5) with respect to $\mathbf{x}$ leads to

$$\frac{d}{d\mathbf{x}}\left(\boldsymbol{\rho}(\mathbf{e})+\lambda\|\mathbf{x}\|_1\right)=\sum_{k=1}^{d}\frac{d}{d\mathbf{x}}\rho(e_k)+\lambda\partial\|\mathbf{x}\|_1, \tag{7}$$

where $\partial\|\mathbf{x}\|_1$ represents the derivative of $\|\mathbf{x}\|_1$. Using the chain rule of derivatives, (7) can be expressed as

$$\begin{aligned}&\sum_{k=1}^{d}\frac{d\rho(e_k)}{de_k}\frac{de_k}{d\mathbf{x}}+\lambda\partial\|\mathbf{x}\|_1\\ &=\frac{1}{2}\sum_{k=1}^{d}\frac{d\rho(e_k)}{de_k}\frac{1}{e_k}\frac{de_k^2}{d\mathbf{x}}+\lambda\partial\|\mathbf{x}\|_1\\ &=\frac{1}{2}\sum_{k=1}^{d}w(e_k)\frac{de_k^2}{d\mathbf{x}}+\lambda\partial\|\mathbf{x}\|_1,\end{aligned} \tag{8}$$

where

$$w(e_k)=\frac{d\rho(e_k)}{de_k}\frac{1}{e_k}=\frac{\exp(-\mu e_k^2+\mu\delta)}{1+\exp(-\mu e_k^2+\mu\delta)}. \tag{9}$$

If $w(e_k)$ in (8) is fixed as a constant, then (8) becomes the derivative of

$$\frac{1}{2}\sum_{k=1}^{d}w(e_k)e_k^2+\lambda\|\mathbf{x}\|_1=\frac{1}{2}\|\mathbf{We}\|_2^2+\lambda\|\mathbf{x}\|_1, \tag{10}$$

where $\mathbf{e} = \mathbf{y} - [\mathbf{D}, \mathbf{A}]\mathbf{x}$ and

$$\mathbf{W} = \text{diag}(w(e_1), w(e_2), \ldots, w(e_d))^{1/2}. \quad (11)$$

From the above derivation, we know that the solution of (5) can be calculated by repeatedly solving

$$\min_{\mathbf{x}} \left\| \mathbf{W} \left(\mathbf{y} - [\mathbf{D},\ \mathbf{A}] \begin{bmatrix} \mathbf{x_d} \\ \mathbf{x_a} \end{bmatrix} \right) \right\|_2^2 + \lambda \|\mathbf{x}\|_1, \quad (12)$$

and updating $\mathbf{W}$ according to (11), where e_k is the kth entry of $\mathbf{e}$. Notice that with $\mathbf{W}$ fixed, (12) is in the form of the standard L1-minimization problem, and one can apply existing techniques such as Homotopy, Iterative Shrinkage-Thresholding, or Alternating Direction Method for solving (12). In our work, we choose the Homotopy method because of its effectiveness and efficiency as suggested in [26].

We see from (10) that $w(e_k)$ is multiplied with e_k^2, and thus $w(e_k)$ can be viewed as the weight of e_k^2. We plot the weight function $w(\cdot)$ corresponding to different residual functions $\rho(\cdot)$ in Fig. 3(b). It can be seen from the figure that the weight function of L2-norm is a constant function, while our weight function outputs a smaller value for large $|e_k|$. This property makes our $w(\cdot)$ able to detect occlusion from the test image, since occlusion often leads to large reconstruction errors than ordinary pixels do. Although the Welsch function in Fig. 3(b) also possesses this property, it is more sensitive to the magnitude of e_k than our weight function is. When e_k slightly deviates from zero, the output of the Welsch function quickly drops, while the output of our weight function remains unchanged.

After obtaining the optimal solution of (5), denoted by $\mathbf{x}^*$, and its corresponding weight matrix $\mathbf{W}^*$, we propose the following classification rule to classify $\mathbf{y}$:

$$\ell^* = \arg\min_{\ell} \left\| \mathbf{W}^* \left(\mathbf{y} - [\mathbf{D},\ \mathbf{A}] \begin{bmatrix} \delta_\ell(\mathbf{x}_\mathbf{d}^*) \\ \mathbf{x}_\mathbf{a}^* \end{bmatrix} \right) \right\|_2, \quad (13)$$

where $\mathbf{x}^* = [\mathbf{x}_\mathbf{d}^*;\ \mathbf{x}_\mathbf{a}^*]$. Namely, we assign $\mathbf{y}$ to the class with the smallest reconstruct error. Different from the classification rule of ESRC in (4), the weight matrix $\mathbf{W}^*$ in (13) lowers the influence of pixels that are poorly reconstructed. As a result, (13) achieves better recognition performance than ESRC, especially when the test image $\mathbf{y}$ is occluded or corrupted. Algorithm 1 summarizes our algorithm for classifying $\mathbf{y}$.

4) Parameter Selection: Next, we discuss how to choose the parameters μ and δ for the weight function in (9). The goal is to select μ and δ such that the output of the weight function in (9) is similar to the red curve in Fig. 3(b). Notice that when e_k approaches zero, we have $w(e_k) \approx \exp(\mu\delta)/(1+\exp(\mu\delta))$. If the product $\mu\delta$ is large enough, then $w(e_k) \approx \exp(\mu\delta)/\exp(\mu\delta) = 1$. To this end, we let $\mu\delta = C_{\mu\delta}$, where $C_{\mu\delta}$ is a constant whose value is equal to or larger than 8. Next, we show how to determine δ. Notice that $w(e_k)$ in (9) can be expressed as

$$w(e_k) = \frac{\exp(\mu(\delta - e_k^2))}{1 + \exp(\mu(\delta - e_k^2))},$$

and thus $w(e_k) = 1/2$ when $\delta = e_k^2$. That is, δ determines when the output of $w(e_k)$ will pass through $1/2$. To decide the value of δ, we sort the vector $[e_1^2, e_2^2, \ldots, e_d^2]$ in descending order and denote the sorted vector by $\mathbf{e}_s$. We let δ be the jth largest element of $\mathbf{e}_s$, where j is the nearest integer to τd with $\tau \in [0.6, 0.8]$ and $d = \text{length}(\mathbf{e}_s)$. Once δ is obtained, μ can be readily calculated as $\mu = C_{\mu\delta}/\delta$. This mechanism for adjusting μ and δ has been utilized in [12] and [25].

Algorithm 1 Undersampled Face Recognition via RADL

Input: Training data $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_L]$ from L subjects, intra-class dictionary $\mathbf{A}$, and the test input $\mathbf{y}$

Step 0: Normalize $\mathbf{y}$ and the columns of $\mathbf{D}$ to have unit ℓ_2-norm

Step 1: Initialize $\mathbf{W} = \mathbf{I}$

Step 2: Calculate the optimal solution of (5), denoted by $\mathbf{x}^*$, and the associated weight matrix $\mathbf{W}^*$

while not converged **do**

$\mathbf{x} \leftarrow \arg\min_{\mathbf{x}} \|\mathbf{W}(\mathbf{y} - [\mathbf{D},\ \mathbf{A}]\mathbf{x})\|_2^2 + \lambda\|\mathbf{x}\|_1$

$\mathbf{e} \leftarrow \mathbf{y} - [\mathbf{D}, \mathbf{A}]\mathbf{x}$

$\mathbf{W} \leftarrow \text{diag}(w(e_1), \ldots, w(e_d))^{1/2}$ with $w(\cdot)$ defined in (9)

end while

Step 3: Classify $\mathbf{y}$ via weighted reconstruction errors

$$\ell^* = \arg\min_{\ell \in \{1,2,\ldots,L\}} \left\| \mathbf{W}^* \left(\mathbf{y} - [\mathbf{D},\ \mathbf{A}] \begin{bmatrix} \delta_\ell(\mathbf{x}_\mathbf{d}^*) \\ \mathbf{x}_\mathbf{a}^* \end{bmatrix} \right) \right\|_2$$

Output: identity$(\mathbf{y}) \leftarrow \ell^*$

B. Robust Auxiliary Dictionary Learning (RADL)

1) Our Proposed Algorithm for RADL: In Section III-A, we present an ESRC-based algorithm for undersampled face recognition, with the introduced residual function can be applied to identify and disregard corrupted image regions due to occlusion. We now discuss how we learn the auxiliary dictionary $\mathbf{A}$ in (5) for properly handling intra-class variants of interest. Inspired by [9]–[11], we utilize images collected from external data to learn the auxiliary dictionary. More specifically, our objective function integrates dictionary learning and the classification rule of (13) for improved and robust recognition performance.

We now detail our proposed algorithm for RADL, which is depicted in the lower part of Fig. 2. Suppose the external dataset contains p subjects. We partition these external images into a probe set $\mathbf{Y}_e$ and a gallery set $\mathbf{D}_e$ (note that the subscript e indicates external data). The probe matrix $\mathbf{Y}_e = [\mathbf{y}_e^1, \mathbf{y}_e^2, \ldots, \mathbf{y}_e^N] \in \mathbb{R}^{d \times N}$ consists of N images in $\mathbb{R}^d$ with different intra-class variations to be modeled. The gallery matrix $\mathbf{D}_e \in \mathbb{R}^{d \times rp}$ contains only one or few face images per subject, where r is the number of images per subject. If each subject in the gallery set only has one face image, then $\mathbf{D}_e \in \mathbb{R}^{d \times p}$. To learn an auxiliary dictionary for modeling intra-class image variants, we propose to solve the following minimization problem during training:

$$\min_{\mathbf{A},\mathbf{X}} \sum_{i=1}^{N} \boldsymbol{\rho}\left(\mathbf{y}_e^i - [\mathbf{D}_e, \mathbf{A}] \begin{bmatrix} \mathbf{x}_\mathbf{d}^i \\ \mathbf{x}_\mathbf{a}^i \end{bmatrix} \right) + \lambda \left\| \mathbf{x}^i \right\|_1 + \eta\, \boldsymbol{\rho}(\mathbf{y}_e^i - \mathbf{D}_e \delta_{i\ell}(\mathbf{x}_\mathbf{d}^i) - \mathbf{A}\mathbf{x}_\mathbf{a}^i), \quad (14)$$

where the auxiliary dictionary $\mathbf{A} \in \mathbb{R}^{d\times m}$ is to be learned (m specifies the number of dictionary atoms), and the function $\boldsymbol{\rho}(\cdot)$ is defined as in (6). The vector $\mathbf{x}^i = [\mathbf{x}^i_{\mathbf{d}};\ \mathbf{x}^i_{\mathbf{a}}]$ is the sparse coefficient of $\mathbf{y}^i_e$, in which $\mathbf{x}^i_{\mathbf{d}} \in \mathbb{R}^{p\times 1}$ and $\mathbf{x}^i_{\mathbf{a}} \in \mathbb{R}^{m\times 1}$ indicate the coefficients associated with the gallery $\mathbf{D}_e$ and the auxiliary dictionary $\mathbf{A}$, respectively. We denote by $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N] \in \mathbb{R}^{(m+p)\times N}$ the sparse coefficient matrix for $\mathbf{Y}_e$. The function $\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}})$ outputs a vector whose only nonzero entries are the entries in $\mathbf{x}^i_{\mathbf{d}}$ that are associated with class i_ℓ (i_ℓ denotes the label of $\mathbf{y}^i_e$ in the external data set). Parameters λ and η control the weights of the sparsity and the class-wise reconstruction error, respectively.

In (14), the first term indicates data representation, the second term introduces the sparsity constraint, while the last term $\boldsymbol{\rho}(\mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \mathbf{A}\mathbf{x}^i_{\mathbf{a}})$ is the reconstruction error for class i_ℓ. Notice that our classification rule in (13) assigns the test image to the class with the minimum reconstruction error. Since the label of $\mathbf{y}^i_e$ is i_ℓ, we introduce the last term in (14) to minimize the reconstruction error for class i_ℓ. This explains how we effectively integrate both robust auxiliary dictionary learning and classification into a unified framework.

2) Optimization for RADL: We now provide optimization details for solving (14) during training. The objective function in (14) is nonlinear with respect to variables $\mathbf{X}$ and $\mathbf{A}$. To solve (14), we employ the alternating direction method [27], which iterates between the stages of sparse coding and dictionary update for obtaining the optimal solution of (14).

a) Sparse coding for updating X: In the sparse coding stage, we fix $\mathbf{A}$ and optimize (14) with respective to $\mathbf{X}$, which is equivalent to solving the following problem:

$$\min_{\mathbf{x}^i} \boldsymbol{\rho}\big(\mathbf{y}^i_e - [\mathbf{D}_e, \mathbf{A}]\mathbf{x}^i\big) + \lambda\big\|\mathbf{x}^i\big\|_1 + \eta\,\boldsymbol{\rho}(\mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \mathbf{A}\mathbf{x}^i_{\mathbf{a}}) \tag{15}$$

for $i = 1, 2, \ldots, N$. Following similar steps as in Section III-A3, we can obtain the solution of (15) by iteratively solving

$$\min_{\mathbf{x}^i} \big\|\mathbf{W}_g(\mathbf{y}^i_e - [\mathbf{D}_e, \mathbf{A}]\mathbf{x}^i)\big\|_2^2 + \lambda\big\|\mathbf{x}^i\big\|_1 + \eta\,\big\|\mathbf{W}_c(\mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \mathbf{A}\mathbf{x}^i_{\mathbf{a}})\big\|_2^2 \tag{16}$$

with

$$\mathbf{W}_g = \mathrm{diag}(w(g_1), w(g_2), \ldots, w(g_d))^{1/2},$$
$$\mathbf{W}_c = \mathrm{diag}(w(c_1), w(c_2), \ldots, w(c_d))^{1/2}, \tag{17}$$

where $w(\cdot)$ is defined as in (9), and g_k and c_k are the kth entries of

$$\mathbf{g} = \mathbf{y}^i_e - [\mathbf{D}_e, \mathbf{A}]\mathbf{x}^i,$$
$$\mathbf{c} = \mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \mathbf{A}\mathbf{x}^i_{\mathbf{a}}, \tag{18}$$

respectively. Notice that (16) can be written as the following L1 minimization problem:

$$\min_{\mathbf{x}^i} \left\|\begin{bmatrix}\mathbf{W}_g\mathbf{y}^i_e \\ \gamma\,\mathbf{W}_c\mathbf{y}^i_e\end{bmatrix} - \begin{bmatrix}\mathbf{W}_g\mathbf{D}_e & \mathbf{W}_g\mathbf{A} \\ \gamma\,\mathbf{W}_c\delta_{i_\ell}(\mathbf{D}_e) & \gamma\,\mathbf{W}_c\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{x}^i_{\mathbf{d}} \\ \mathbf{x}^i_{\mathbf{a}}\end{bmatrix}\right\|_2^2 + \lambda\left\|\mathbf{x}^i\right\|_1, \tag{19}$$

where $\gamma = \eta^{1/2}$ and $\delta_{i_\ell}(\mathbf{D}_e) \in \mathbb{R}^{d\times p}$ whose only nonzero columns are those columns of $\mathbf{D}_e$ that are associated with class i_ℓ. Hence, one can utilize existing techniques mentioned in Section III-A3 to solve (19).

b) Dictionary update for A: In the dictionary update stage, we fix $\mathbf{X}$ and optimize (14) with respect to $\mathbf{A}$, whose solution can be obtained by solving the following problem:

$$\min_{\boldsymbol{\alpha}^j} \sum_{i=1}^{N} \boldsymbol{\rho}(\mathbf{y}^i_e - [\mathbf{D}_e, \mathbf{A}]\mathbf{x}^i) + \eta\,\boldsymbol{\rho}(\mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \mathbf{A}\mathbf{x}^i_{\mathbf{a}}) \tag{20}$$

for $j = 1, 2, \ldots, m$, where $\boldsymbol{\alpha}^j$ is the jth column of $\mathbf{A}$, i.e. $\mathbf{A} = [\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \ldots, \boldsymbol{\alpha}^m]$. Following similar steps as in Section III-A3, we calculate the solution of (20) by iteratively solving

$$\min_{\boldsymbol{\alpha}^j} \sum_{i=1}^{N} \big\|\mathbf{W}_g(\mathbf{y}^i_e - [\mathbf{D}_e, \mathbf{A}]\mathbf{x}^i)\big\|_2^2 + \eta\,\big\|\mathbf{W}_c(\mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \mathbf{A}\mathbf{x}^i_{\mathbf{a}})\big\|_2^2, \tag{21}$$

where $\mathbf{W}_g$ and $\mathbf{W}_c$ are defined as in (17). Once the solution of (21), denoted by $\boldsymbol{\alpha}^{j*}$, is obtained, the jth column of $\mathbf{A}$ is updated as

$$\mathbf{A}(:, j) = \boldsymbol{\alpha}^{j*}. \tag{22}$$

Repeating the above process for $j = 1, 2, \ldots, m$, we finish the update of $\mathbf{A}$.

Next, we show how to derive the solution of (21). The objective function of (21) can be written as

$$\sum_{i=1}^{N} \left\|\boldsymbol{\Phi}^i_g - \tilde{\mathbf{W}}^i_g\boldsymbol{\alpha}^j\right\|_2^2 + \eta\left\|\boldsymbol{\Phi}^i_c - \tilde{\mathbf{W}}^i_c\boldsymbol{\alpha}^j\right\|_2^2, \tag{23}$$

where

$$\boldsymbol{\Phi}^i_g = \mathbf{W}_g\left(\mathbf{y}^i_e - \mathbf{D}_e\mathbf{x}^i_{\mathbf{d}} - \sum_{k\neq j}^{m}\boldsymbol{\alpha}^k x^i_{\mathbf{a},k}\right),$$
$$\boldsymbol{\Phi}^i_c = \mathbf{W}_c\left(\mathbf{y}^i_e - \mathbf{D}_e\delta_{i_\ell}(\mathbf{x}^i_{\mathbf{d}}) - \sum_{k\neq j}^{m}\boldsymbol{\alpha}^k x^i_{\mathbf{a},k}\right), \tag{24}$$

and

$$\tilde{\mathbf{W}}^i_g = x^i_{\mathbf{a},j}\mathbf{W}_g, \quad \tilde{\mathbf{W}}^i_c = x^i_{\mathbf{a},j}\mathbf{W}_c, \tag{25}$$

where $x^i_{\mathbf{a},j}$ is the jth entry of $\mathbf{x}^i_{\mathbf{a}}$. Since (23) is a quadratic function of $\boldsymbol{\alpha}^j$, the solution of (21) can be obtained by setting the partial derivative of (23) with respect to $\boldsymbol{\alpha}^j$ equal to zero, i.e.,

$$2\sum_{i=1}^{N}\Big((\tilde{\mathbf{W}}^i_g)^T\boldsymbol{\Phi}^i_g - (\tilde{\mathbf{W}}^i_g)^T\tilde{\mathbf{W}}^i_g\boldsymbol{\alpha}^j + \eta(\tilde{\mathbf{W}}^i_c)^T\boldsymbol{\Phi}^i_c - \eta(\tilde{\mathbf{W}}^i_c)^T\tilde{\mathbf{W}}^i_c\boldsymbol{\alpha}^j\Big) = 0. \tag{26}$$

Fig. 4. Example images of the Extended Yale B database.

In view of (26), the optimal solution of (21) is

$$\boldsymbol{\alpha}^{j*} = \left(\sum_{i=1}^{N}(\tilde{\mathbf{W}}_g^i)^T\tilde{\mathbf{W}}_g^i + \eta(\tilde{\mathbf{W}}_c^i)^T\tilde{\mathbf{W}}_c^i\right)^{-1} \times\left(\sum_{i=1}^{N}(\tilde{\mathbf{W}}_g^i)^T\boldsymbol{\Phi}_g^i + \eta(\tilde{\mathbf{W}}_c^i)^T\boldsymbol{\Phi}_c^i\right). \quad (27)$$

We summarize our algorithm for learning the auxiliary dictionary in Algorithm 2. Note that the coefficient matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where $\mathbf{x}_i$ is initially set as

$$\mathbf{x}_i = [\overbrace{1, 1, \ldots, 1}^{p}, \overbrace{0, 0, \ldots, 0}^{m}]^T$$

for $i = 1, 2, \ldots, N$. On the other hand, we apply the settings of ESRC for initializing the auxiliary dictionary $\mathbf{A}$ (as detailed later in Section IV).

IV. Experimental Results

A. Extended Yale B Database

First, we consider the Extended Yale B database [28] for our experiments. This database contains 38 subjects with about 64 frontal face images for each (see example images in Fig. 4), and the face images are taken under various illumination conditions [29]. All images are converted into grayscale and are downsampled to 34×30 pixels prior to our experiments. We select 32 subjects from the database to be recognized, and the remaining 6 subjects are considered as external data (i.e., subjects not of interest) for robust auxiliary dictionary learning.

For the 32 subjects of interest, we select 3 images from each of the 32 subjects as the gallery $\mathbf{D}$, and the remaining 61 images for testing. The three gallery images correspond to the three illumination conditions: A+000E+00, A-085E+20, and A+085E+20 (A+085 refers to 85 degrees azimuth, and E+20 refers to 20 degrees elevation [28]). For the training stage of robust auxiliary dictionary learning using external data (i.e., the six subjects *not* of interest), we choose the same images corresponding to A+000E+00, A-085E+20, and A+085E+20 as the gallery $\mathbf{D}_e$, and thus $\mathbf{D}_e$ contains a total of 6×3 images. The probe $\mathbf{Y}_e$ consists of the random selection of 29 images from the remaining images of these 6 subjects. We will vary the number of dictionary atoms m and evaluate the performance of our approach.

For comparison purposes, we consider several SRC-based approaches: SRC [14], RSC [12], ESRC [9], and ADL [10]. To construct the auxiliary dictionary of ESRC, first we follow the procedure in [9] to build an intra-class variant dictionary $\mathbf{A}_0$ from an external dataset. Then, we randomly select the columns of $\mathbf{A}_0$ to form the auxiliary dictionary $\mathbf{A}$ with the desired number of columns m. Throughout our experiments, we let m be a multiple of r, where r is the number of images per subject. Note that when randomly selecting the columns of $\mathbf{A}_0$, we choose r images of the same subject at a time. We also test our method without dictionary learning, which is denoted by `Ours w/o DL`, i.e., we use Algorithm 1 as the classification method with $\mathbf{A}$ derived from ESRC instead of from Algorithm 2. For our RADL, we utilize the auxiliary dictionary of ESRC as the initial value of $\mathbf{A}$ in Algorithm 2. For this and all subsequent experiments, the parameter λ in (5) is set as 10^{-4}, and the parameters λ and η in (15) are set as 10^{-4} and 1, respectively. By varying the number of atoms m of the auxiliary dictionary $\mathbf{A}$, we show the performance comparisons in Fig. 5. We have $m = 0$ for

Algorithm 2 Robust Auxiliary Dictionary Learning

Input: The gallery matrix $\mathbf{D}_e \in \mathbb{R}^{d\times p}$ and the probe $\mathbf{Y}_e \in \mathbb{R}^{d\times N}$

Step 0: Normalize the columns of $\mathbf{D}_e$ and $\mathbf{Y}_e$ to have unit ℓ_2-norm

Step 1: Initialize $\mathbf{X} \in \mathbb{R}^{(p+m)\times N}$ and $\mathbf{A} \in \mathbb{R}^{d\times m}$

Step 2: Calculate the optimal solution of (14)

while not converged **do**

  Sparse Coding Stage: update $\mathbf{X}$

  for $i = 1 : N$ **do**

    Calculate $\mathbf{W}_g$ and $\mathbf{W}_c$ by (17) with $\mathbf{g}$ and $\mathbf{c}$ in (18)

    Obtain $\mathbf{x}^i$ via solving (19)

  end for

  Dictionary Update Stage: update $\mathbf{A}$

  for $j = 1 : m$ **do**

    for $i = 1 : N$ **do**

      Calculate $\boldsymbol{\Phi}_g^i$, $\boldsymbol{\Phi}_c^i$ in (24) and $\tilde{\mathbf{W}}_g^i$, $\tilde{\mathbf{W}}_c^i$ in (25)

    end for

    Obtain $\boldsymbol{\alpha}^j$ via solving (27)

    Update the jth column of $\mathbf{A}$, i.e. $\mathbf{A}(:, j) = \boldsymbol{\alpha}^j$

  end for

end while

Output: Auxiliary dictionary $\mathbf{A}$

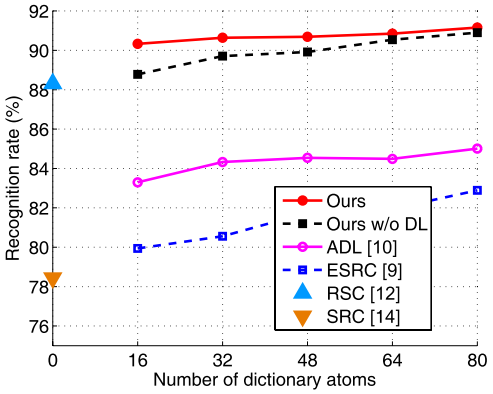


Fig. 5. Performance comparisons on the Extended Yale B database with different numbers of dictionary atoms in $\mathbf{A}$.

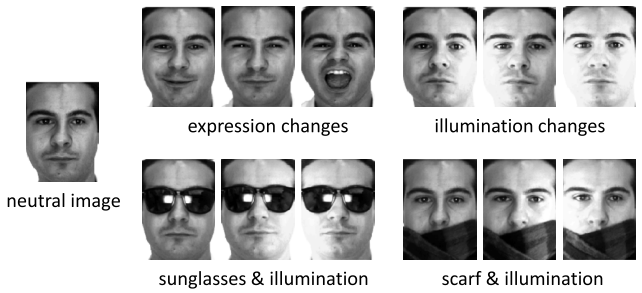


Fig. 6. Example images of the AR database. Note that only the neural image of each subject is included in the gallery set, while the rest are viewed as query images to be recognized.

SRC and RSC, since they do not consider any external data. It is worth noting that, if no external data is available, methods of ESRC and ADL are equivalent to SRC, and our method turns into RSC.

Note that the Extended Yale B dataset contains some face images taken at extreme illumination conditions. Hence, it is likely that some pixels of these images have large residuals, which can result in inaccurate classification results. Our methods assign small weights to pixels that lead to large residuals, and thus better recognition performance can be expected. As shown in Fig. 5, our method clearly outperformed other SRC-based (with and without learning) approaches when different numbers of auxiliary dictionary atoms were considered. In the following parts of our experiments, we consider more challenging databases which contain not only face images with illumination and expressions variations, but also the occluded ones for recognition.

B. AR Database

1) Face Recognition and RADL With the Same Domain: The AR database [30] consists of over 4,000 frontal face images of 126 individuals. The images are taken under different variations, including illumination, expression, and facial occlusion/disguise in two separate sessions. For each session there are thirteen images, in which three images are with sunglasses, another three are with scarfs, and the remaining seven are with illumination and expressions variations. In our experiments, we consider a subset of AR consisting of 50 men and 50 women. All images are converted to grayscale and cropped to 165×120 pixels. We select 80 subjects of interest for training and testing, and the remaining 20 subjects are considered as external data for robust auxiliary dictionary learning.

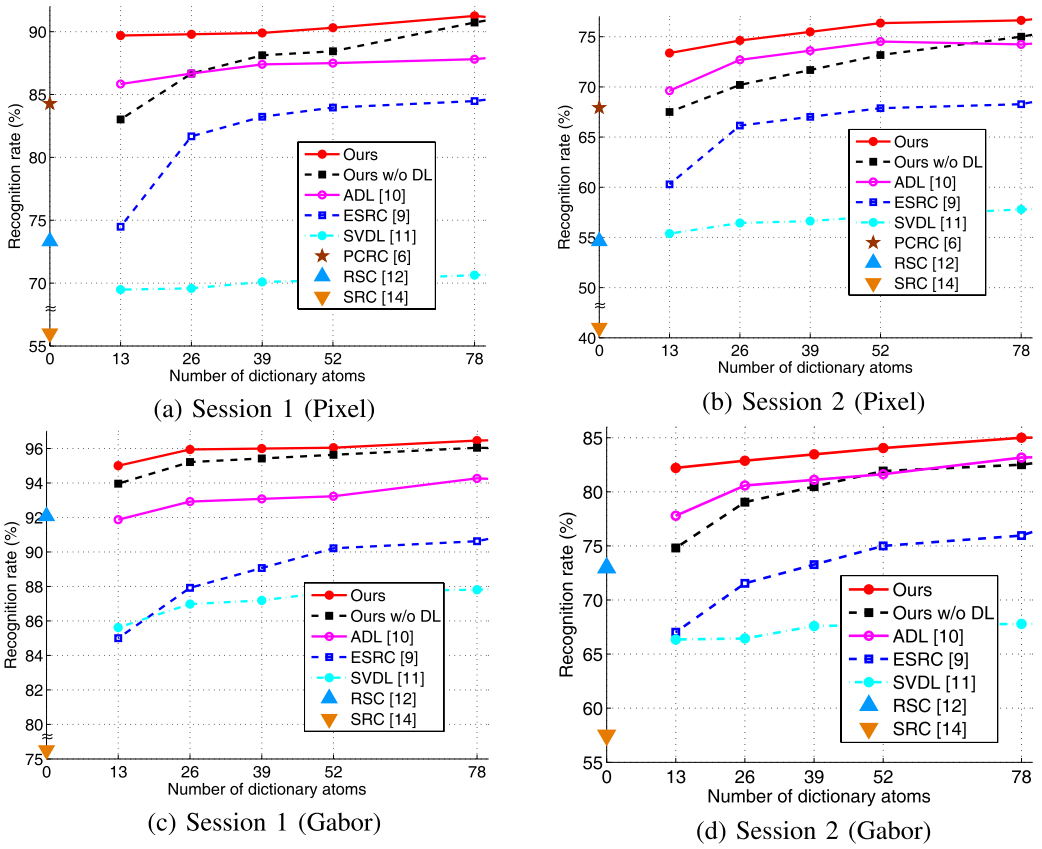
For the scenario of undersampled face recognition, we choose only the *neutral* image of each of the 80 subjects (40 men and 40 women) in Session 1 as the gallery, and the rest images in Sessions 1 and 2 are for testing, see Fig. 6 for example. It is worth noting that the setting for the AR database is more challenging than that of the Extended Yale B database. We not only have to deal with image variants of illumination, expression, and occlusion, but we also require only one face image for each person as the gallery for recognition. To learn the auxiliary dictionary $\mathbf{A}$ from the external data, we form the gallery matrix $\mathbf{D}_e$ by calculating the mean of non-occluded images of each individual, and thus $\mathbf{D}_e$ contains a total of 20 images. The probe matrix $\mathbf{Y}_e$ consists of all images from the above 20 external subjects, i.e., $\mathbf{Y}_e$ includes a total of 520 images.

We consider several recent approaches (using pixel-based or Gabor features) for comparisons: SRC [14], RSC [12], PCRC [6], ESRC [9], ADL [10], and SVDL [11]. Same as in the previous subsection, the random sampling technique is applied for ESRC and `Our w/o DL` to obtain the auxiliary dictionary. The pixel-based feature vector is obtained by downsampling the original image to 38×28 pixels. The Gabor feature vector of length 2,304 is derived by evaluating the Gabor kernel at three scales and four orientations (see [31] for more detailed information). By varying the number of atoms m of the auxiliary dictionary $\mathbf{A}$, we show the performance comparisons in Fig. 7. The gallery matrix $\mathbf{D}$ is collected from Session 1, while the query image $\mathbf{y}$ can be chosen from Sessions 1 or 2, which corresponds to the left and the right columns of Fig. 7, respectively. As a result, the scenario of the right column is more challenging than that of the left column. It can be seen from Fig. 7 that our method outperformed other SRC-based approaches across different features and sessions.

While AGL [7] has also been applied to solve undersampled face recognition problems, it is not particularly designed to recognize face images with occlusion. In addition, it requires a sufficient amount of external data for handling image variants (i.e., within-class variations). As a result, if applying the same setting as those in Fig. 7(d), AGL would achieve a lower recognition rate of 60.58%. We note that, as shown in Fig. 7, recognition performances of ESRC and `Ours w/o DL` degraded remarkably when the number of dictionary atoms became small. On the other hand, dictionary learning based methods like ADL and ours did not suffer from this problem. This illustrates the importance of the learning of dictionary atoms for obtaining satisfactory recognition performance when a compact dictionary is required. We note that it is expected that the difference between our method with and without DL would become smaller as the number of dictionary atoms increases. This is because the use of more external data can give comparable performance as learning-based approaches do (but is more expensive in terms of both computation and storage costs).

Finally, we compare the performance of ESRC [9], KSVD [13], ADL [10], and ours over a range of feature dimensions. For KSVD, we directly apply its algorithm to the gallery $\mathbf{D}_e$ for learning the auxiliary dictionary $\mathbf{A}$ with $m = 13$, and use our Algorithm 1 as the classification method. We plot the performance comparisons using pixel-based features with $m = 13$ in Fig. 8. Note that KSVD only considers the representation ability of dictionaries, while our formulation (14) incorporates the classification rule into the objective function. It can be observed from Fig. 8 that our approach clearly outperformed KSVD and others, which supports the use of our method even when lower feature dimensions are of interest.

2) Face Recognition and RADL With Different Domains: In the previous experiments, the external dataset for building auxiliary dictionaries is a disjoint subset of the same database



(a) Session 1 (Pixel) (b) Session 2 (Pixel)

(c) Session 1 (Gabor) (d) Session 2 (Gabor)

Fig. 7. Performance comparisons on the AR database with different numbers of dictionary atoms in $\mathbf{A}$ using pixel-based (shown in (a) and (b)) and Gabor features (shown in (c) and (d)). For each row, the left and right figures present the recognition rates using query images from Sessions 1 and 2, respectively.

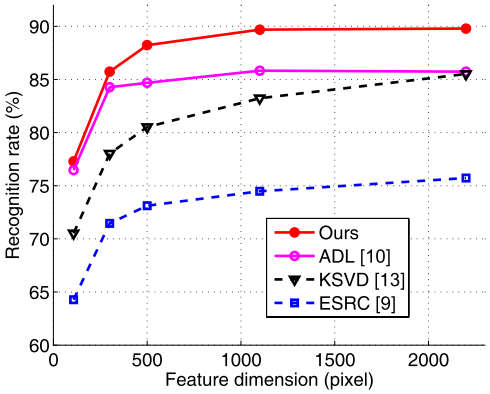


Fig. 8. Performance comparisons on the AR database with different pixel-based feature dimensions.

(i.e., images of the same dataset but different from those for training and testing). To evaluate the generalization ability of our approach, we conduct a new experiment on the AR database with auxiliary dictionaries learned from a subset of the Multi-PIE database [32]. We randomly choose 20 subjects from the Multi-PIE database, and each of the subject has 20 frontal face images. Note that the subset of Multi-PIE only includes illumination changes, while the subset of AR contains intra-class variations due to illumination, expression, and occlusion. Using Multi-PIE for learning auxiliary dictionaries makes the recognition problem more challenging. The experimental setting for training and testing is the same as that in Section IV-B1. The number of dictionary atoms is set as 26, and Gabor filters are used to extract the image features.

We compare our methods with recent SRC-based approaches: SRC [14], RSC [12], ESRC [9], ADL [10], and SVDL [11]. Table II lists and compares the recognition results, in which the first row indicates the session number of test data (training data is from Session 1), and the second row indicates the subsets for learning auxiliary dictionaries. It can be seen that, since SRC requires an overcomplete dictionary for handling occluded test inputs (i.e., an oversampled instead of undersampled setting), it was not able to achieve satisfactory performance. As for RSC, while it well recognized test images of Session 1, its recognition performance degraded rapidly (about 19%) when the test images were from a different session (i.e., Session 2). From Table II, we see that the recognition rates of ESRC and ADL degraded remarkably when the external data was selected from Multi-PIE instead of AR. This is because that both ESRC and ADL directly applied external data as (or for learning) the auxiliary dictionary to model intra-class variations, including occlusion. If such data does not contain the information about occlusion (such as Multi-PIE), ESRC and ADL will not be able to achieve satisfactory recognition performance. In contrast, our method does not suffer from this problem. Our approach not only performs dictionary learning for dealing with image variants, it is also able to identify occluded pixels with large reconstruction errors as outliers.

TABLE II

PERFORMANCE COMPARISONS ON THE AR DATABASE. NOTE THAT THE GALLERY SET IS FROM SESSION 1, WHILE THE PROBE IMAGES ARE FROM SESSIONS 1 OR 2. FOR EACH EXPERIMENT, THE AUXILIARY DICTIONARY IS LEARNED FROM AR OR MULTI-PIE DATABASES. NOTE THAT $*$ INDICATES THE METHODS WITHOUT USING ANY EXTERNAL DATA

Methods	Session 1			Session 2		
	AR	Multi-PIE		AR	Multi-PIE	
SRC* [14]	75.31	75.31		57.50	57.50	
RSC* [12]	92.08	92.08		72.98	72.98	
ESRC [9]	87.92	77.60	↓ 10.32	71.54	59.23	↓ 12.31
ADL [10]	92.92	80.31	↓ 12.61	80.58	61.83	↓ 18.75
SVDL [11]	86.98	83.44	↓ 3.54	66.44	63.08	↓ 3.36
Ours w/o DL	95.21	93.54	↓ 1.67	79.04	76.92	↓ 2.12
Ours	**95.94**	**94.00**	↓ 1.94	**82.88**	**80.10**	↓ 2.78

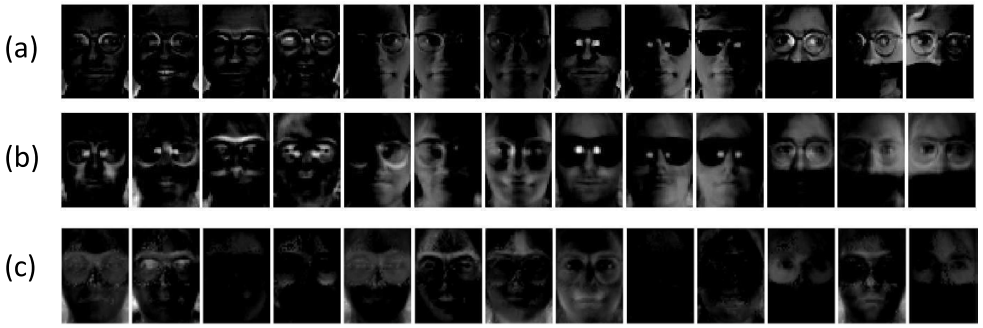


Fig. 9. The auxiliary dictionaries learned or selected from a subset of the AR database by (a) ESRC [9] (b) ADL [10], and (c) our method.

Fig. 9 shows the auxiliary dictionaries learned or selected from the AR subset by ESRC, ADL, and our method. From this figure, we see that the auxiliary dictionaries of ESRC and ADL include the intra-class variations due to sunglasses and scarfs, while ours does not depend on the occlusion presented in the AR subset.

We note that our experimental setup is actually different from that of SVDL [11]. In [11], SVDL consistently outperformed ESRC while the external data did not contain any corrupted images. In our work, we consider the cases when the external dataset is with or without occluded data. Take Fig. 7 for example, we have training/test and external data from the same AR database, and occluded images (i.e., those with sunglasses and scarves) are presented in both test and external datasets. ESRC performed favorably against SVDL in this experiment, since ESRC directly applied such external data in which the image variants exhibit exactly the same types of image corruption. On the other hand, we have additional experiments shown in Table II where we take AR or Multi-PIE as external datasets (while training/test images are from AR). We see that, when applying images of Multi-PIE as external data, ESRC was not able to handle occluded test images as expected. This is due to its direct use of non-occluded images as image variants. In this case, SVDL still achieved improved performance than the ESRC did (e.g., 83.44% vs. 77.6%, and 63.08% vs. 59.23%). Therefore, our results and observations are still consistent with those reported in [11].

It is worth repeating that, SVDL [11] characterizes occlusion as sparse errors during classification, which could also recognize occluded test images without the prior knowledge of occlusion. However, as indicated in [12], such characterization might not be sufficiently accurate, and thus would be difficult to describe real-world occluded face images. From Table II, we see that the recognition performance of SVDL was inferior to ours in both sessions. It is of practical interest to know whether our approach can generalize well to the case, in which the auxiliary dictionary is learned across different datasets. From Table II, we see that our method achieved the best generalization ability among all the methods considered, and thus the robustness of our approach can be successfully verified.

C. CAS-PEAL Database

Finally, we consider the CAS-PEAL database [33]. This database contains 1,040 individuals with variations including facing direction, expression, accessory, lighting, time, background, and distance. Every subject is captured under at least two kinds of these variations. To the best of our knowledge, CAS-PEAL is currently the largest public face database with corrupted face images available. We utilize all 434 subjects from the Normal and the Accessory categories of CAS-PEAL (recall that AR only has face images of 100 subjects). Thus, each subject has 1 neutral image, 3 images with hats, and 3 images with glasses/sunglasses. We select 374 subjects of interest for training and testing, and the remaining 60 subjects are considered as external data for robust auxiliary dictionary learning.

In our experiments, we choose only the *neutral* image of each of the 374 subjects as the gallery, and the rest images for testing, see Fig. 10 for example. To learn the auxiliary dictionary $\mathbf{A}$ from the external data, we choose the neutral image of every subject in the external data to form the gallery matrix $\mathbf{D}_e$, and thus $\mathbf{D}_e$ contains a total of 60 images.

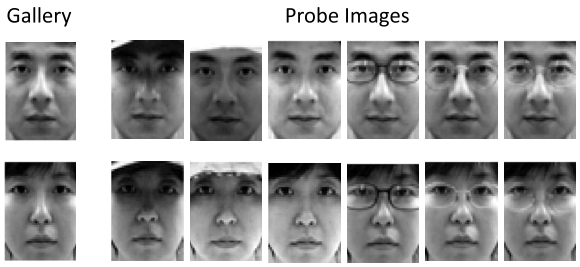


Fig. 10. Example images of the CAS-PEAL database.

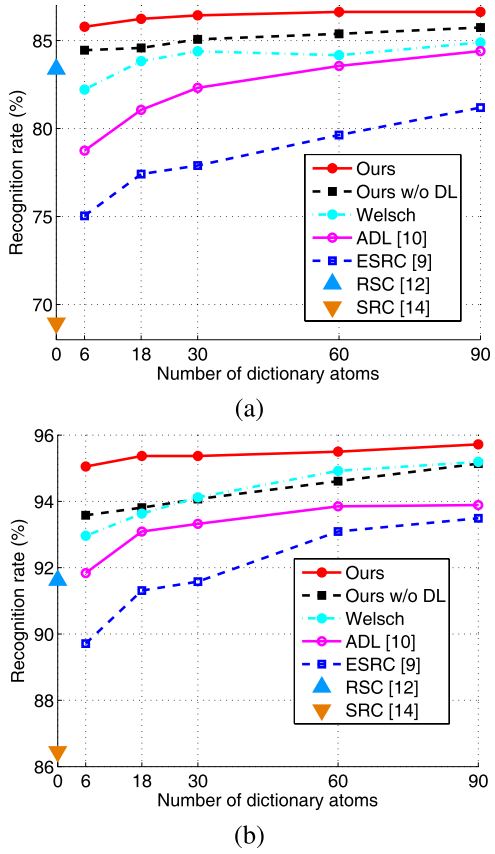


Fig. 11. Performance comparisons on the CAS-PEAL database with different numbers of dictionary atoms in $\mathbf{A}$ using (a) pixel-based and (b) Gabor features.

The probe matrix $\mathbf{Y}_e$ consists of the remaining images from the above 60 external subjects, i.e., $\mathbf{Y}_e$ includes a total of 360 images.

Similarly, using pixel-based or Gabor features, we consider several recent SRC-based approaches for comparisons: SRC [14], RSC [12], ESRC [9], and ADL [10]. We also compare our weight function with the Welsch function, i.e., replace the weight functions in Algorithms 1 and 2 with $w(e_k) = \exp(-(e_k/c)^2)$, and denote this face recognition method by Welsh (the parameter c of the Welsch function is adjusted to achieve the best recognition performance). The pixel-based feature vector is obtained by downsampling the original image to 35×28 pixels. The other settings are the same as those in the previous subsections. By varying the number of atoms m of the auxiliary dictionary $\mathbf{A}$, we show the performance comparisons in Fig. 11. It can be seen that our method outperformed other baseline and

TABLE III
PERFORMANCE COMPARISONS ON CAS-PEAL WITH AUXILIARY DICTIONARIES LEARNED FROM ITS ACCESSORY AND EXPRESSION SUBSETS (I.E., WITH AND WITHOUT OCCLUDED IMAGES). THE GALLERY SET CONTAINS ONLY THE NEUTRAL IMAGE OF EACH SUBJECT, WHILE THE QUERY IMAGES ARE FROM THE ACCESSORY CATEGORY. NOTE THAT $*$ INDICATES METHODS WITHOUT USING ANY EXTERNAL DATA

Methods	Accessory	Expression	
SRC* [14]	86.45	86.45	
RSC* [12]	91.62	91.62	
ESRC [9]	89.71	86.63	↓ 3.08
ADL [10]	91.84	87.21	↓ 4.63
SVDL [11]	93.98	91.58	↓ 2.40
Ours w/o DL	93.58	92.91	↓ 0.67
Ours	**95.05**	**93.63**	↓ 1.42

state-of-the-art approaches. Therefore, we conclude that a joint optimization framework which considers both auxiliary dictionary learning and classification (like ours) would be preferable for addressing undersampled face recognition problems.

Next, we provide additional experiments (with the same Gabor features), in which the external data are selected from either a subset of its Expression category or from a subset of the Accessory category (from 60 subjects not of interest). The Expression category includes non-occluded images with only expression changes, while the Accessory category contains occluded images due to hats or glasses. As a result, if the Expression category is considered, the external data will consist of 5 images with different facial expressions for each subject not of interest; if the Accessory category is used, each subject in the external dataset will consist of 3 images with hats and 3 images with glasses. The gallery and probe sets are the same as those used in Fig. 11, and the number of dictionary atoms is set as 6.

With the above experimental setting, we compare our methods with recent SRC-based approaches: SRC [14], RSC [12], ESRC [9], ADL [10], and SVDL [11]. Table III lists and compares the recognition results. From this table, we see that our method was able to achieve comparable results while the performances of ESRC, ADL and SVDL degraded when the external dataset was changed from Accessory to Expression. In other words, even with no occlusion information observed in external data, our method still performs favorably against recent SRC and dictionary learning based approaches.

Finally, we present two examples recognition results in Fig. 12. For both examples shown in this figure, our method successfully determined the correct identity for the query input while other SRC-based approaches failed. We note that, the query image in the first example is with a pair of sunglasses, which is viewed as occlusion. For the methods of ADL and ESRC, they simply selected the subjects wearing similar glasses for recognition. Note that the weighting matrix derived by RSC contained more extreme errors (dark pixels) than ours did. This is because that RSC does not have

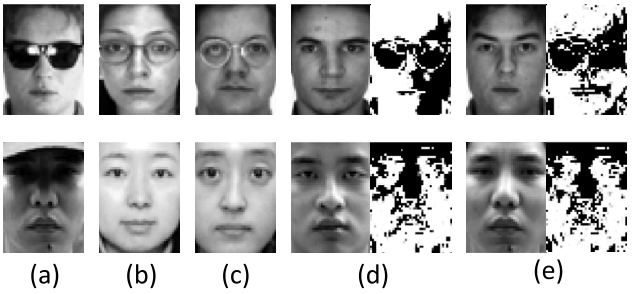


Fig. 12. Example results of AR (top row) and CAS-PEAL (bottom row). The query images are shown in (a), while the subjects identified by ADL, ESRC, RSC, and ours are in (b) to (e), respectively. Note that the weighting matrixes of RSC and ours are also illustrated.

a mechanism to model intra-class variations. On the other hand, the query image in the second example is with a hat, which also results in occluded image regions. Although both weighting matrices of RSC and ours were very similar to each other (i.e., the hat regions were successfully treated as outliers), RSC did not correctly identify the query, which again is due to its lack of ability in handling intra-class variations. With the introduction of robust auxiliary dictionary learning, our method overcame the aforementioned problems and achieved improved recognition. From the above experiments, the effectiveness and robustness of our proposed algorithm can be successfully verified.

D. Multi-PIE Database

As noted in previous subsections, the use of external data is to cover intra-class variations such as expression and illumination changes. Therefore, it would be best to have the collected external dataset contain the same image variants as the training and test ones do. However, in practical scenarios, one cannot expect the type of image variants to be fixed or known in advance. Therefore, our strategy is to select external data which exhibit intra-class variations of interest, so that the image variants of the test set can be covered via linearly approximation/reconstruction.

To verify our claim, we now conduct experiments on Multi-PIE with external data collected from AR or Multi-PIE. For training and testing, 80 subjects from Multi-PIE are selected to be recognized, while a different set of 20 subjects from AR or Multi-PIE are chosen for learning the auxiliary dictionary. For the training set, we choose the image with neutral expression captured by camera `05_1` of each of the 80 subjects. On the other hand, 13 images with a neutral expression under different illumination conditions of each of the 80 subjects are selected for testing. The external dataset from Multi-PIE contains 14 face images with illumination variations for each of the 20 subjects. If the external dataset is from AR, then there are two scenarios to be considered. In the first scenario, denoted by AR_{14}, the external dataset consists of 14 face images (7 images from Session 1, and 7 images from Session 2) with expression and illumination variations for each subject not of interest. The second scenario, denoted by AR_8, is similar to the first scenario except that each subject has 8 face images with only expression variations.

TABLE IV
RECOGNITION PERFORMANCE ON MULTI-PIE USING DIFFERENT EXTERNAL DATASETS. EXTERNAL DATA MULTI-PIE_{14} DENOTES A SUBSET OF MULTI-PIE, WHICH CONTAINS 14 IMAGES WITH ILLUMINATION VARIATIONS FOR EACH SUBJECT NOT OF INTEREST. AR_{14} DENOTES A SUBSET OF AR, WHICH CONTAINS 14 IMAGES WITH EXPRESSION AND ILLUMINATION VARIATIONS. FINALLY, AR_8 CONTAINS ONLY IMAGES WITH EXPRESSION VARIATIONS

External data	None	AR_8	AR_{14}	Multi-PIE_{14}
Accuracy	80.19	82.88	91.73	94.42

Table IV lists and compares the recognition rates using different external datasets with our method RADL. The baseline method is RSC (i.e., the first entry of Table IV, denoted as *None*), which does not utilize any external data. Recall that each subject in the test set has 13 images with illumination variations. To achieve satisfactory recognition performance, the external data should contain sufficient information about illumination variations. As expected, the use of Multi-PIE_{14} as external data lead to the best recognition rate, since both training/test and external data were from the same dataset. Compared to Multi-PIE_{14}, the recognition rate of AR_{14} was slightly dropped by 2.69%, while the recognition rate of AR_8 decreased 11.54%. This is because that AR_8 only contained image variants of expression changes and thus failed to cover illumination variations presented in the test data. The above experimental results verify that one should properly select image variants as external data for performance guarantees.

V. CONCLUSION

We presented a novel learning-based algorithm for undersampled face recognition. We advocated the learning of an auxiliary dictionary from external data for modeling intra-class image variants of interest, and utilized a residual function in a joint optimization formulation for identifying and disregarding corrupted image regions due to occlusion. As a result, the proposed algorithm allows one to recognize occluded face images, or those with illumination and expressions variations, even only one or few gallery images per subject are available during training. Experimental results on four different face image datasets confirmed the effectiveness and robustness of our method, which was shown to outperform state-of-the-art sparse representation and dictionary learning based approaches with or without using external face data.

REFERENCES

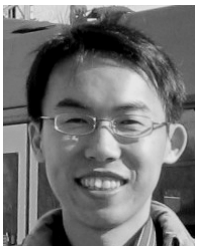
[1] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, Sep. 2006.

[2] T. Ahonen, A. Hadid, and M. Pietikainen, "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.

[3] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2617–2628, Oct. 2007.

[4] 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.

[5] R. Kumar, A. Banerjee, B. C. Vemuri, and H. Pfister, "Maximizing all margins: Pushing face recognition with kernel plurality," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2375–2382.
[6] P. Zhu, L. Zhang, Q. Hu, and S. C. K. Shiu, "Multi-scale patch based collaborative representation for face recognition with margin distribution optimization," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 822–835.
[7] Y. Su, S. Shan, X. Chen, and W. Gao, "Adaptive generic learning for face recognition from a single sample per person," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2699–2706.
[8] M. Kan, S. Shan, Y. Su, D. Xu, and X. Chen, "Adaptive discriminant learning for face recognition," *Pattern Recognit.*, vol. 46, no. 9, pp. 2497–2509, Sep. 2013.
[9] 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.
[10] C.-P. Wei and Y.-C. F. Wang, "Learning auxiliary dictionaries for undersampled face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.
[11] M. Yang, L. Van Gool, and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 689–696.
[12] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 625–632.
[13] 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.
[14] 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.
[15] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2618–2625.
[16] I. Tošić and P. Frossard, "Dictionary learning: What is the right representation for my signal?" *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[17] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1999, pp. 2443–2446.
[18] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3501–3508.
[19] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 543–550.
[20] D.-S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[21] J. Mairal, F. R. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 1033–1040.
[22] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2691–2698.
[23] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1697–1704.
[24] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *Int. J. Comput. Vis.*, vol. 19, no. 1, pp. 57–91, Jul. 1996.
[25] 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.
[26] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, "Fast ℓ_1-minimization algorithms and an application in robust face recognition: A review," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1849–1852.
[27] J. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.
[28] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[29] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
[30] A. Martínez and R. Benavente, "The AR face database," Centre Visió Comput., Univ. Autònoma Barcelona, Bellaterra, Barcelona, Spain, Tech. Rep. 24, 1998.
[31] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 448–461.
[32] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.
[33] 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.



Chia-Po Wei received the B.S. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2004 and 2011, respectively. He is currently a Post-Doctoral Researcher with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include face recognition, dictionary learning, and computer vision.

Yu-Chiang Frank Wang (M'04) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, USA, in 2004 and 2009, respectively.

He joined the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan, in 2009. He is currently a Tenure-Track Associate Research Fellow, and leads the Multimedia and Machine Learning Laboratory with CITI. His research interests span the fields of computer vision, pattern recognition, and machine learning. He and his team received the First Place Award at Taiwan Tech Trek by the National Science Council (NSC) of Taiwan in 2011. In 2013, he was selected as the Outstanding Young Researcher by NSC.