Sparse Variation Dictionary Learning for Face Recognition with A Single Training Sample Per Person

Meng Yang, Luc Van Gool
ETH Zurich
Switzerland
{yang,vangool}@vision.ee.ethz.ch

Lei Zhang
The Hong Kong Polytechnic University
Hong Kong
cslzhang@comp.polyu.edu.hk

Abstract

Face recognition (FR) with a single training sample per person (STSPP) is a very challenging problem due to the lack of information to predict the variations in the query sample. Sparse representation based classification has shown interesting results in robust FR; however, its performance will deteriorate much for FR with STSPP. To address this issue, in this paper we learn a sparse variation dictionary from a generic training set to improve the query sample representation by STSPP. Instead of learning from the generic training set independently w.r.t. the gallery set, the proposed sparse variation dictionary learning (SVDL) method is adaptive to the gallery set by jointly learning a projection to connect the generic training set with the gallery set. The learnt sparse variation dictionary can be easily integrated into the framework of sparse representation based classification so that various variations in face images, including illumination, expression, occlusion, pose, etc., can be better handled. Experiments on the large-scale CMU Multi-PIE, FRGC and LFW databases demonstrate the promising performance of SVDL on FR with STSPP.

1. Introduction

As one of the most visible applications in computer vision, face recognition (FR) has been receiving significant attention in the community [30]. In the past decade, researchers have been devoting themselves to addressing the various problems emerging in practical FR scenarios such as face identification/verification in uncontrolled or less controlled environment [7][26][9][28]. In many practical applications of FR (e.g., law enforcement, e-passport, driver license, etc.), we can only have a single training face image per person. This makes the problem of FR particularly hard since there is very limited information we can use to predict the variations in the query sample. How to achieve robust FR performance in the scenario of single training sample per person (STSPP) is an important yet one of the most challenging problems in FR.

The number of training samples per person will greatly affect the performance of FR [22]. In the case of STSPP, many discriminant subspace and manifold learning algorithms (e.g., LDA and its variants [2]) cannot be directly applied. The recently developed representation based FR methods such as sparse representation based classification (SRC) [27] cannot be easily applied to STSPP, either, since SRC needs multiple training samples per person to reasonably represent the query face. To address the problem of STSPP, many specially designed FR methods have been developed [22]. According to the availability of an additional generic training set, the FR methods for STSPP can be classified into two categories: methods without using a generic training set, and methods with generic learning.

The STSPP methods without generic learning often extract robust local features (e.g., gradient orientation [23] and local binary pattern [20]), generate additional virtual training samples (e.g., via singular value decomposition [29], geometric transform and photometric changes [18]), or perform image partitioning (e.g., local patch based LDA [3], self-organizing maps of local patches [21], and multi-manifold learning from local patches [11]). Although these methods have led to improved FR results, local feature extraction and discriminative learning from local patches can be sensitive to image variations (e.g., extreme illumination and expression), while the new information introduced by virtual training sample generation can be rather limited.

Considering the fact that face variations for different subjects share much similarity, an additional generic training set with multiple samples per person could bring new and useful information (e.g., generic intra-class variation) to the STSPP gallery set. Therefore, a generic training set can be employed to extract discriminant information for FR with STSPP[14][25][8]. For instance, the expression-invariant subspace and pose-invariant subspace were learned from a collected generic training set to solve the expression-invariant [14] and pose-invariant [8] FR problems, respec-

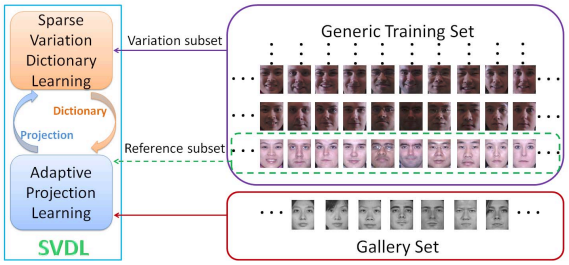

Figure 1. Flowchart of the proposed sparse variation dictionary learning (SVDL) method for face recognition with a single training sample per person.

tively. Deng *et al.* [5] extended the SRC method to FR with STSPP. The so-called Extended SRC (ESRC) computes the intra-class variation from a generic training set and then uses the generic variation matrix to code the difference between the query and gallery samples.

Although much improvement has been reported, several critical issues remain in ESRC and other generic training based methods for FR with STSPP. First, the generic intra-class variation may not be similar to that of gallery subjects, so the extraction of discrimination information from the generic training set may not be guaranteed. Second, the learned variation matrix can be very big and redundant since many subjects in the generic training set are involved. This will increase the computational burden of the final FR algorithm. Third, the learned variation matrix cannot represent the unknown occlusion in query images due to the randomness of location and intensity of occluded pixels.

To solve the above problems, we propose to learn a compact dictionary with powerful variation representation ability, jointly with an adaptive projection from the generic training set to the gallery set. Dictionary learning has been extensively studied in image processing and computer vision [16][12][1]. To the best of our knowledge, however, the dictionary learning for pattern classification tasks is mostly conducted on the gallery set with multiple samples per class. If we apply some dictionary learning method (e.g., K-SVD [1]) to the generic training set to learn a dictionary for variation representation, it would ignore the correlation between the generic training set and the gallery set. Although the correlation between generic training set and gallery set has been studied in subspace learning [19], how to learn a gallery-set adaptive dictionary to exploit the variation in the generic training set is a new problem.

With the above considerations, in this paper we propose a novel sparse variation dictionary learning (SVDL) method for FR with STSPP. As shown in Fig.1, the proposed SVDL is a joint learning framework of adaptive projection and a sparse variation dictionary. By extracting from the generic training set a reference subset and a variation subset, the adaptive projection learning aims to exploit the correlation between the reference subset and the gallery set, while the variation dictionary learning aims to learn a compact dictionary with sparse bases from a big variation matrix, which is the projection of the intra-class variation of generic training set over the learned projection matrix. Compared with previous methods [5][14][25][8][19], the proposed joint learning of adaptive projection and sparse variation dictionary more effectively exploits the information of gallery set and generic training set. Extensive experiments on large-scale face databases with various variations, including illumination, expression, pose, session, occlusion and blur, show that the proposed SVDL achieves state-of-the-art performance for FR with STSPP.

2. Sparse Variation Dictionary Learning

2.1. Representation of face images with variation

Suppose that we are given a sufficiently large generic training set. Each subject in the generic training set has multiple face images, each with one type of variation (e.g., illumination, expression, pose). The number of subjects should be large enough to enhance generalization. Since the high-dimensional face images usually lie on a lower-dimensional subspace or sub-manifold, a sample in the gallery set, denoted by $\boldsymbol{g}$, could be represented as $\boldsymbol{g} = \boldsymbol{R}\boldsymbol{\gamma}$, where $\boldsymbol{R}$ is a subset of the generic training set and $\boldsymbol{\gamma}$ is the representation coefficient of $\boldsymbol{g}$ over $\boldsymbol{R}$. Each column of $\boldsymbol{R}$ is a vectorized training sample of a generic subject, and we assume that the subjects in $\boldsymbol{R}$ have similar facial variations as $\boldsymbol{g}$ (e.g., illumination, expression, pose, etc.; random corruption and occlusion are not considered in the generic training set.)

Let $\boldsymbol{g}_{(v)}$ denote a sample which has the same identity as $\boldsymbol{g}$ but has some variations of illumination, expression, or pose w.r.t. $\boldsymbol{g}$, where subscript v indicates the type of variation. Similarly, we could represent $\boldsymbol{g}_{(v)}$ as

$$\boldsymbol{g}_{(v)} = \boldsymbol{R}_{(v)}\boldsymbol{\gamma}_{(v)} \tag{1}$$

where $\boldsymbol{R}_{(v)}$ is the counterpart of $\boldsymbol{R}$ with variation type v. Since $\boldsymbol{g}_{(v)}$ and $\boldsymbol{R}_{(v)}$ show similar variations to $\boldsymbol{g}$ and $\boldsymbol{R}$, the representation coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_{(v)}$ should also be similar:

$$\boldsymbol{\gamma}_{(v)} \approx \boldsymbol{\gamma} \tag{2}$$

Eq. (2) is actually based on the fact that people with similar normal frontal appearance should also have similar appearance in other variations. This assumption has been successfully used in illumination-invariant, expression-invariant and pose-invariant FR [4][14][10], and speech animation [13], respectively.

A testing sample from the subject associated with $\boldsymbol{g}$, denoted by $\boldsymbol{y}$, could be well represented as a linear combination of the samples from this subject:

$$\boldsymbol{y} = \boldsymbol{g}\beta_g + \sum\nolimits_{(v)} \boldsymbol{g}_{(v)}\beta_{(v)} + \boldsymbol{e} \tag{3}$$

where β_g and $\beta_{(v)}$ are the representation coefficients, and $\boldsymbol{e}$ is a sparse vector of representation residual.

Suppose that there are n types of variations available in the generic training set, and we denote by

$$\boldsymbol{X}_{(v)} = \boldsymbol{R}_{(v)} - \boldsymbol{R}, \quad v = 1, \cdots, n \tag{4}$$

the n variation matrices obtained from the generic training set. The face variation representation model in Eq. (3) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{g}\beta'_g + \sum\nolimits_{(v)} \beta_{(v)} \boldsymbol{X}_{(v)}\boldsymbol{\gamma} + \boldsymbol{e} \tag{5}$$

where $\beta'_g = \beta_g + \sum_{(v)} \beta_{(v)}$.

A reference subset $\boldsymbol{R}_i$ for a gallery individual $\boldsymbol{g}_i$ can typically be extracted through group sparse coding [24]: $\boldsymbol{R}_i = \boldsymbol{R}_{(\hat{v})}$, where $\hat{v} = \arg\min_v ||\boldsymbol{g}_i - \boldsymbol{R}_{(v)}\hat{\boldsymbol{\gamma}}_{(v)}||_2$ and $\hat{\boldsymbol{\gamma}}_{(v)} = \arg\min_{\boldsymbol{\gamma}_{(v)}} \sum_{(v)} ||\boldsymbol{\gamma}_{(v)}||_2$ s.t. $\boldsymbol{g}_i \approx \sum_{(v)} \boldsymbol{R}_{(v)}\boldsymbol{\gamma}_{(v)}$.

2.2. Sparse variation dictionary learning model

For the i^{th} gallery subject, $\boldsymbol{g}_i, i = 1, \cdots, c$, we thus can extract from the generic training set two subsets (see the end of previous Section). One is the *reference subset*, denoted by $\boldsymbol{R}_i$. The other is the *variation subset*, denoted by $\boldsymbol{X}_i = \left[\boldsymbol{X}_{i,(1)}, \cdots, \boldsymbol{X}_{i,(v)}, \cdots, \boldsymbol{X}_{i,(n)}\right]$, where $\boldsymbol{X}_{i,(v)}$ is the v^{th} generic variation matrix of the i^{th} gallery subject (refer to Eq. (4)). Different subject i usually has the same $\boldsymbol{R}_i$ and $\boldsymbol{X}_i$ since the extraction of $\boldsymbol{R}_i$ only depends on the variation type of $\boldsymbol{g}_i$ and the gallery images have limited facial variations. In order to learn a compact variation dictionary adaptively associated to the gallery set, we propose the following sparse variation dictionary learning (SVDL) model:

$$\min_{\boldsymbol{D}} \sum\nolimits_{i=1}^{c} \left\{p\left(\boldsymbol{g}_i, \boldsymbol{R}_i, \boldsymbol{\gamma}_i\right) + q\left(\boldsymbol{D}, \boldsymbol{X}_i, \boldsymbol{\gamma}_i\right)\right\} \tag{6}$$

where $\boldsymbol{D}$ is the dictionary to be learned, $\boldsymbol{\gamma}_i$ is the coding vector of $\boldsymbol{g}_i$ on $\boldsymbol{R}_i$, $p\left(\boldsymbol{g}_i, \boldsymbol{R}_i, \boldsymbol{\gamma}_i\right)$ is the adaptive projection learning term, $q\left(\boldsymbol{D}, \boldsymbol{X}_i, \boldsymbol{\gamma}_i\right)$ is the variation dictionary learning term. With projection $\boldsymbol{\gamma}_i$, the variation matrix $\boldsymbol{X}_i$ could be projected onto the dictionary to obtain the gallery set's variation matrix. Therefore, the projection serves as a bridge to connect the generic training dataset with the gallery set so that the learned dictionary is adaptive to, instead of being independent of, the gallery set. Next lets discuss the design of $p\left(\boldsymbol{g}_i, \boldsymbol{R}_i, \boldsymbol{\gamma}_i\right)$ and $q\left(\boldsymbol{D}, \boldsymbol{X}_i, \boldsymbol{\gamma}_i\right)$.

2.2.1 Adaptive projection term

From the analysis in Section 2.1, we know that the projected variation of the i^{th} subject on the gallery set, denoted by $\boldsymbol{Y}_i$, could be approximately represented as $\boldsymbol{Y}_i = \boldsymbol{X}_i \odot \boldsymbol{\gamma}_i$, where the operator $\odot$ is defined as:

$$\boldsymbol{X}_i \odot \boldsymbol{\gamma}_i = \left[\boldsymbol{X}_{i,(1)}\boldsymbol{\gamma}_i, \cdots, \boldsymbol{X}_{i,(v)}\boldsymbol{\gamma}_i, \cdots, \boldsymbol{X}_{i,(n)}\boldsymbol{\gamma}_i\right] \tag{7}$$

The adaptive projection learning term could be designed as:

$$p\left(\boldsymbol{g}_i, \boldsymbol{R}_i, \boldsymbol{\gamma}_i\right) = \left\|\boldsymbol{g}_i - \boldsymbol{R}_i\boldsymbol{\gamma}_i\right\|_F^2 + \lambda_1 \left\|\boldsymbol{\gamma}_i\right\|_F^2 \tag{8}$$

where λ_1 is a scalar constant. Here we use the l_2-norm to regularize the projection coefficients $\boldsymbol{\gamma}_i$ since no subject in $\boldsymbol{R}_i$ will have the same identity as $\boldsymbol{g}_i$, and thus more subjects in $\boldsymbol{R}_i$ should be involved to represent $\boldsymbol{g}_i$.

2.2.2 Variation dictionary learning term

The whole variation matrix of all subjects in the gallery set can be simply set as the concatenation of all $\boldsymbol{Y}_i, i = 1, \cdots, c$. However, such a variation matrix for the whole gallery set will have a very big size and is very redundant. Intuitively, we could learn a much more compact variation dictionary $\boldsymbol{D} = [\boldsymbol{d}_1, , \boldsymbol{d}_j, \cdots, \boldsymbol{d}_m]$ from all $\boldsymbol{Y}_i$, where $\boldsymbol{d}_j$ is a basis (or atom) in the dictionary $\boldsymbol{D}$.

We propose the following learning term:

$$\begin{aligned} q\left(\boldsymbol{D}, \boldsymbol{X}_i, \boldsymbol{\gamma}_i\right) = &\left\|\boldsymbol{Y}_i - \boldsymbol{D}\boldsymbol{B}_i\right\|_F^2 + \lambda_2 \|\boldsymbol{B}_i\|_1 + \lambda_3 \textstyle\sum_j \|\boldsymbol{d}_j\|_1 \\ &\text{s.t.} \quad \boldsymbol{Y}_i = \boldsymbol{X}_i \odot \boldsymbol{\gamma}_i; \quad \|\boldsymbol{d}_j\|_2 = 1 \end{aligned} \tag{9}$$

where $\boldsymbol{B}_i = \left[\boldsymbol{\beta}_{i,(1)}, \cdots, \boldsymbol{\beta}_{i,(v)}, \cdots, \boldsymbol{\beta}_{i,(n)}\right]$ and $\boldsymbol{\beta}_{i,(v)}$ is the coding vector of $\boldsymbol{X}_{i,(v)}\boldsymbol{\gamma}_i$ over $\boldsymbol{D}$, λ_2 and λ_3 are constants to balance the representation fidelity term, the sparse coefficient term, and the sparse dictionary basis term. We let $\|\boldsymbol{d}_j\|_2 = 1$ to avoid that $\boldsymbol{D}$ has arbitrarily large l_2-norm, resulting in trivial values of the coding coefficients in $\boldsymbol{B}_i$. Apart from the l_1-norm sparsity imposed on the coefficients $\boldsymbol{B}_i$, we also impose the l_1-norm sparsity on the variation dictionary atoms $\boldsymbol{d}_j$. This is useful because face variations (e.g., expression, pose, session difference, and some illumination) usually lead to sparse changes of face images.

2.2.3 The overall SVDL model

By integrating Eqs. (8) and (9) into Eq. (6), we have the following SVDL model:

$$\begin{aligned} \min_{\boldsymbol{D}, \boldsymbol{\gamma}_i, \boldsymbol{B}_i} \sum_{i=1}^{c} &\left\{ \begin{array}{l} \left\|\boldsymbol{g}_i - \boldsymbol{R}_i\boldsymbol{\gamma}_i\right\|_F^2 + \left\|\boldsymbol{X}_i \odot \boldsymbol{\gamma}_i - \boldsymbol{D}\boldsymbol{B}_i\right\|_F^2 \\ +\lambda_1 \left\|\boldsymbol{\gamma}_i\right\|_F^2 + \lambda_2 \|\boldsymbol{B}_i\|_1 + \lambda_3 \textstyle\sum_j \|\boldsymbol{d}_j\|_1 \end{array} \right\} \\ &\text{s.t.} \quad \|\boldsymbol{d}_j\|_2 = 1 \end{aligned} \tag{10}$$

The objective function in Eq. (10) is not jointly convex to $(\boldsymbol{D}, \boldsymbol{\gamma}_i, \boldsymbol{B}_i)$. Therefore, we solve this problem by breaking it into several sub-problems, and alternatively solving these unknown variables. Detailed optimization procedures are presented next in Section 3.

3. Optimization of SVDL

The minimization of SVDL in Eq. (10) can be divided into two sub-problems: adaptive projection learning by fixing $\boldsymbol{D}$ and $\boldsymbol{B}_i$, and sparse variation dictionary learning by

fixing $\boldsymbol{\gamma}_i$. Before the alternative optimization, we should first initialize $\boldsymbol{\gamma}_i$. By minimizing Eq. (8), we have

$$\hat{\boldsymbol{\gamma}}_i = \left(\boldsymbol{R}_i^T \boldsymbol{R}_i + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{R}_i^T \boldsymbol{g}_i \tag{11}$$

We take the above value as the initialization of $\boldsymbol{\gamma}_i$.

3.1. The update of sparse variation dictionary

For the convenience of expression, we let $\boldsymbol{Y} = [\boldsymbol{X}_1 \odot \boldsymbol{\gamma}_1, \boldsymbol{X}_2 \odot \boldsymbol{\gamma}_2, \cdots, \boldsymbol{X}_c \odot \boldsymbol{\gamma}_c]$ and $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \cdots, \boldsymbol{B}_c]$. When $\boldsymbol{\gamma}_i, i = 1, \cdots, c$, are fixed, the SVDL model in Eq. (10) is reduced to the following sparse dictionary learning problem:

$$\min_{\boldsymbol{D},\boldsymbol{B}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{B}\|_F^2 + \lambda_2 \|\boldsymbol{B}\|_1 + \lambda_3 \sum\nolimits_j \|\boldsymbol{d}_j\|_1 \text{s.t. } \|\boldsymbol{d}_j\|_2 = 1 \tag{12}$$

The minimization in Eq. (12) could be solved by alternatively solving $\boldsymbol{B}$ and $\boldsymbol{D}$. When $\boldsymbol{D}$ is fixed, the solving of $\boldsymbol{B}$ is a standard sparse coding problem, which could be easily solved by algorithms such as [27]. However, when $\boldsymbol{B}$ is fixed, the update of sparse dictionary $\boldsymbol{D}$ needs a little more effort. We update the variation dictionary atom by atom, as described below.

We rewrite $\boldsymbol{B}$ as $\boldsymbol{B} = [\boldsymbol{b}_1; \cdots; \boldsymbol{b}_j; \cdots; \boldsymbol{b}_m]$, where $\boldsymbol{b}_j$ is the j^{th} row of $\boldsymbol{B}$ and m is the number of dictionary atoms. Let $\boldsymbol{Z} = \boldsymbol{Y} - \sum_{j \neq k} \boldsymbol{d}_j \boldsymbol{b}_j$. By fixing all the other atoms $\boldsymbol{d}_j, j \neq k$, the updating of $\boldsymbol{d}_k$ could be rewritten as

$$\min_{\boldsymbol{d}_k} \|\boldsymbol{Z} - \boldsymbol{d}_k \boldsymbol{b}_k\|_F^2 + \lambda_3 \|\boldsymbol{d}_k\|_1 \quad \text{s.t. } \|\boldsymbol{d}_k\|_2 = 1 \tag{13}$$

Based on *Lemma* 1 of [17], Eq. (13) could be rewritten as

$$\min_{\boldsymbol{d}_k} \left\| \boldsymbol{Z}\boldsymbol{b}_k^T \big/ l^2 - \boldsymbol{d}_k \right\|_F^2 + \frac{\lambda_3}{\sqrt{l}} \|\boldsymbol{d}_k\|_1 \text{ s.t. } \|\boldsymbol{d}_k\|_2 = 1 \tag{14}$$

where $l = \|\boldsymbol{d}_k\|_2$. If l is very close to 0 (e.g., $< 1e-6$), the atom $\boldsymbol{d}_k$ could be removed from the dictionary since it is useless to represent $\boldsymbol{Z}$; otherwise, $\boldsymbol{d}_k$ could be updated as

$$\boldsymbol{d}_k = T_{\frac{\lambda_3}{2\sqrt{l}}} \left(\boldsymbol{Z}\boldsymbol{b}_k^T \big/ l^2 \right) \Big/ \left\| T_{\frac{\lambda_3}{2\sqrt{l}}} \left(\boldsymbol{Z}\boldsymbol{b}_k^T \big/ l^2 \right) \right\|_2 \tag{15}$$

where T_τ is a soft thresholding operator defined as

$$[T_\tau(\boldsymbol{x})]_\eta = \begin{cases} 0 & |x_\eta| \leq \tau \\ x_\eta - \text{sign}(x_\eta)\,\tau & \text{otherwise} \end{cases} \tag{16}$$

The dictionary $\boldsymbol{D}$ is updated once all atoms $\boldsymbol{d}_k$ are updated.

3.2. The update of adaptive projection

With $\boldsymbol{D}$ and $\boldsymbol{B}$ fixed, the model of SVDL changes to

$$\min_{\boldsymbol{\gamma}_i} \sum_{i=1}^{c} \|\boldsymbol{g}_i - \boldsymbol{R}_i \boldsymbol{\gamma}_i\|_F^2 + \|\boldsymbol{X}_i \odot \boldsymbol{\gamma}_i - \boldsymbol{D}\boldsymbol{B}_i\|_F^2 + \lambda_1 \|\boldsymbol{\gamma}_i\|_F^2 \tag{17}$$

which has an analytic solution:
$\boldsymbol{\gamma}_i = (\boldsymbol{R}_i^T \boldsymbol{R}_i + \sum_{v=1}^{n} \boldsymbol{X}_{i,(v)}^T \boldsymbol{X}_{i,(v)} + \lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{R}_i^T \boldsymbol{g}_i + \sum_{v=1}^{n} \boldsymbol{X}_{i,(v)}^T \boldsymbol{D}\boldsymbol{\beta}_{i,(v)})$.

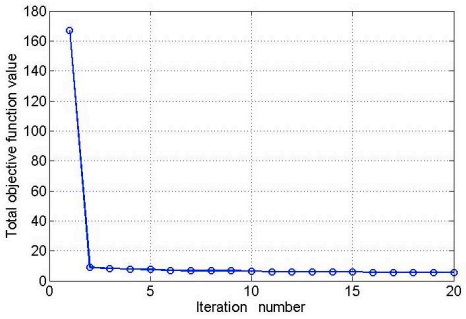

Figure 2. The convergence curve of SVDL objective function on the CMU Multi-PIE database.

3.3. Algorithm and classifier

Algorithm 1 Sparse Variation Dictionary Learning(SVDL)

1: **Initialization** $\boldsymbol{\gamma}_i, i = 1, \cdots, c$.
$\boldsymbol{\gamma}_i$ is set as $\hat{\boldsymbol{\gamma}}_i$ in Eq. (11).
2: **Sparse Variation Dictionary Learning**
While not converge **do**
Update $\boldsymbol{B}$ via standard sparse coding with fixed $\boldsymbol{D}$.
Update dictionary $\boldsymbol{D}$ atom by atom by solving Eq. (13).
End while
3: **Adaptive Projection Learning**
Update the projection matrix $\boldsymbol{\gamma}_i$ via Eq. (17).
4: **Output**
Return to step 2 until the values of the objective function in Eq. (10) in adjacent iterations are close enough or the maximum number of iterations is reached.
Output $\boldsymbol{D}$.

The algorithm of SVDL is summarized in Algoritm 1. Since in each round of alternative minimization, the objective function of SVDL will decrease, the proposed algorithm will converge. Fig. 2 plots the empirical convergence curve of SVDL on the CMU Multi-PIE database, from which we see that the proposed SVDL algorithm converges quickly.

With the learnt dictionary $\boldsymbol{D}$, the testing sample $\boldsymbol{y}$ could be coded as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - [\boldsymbol{G}, \boldsymbol{D}]\,\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \tag{18}$$

where $\boldsymbol{G} = [\boldsymbol{g}_1, \cdots, \boldsymbol{g}_i, \cdots, \boldsymbol{g}_c]$ is the gallery set. In the case that $\boldsymbol{y}$ is occluded, similar to SRC [27] which imposes l_1-norm on representation residual to tolerate outliers, we code $\boldsymbol{y}$ as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - [\boldsymbol{G}, \boldsymbol{D}]\,\boldsymbol{\alpha}\|_1 + \lambda \|\boldsymbol{\alpha}\|_1 \tag{19}$$

Let $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_1; \hat{\alpha}_2; \cdots; \hat{\alpha}_c; \hat{\boldsymbol{\alpha}}_D]$, and $\hat{\alpha}_i$ is the coefficient associated with class i. The classification is conducted via

$$\text{identity}(\boldsymbol{y}) = \arg\min_i \{e_i\} \tag{20}$$

where $e_i = \|\boldsymbol{y} - \boldsymbol{g}_i\hat{\alpha}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}}_D\|_2$ (for FR without occlusion), or $\|\boldsymbol{y} - \boldsymbol{g}_i\hat{\alpha}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}}_D\|_1$ (for FR with occlusion).

4. Experimental Results

In this section, we perform FR with STSPP on benchmark face databases, including large-scale CMU Multiple PIE [6], FRGC [15] and LFW [26], to demonstrate the performance of SVDL. We first discuss the parameter setting in Section 4.1; in Section 4.2 we test the robustness of SVDL to various variations; in Section 4.3, we conduct experiments on the challenging FRGC and LFW databases.

We compare the proposed SVDL with state-of-the-art methods on FR with STSPP, including ESRC [5], Adaptive Generic Learning (AGL) for Fisherfaces [19], and Discriminative Multi-Manifold Analysis (DMMA) [11], and baseline classifiers such as SRC [27], Nearest Subspace (NS) and Support Vector Machine (SVM). It should be noted that NS is reduced to Nearest Neighbor (NN) in the case of FR with STSPP. Among these methods, NN, SVM, SRC and DMMA do not use a generic training set, while ESRC, AGL, and SVDL need a generic training set.

For a more comprehensive evaluation and better demonstration of the proposed SVDL, we also report the performance of ESRC by coupling it with a variation dictionary learnt via KSVD [1]. In the so-called ESRC-KSVD, a dictionary that can sparsely represent the generic variation matrix is learned via KSVD, and the classification is conducted in the same way as ESRC except that the generic variation is replaced by the learned dictionary.

4.1. Parameter setting

There are three regularization parameters, λ_1, λ_2 and λ_3, in SVDL. λ_1 regularizes the projection from the generic training set to the gallery set, while λ_2 and λ_3 control the sparsity of representation coefficients and dictionary atoms, respectively. λ_3 is set to a relatively small value to tolerate more global variation. λ_2 is set to a relatively high value to enforce sparse representation over the learned dictionary. In all the experiments, we fix $\lambda_1 = 0.001$, $\lambda_2 = 0.01$ and $\lambda_3 = 0.0001$. In addition, the number of dictionary atoms should be set beforehand. Because the proposed SVDL algorithm could adaptively remove the redundant basis (refer to the sentence below Eq. (14)), we set the number of dictionary atoms as 400 in the initialization.

4.2. Robustness to various variations

We first test the robustness of all the competing methods by using the large-scale CMU Multi-PIE database [6], whose images were captured in four sessions with simultaneous variations of pose, expression, and illumination. For each subject in each session, there are 20 illuminations with indices from 0 to 19 per pose per expression. Among the 249 subjects in Session 1, the first 100 subjects were used for gallery training, with the remaining subjects for generic training. For the gallery set, we used the single frontal image with illumination 7 and neutral expression. In the following tests with various variations, the images in the generic training set include all the face images with corresponding expression or pose variation, and the frontal face image with neutral expression in Session 1. The image is cropped to 100×82. Except for AGL [19] and DMMA [11] which learn their own features, all the other competing methods use 90-dimensional Eigenface [2] features in the experiments of FR with illumination, expression and pose variations, and use down-sampled images (size: 25×20) as the feature in the experiments of FR with occlusion.

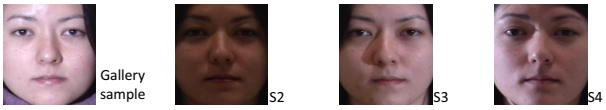

Figure 3. Images with illumination variations in different sessions.

Table 1. The recognition rates (%) on Multi-PIE database with illumination variations.

Session	S2	S3	S4
NN	45.3	40.2	43.7
SVM	45.3	40.2	43.7
SRC[27]	52.4	46.7	49.5
DMMA[11]	63.2	55.4	60.4
AGL[19]	84.9	79.4	78.3
ESRC[5]	92.6	84.9	86.7
ESRC-KSVD	92.7	84.9	86.7
SVDL	**94.8**	**87.7**	**91.0**

1) Illumination variation: We use all the frontal face images with neutral expression in Sessions 2, 3, and 4 for testing. Fig. 3 shows some samples of one subject, including a gallery sample and three testing samples (e.g., S2, S3 and S4 for Sessions 2, 3, and 4, respectively). Table 1 lists the recognition rates in the three sessions by the competing methods.

From Table 1, we can see that SVDL achieves the best results in all cases. ESRC and ESRC-KSVD perform the second best, followed by AGL. SRC does not get good result since the single training sample of each class has very low representation ability. DMMA is the best method without generic training; nonetheless, its recognition rates are not high since the illumination variation cannot be well learned from the gallery set via multi-manifold learning.

2) Expression and illumination variations: In this experiment, the testing samples include the frontal face images with smile in Session 1 (Smi-S1), smile in Session 3 (Smi-S3), surprise in Session 2 (Sur-S2), and squint in Session 2 (Squ-S2) (please refer to Fig. 4 for examples). The recognition rates of all competing methods are listed in Table 2.

We can see that SVDL outperforms all the other methods

Figure 4. Images with expression variations in different sessions.

Table 2. The recognition rates (%) under on Multi-PIE database with expression and illumination variations.

Expression	Smi-S1	Smi-S3	Sur-S2	Squ-S2
NN	46.9	28.8	18.0	25.6
SVM	46.9	28.8	18.0	25.6
SRC[27]	49.6	28.1	20.4	25.7
DMMA[11]	58.2	31.5	22.0	27.5
AGL[19]	84.9	39.3	31.3	23.5
ESRC[5]	81.6	50.5	49.6	41.7
ESRC-KSVD	85.0	50.4	51.2	40.7
SVDL	**88.8**	**58.6**	**54.7**	**47.3**

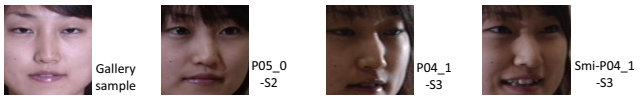

Figure 5. Images with pose variations in different sessions.

Table 3. Face recognition rates (%) on Multi-PIE database with pose, expression and illumination variations.

Pose	P05_0-S2	P04_1-S3	Smi-P04_1-S3
NN	26.0	8.7	12.0
SVM	26.0	8.7	12.0
SRC[27]	25.0	7.3	10.3
DMMA[11]	27.1	5.3	11.0
AGL[19]	66.7	24.9	23.9
ESRC[5]	63.9	31.8	26.9
ESRC-KSVD	67.1	29.9	25.6
SVDL	**77.8**	**38.3**	**34.4**

in all four tests, with at least 3%, 8%, 3% and 5% improvements over the second best, ESRC-KSVD. The variation dictionary learned by KSVD does not improve the recognition accuracy of ESRC much. In addition, all the methods achieve the best results when Smi-S1 is used for testing because the training set is also from Session 1. All the methods have the lowest recognition rate on Squ-S2, probably because a squint expression is more difficult to recognize. Again, the methods with generic training usually have much better performance than the ones without generic training.

3) Pose, illumination and expression variations: In this experiment, the testing samples include face images with pose 05_0 in Session 2 (P05_0-S2), pose 04_1 in Session 3 (P04_1-S3), and pose 04_1 and smile expression in Session 3 (Smi-P04_1-S3) (please refer to Fig. 5 for examples). The recognition rates of all competing methods are listed in Table 3.

From Table 3, we see that SVDL's recognition rates are at least 10%, 6%, and 7% higher than all the other competing methods on the three cases, respectively. The methods with generic learning are significantly better than the ones without generic learning. The improvement of SVDL over ESRC/ESRC-KSVD demonstrates the benefit of jointly learning of the adaptive projection and variation dictionary.

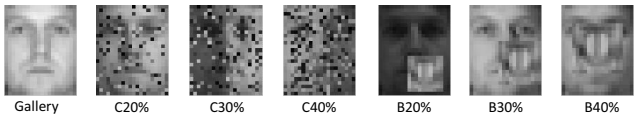

Figure 6. Images with random corruption (e.g., C20%) or block occlusion (e.g., B30%).

Table 4. The recognition rates (%) under different ratios of random corruption.

Corruption ratio	0%	10%	20%	30%	40%
NN	49.1	48.1	44.4	40.4	31.1
SVM	49.1	48.1	44.4	40.4	31.1
SRC[27]	55.5	49.2	44.8	40.1	29.8
DMMA[11]	77.6	65.5	49.3	29.8	15.6
AGL[19]	98.7	9.3	5.1	2.8	2.4
ESRC[5]	99.5	90.1	68.8	42.8	25.3
ESRC-KSVD	99.0	91.1	72.6	47.5	29.1
SVDL	**100**	**98.8**	**100**	**99.3**	**97.2**

Table 5. The recognition rates (%) under different ratios of block occlusion.

Block ratio	0%	10%	20%	30%	40%
NN	49.1	39.8	35.0	29.6	23.3
SVM	49.1	39.8	35.0	29.6	23.3
SRC[27]	55.5	44.2	38.4	33.3	24.5
DMMA[11]	77.6	64.9	51.4	37.5	23.6
AGL[19]	98.7	71.1	53.6	39.8	30.5
ESRC[5]	99.5	85.6	68.4	56.4	41.8
ESRC-KSVD	99.0	84.0	68.9	55.6	42.5
SVDL	**100**	**98.8**	**87.7**	**75.2**	**59.2**

4) Random corruption and block occlusion: In this experiment, the remaining 19 frontal face images of the first 100 subjects in Session 1 with various illuminations (all illuminations but 7) and neutral expression are used as the clean testing images.

We first test the robustness of SVDL to random corruption. As in [27], for each testing image, we replace a certain percentage of its pixels by uniformly distributed random values within [0, 255], where the random locations of corrupted pixels are unknown to all the algorithms. Some corrupted examples are shown in Fig. 6. Table 4 presents the results of all methods under percentages of corrupted pixels from 0% to 40%. It can be seen that in all cases, SVDL could achieve an accuracy of 100% or nearly 100%. Though AGL, ESRC and ESRC-KSVD could get good recognition accuracy when there is no corruption, their recognition rates drop dramatically with the increase of corruption ratio. Compared with ESRC and ESRC-KSVD, SVDL is much more robust to image corruption.

Next we test the robustness of SVDL to block occlusion.

As in [27], we replace a randomly located square block of each test image with an unrelated image, where the location and intensity of occlusion are unknown to all algorithms. Some block-occluded examples are shown in Fig. 6. The recognition results under levels of block occlusion from 0% to 40% are listed in Table 5. Conclusions similar to that of random pixel corruption can be drawn: SVDL performs much better than all the other methods, with at least 18% improvements at 30% and 40% occlusion ratios. In addition, we can see that AGL and DMMA are less sensitive to block occlusion than to pixel corruption.

4.3. FRGC and LFW databases

We then conduct FR with STSPP on the large-scale FRGC database [15] and LFW database [26]. On FRGC, we perform the test on the challenging 4^{th} experiment, which has a target set with 16,028 samples and a query set with 8,014 samples collected from 466 subjects. The samples in the target set were captured under controlled illumination, while the samples in the query set were captured under uncontrolled illumination. For each subject, a single sample with normal illumination and neutral expression in the target set is selected to build the gallery set, and all the 8,014 face images in the query set are used for testing. Some example images in the gallery set and testing set are shown in the last two rows of Fig. 7. We see that this test is very challenging due to the variations of illumination, expression, blur, disguise, time, etc.

The uncontrolled LFW face dataset is usually used to test the face verification methods. Here we adopt it for face recognition by using a subset of aligned LFW [26]. This subset consists of samples from 136 subjects with more than 2 and less than 60 instances. LFW is more challenging than FRGC since it includes various uncontrolled variations of pose and misalignment, etc. We choose a single sample per subject to construct the gallery set, with the remaining images as the testing set.

For the experiment on FRGC, the generic training set is built from the frontal images in the Session 1 of CMU Multi-PIE database. We used the neutral-expression images with illuminations 0,1,7,13,14,16,18 and the smile-expression images with illuminations 0,1,13,14,16,18 for all subjects. For the experiment on LFW, we added face images with pose 05_0, pose 04_1 with smile expression, and their mirrored images in Session 1 to the generic training set. Some examples in the generic training set on the experiment of FRGC are shown in the first row of Fig. 7. The sizes of face images in all sets are cropped to 96×82.

Table 6 shows the rank-1 and rank-10 recognition accuracies on FRGC by using 300-dimensional Eigenfaces, 460-dimensional Eigenfaces and 50×40 down-sampled images as features. AGL [19] utilizes the feature learnt by itself, while DMMA [11] is only conducted on the 50×40

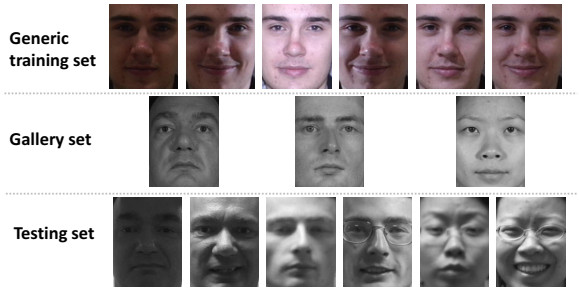

Figure 7. Examples of the face images in the generic training set, gallery set and testing set.

Table 6. Rank-1 (Rank-10) recognition accuracies (%) under different face feature dimensions on the FRGC database.

Dimension	300	460	2000
NN	3.4(13.2)	3.4(13.3)	3.5(13.2)
SVM	3.4(13.2)	3.4(13.3)	3.5(13.2)
SRC[27]	6.9(20.4)	7.3(21.1)	7.3(20.6)
DMMA[11]	—	—	4.1(14.2)
AGL[19]	11.1(25.9)	7.8(22.3)	—
ESRC[5]	12.9(32.9)	13.9(34.9)	14.6(35.6)
ESRC-KSVD	13.6(34.0)	14.0(35.3)	14.8(36.1)
SVDL	**15.2(37.1)**	**16.8(39.2)**	**19.8(42.3)**

down-sampled images. From Table 6, we see that the recognition rates of all the methods are much lower than those obtained in CMU Multi-PIE. This is because the variations produced in uncontrolled environments are much more difficult to process than those produced in controlled environments, especially for the FR problem with STSPP. However, our proposed SVDL could still achieve higher recognition rates than the other methods with all dimensions of face features. The improvement of SVDL over AGL is about 6.5% for rank-1 accuracy and 14% for rank-10 accuracy. The improvement of SVDL over ESRC (ESRC-KSVD) is about 3% for rank-1 accuracy and 4% for rank-10 accuracy. This validates that the proposed joint adaptive projection and sparse variation dictionary learning is more powerful than either of those alone for FR with STSPP.

Table 7 shows the rank-1 recognition accuracies on LFW by using 100-dimensional Eigenfaces, 135-dimensional Eigenfaces, and 50×40 down-sampled images as features. Similar to the results on FRGC, no methods can achieve very high recognition accuracy due to the uncontrolled-challenging face variations. Nevertheless, the proposed SVDL method still achieves the best results among all the competing methods.

5. Conclusion

We proposed a sparse variation dictionary learning (SVDL) method, which learns a sparse variation dictionary from a generic training set to improve face recognition performance with a single training sample per person (STSPP).

Table 7. Rank-1 recognition accuracies (%) under different face feature dimensions on the LFW database.

Dimension	100	135	2000
NN	9.9	10.4	10.6
SVM	9.9	10.4	10.6
SRC[27]	20.4	22.0	22.3
DMMA[11]	—	—	16.2
AGL[19]	19.4	19.9	—
ESRC[5]	22.5	24.0	26.7
ESRC-KSVD	22.4	24.3	26.4
SVDL	**24.2**	**25.7**	**30.2**

The SVDL is adaptive to the gallery set by simultaneously learning a projection from the gallery set to the generic set. Hence, the correlation between the generic set and gallery set can be exploited, and the learned sparse variation dictionary can more effectively aid the single training sample to represent the query image. The extensive experiments with various face variations demonstrated the superiority of SVDL to state-of-the-art face recognition methods with STSPP.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE SP*, 54(11):4311–4322, 2006.

[2] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE PAMI*, 19(7):711–720, 1997.

[3] S. Chen, J. Liu, and Z. Zhou. Making flda applicable to face recognition with one sample per person. *Pattern Recognition*, 37(7):1553–1555, 2004.

[4] T. Chen, W. T. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang. Total variation models for variable lighting face recognition. *IEEE PAMI*, 28(9):1519–1524, 2006.

[5] W. H. Deng, J. N. Hu, and J. Guo. Extended src: Undersampled face recognition via intra-class variant dictionary. *IEEE PAMI*, 34(9):1864–1870, 2012.

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

[7] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[8] T. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE PAMI*, 27(3):318–327, 2005.

[9] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[10] A. N. Li, S. G. Shan, and W. Gao. Coupled biascvariance tradeoff for cross-pose face recognition. *IEEE TIP*, 21(1):305–315, 2012.

[11] J. W. Lu, Y. P. Tan, and G. Wang. Discriminative multimanifold analysis for face recognition from a single training sample per person. *IEEE PAMI*, 35(1):39–51, 2013.

[12] J. Marial, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE PAMI*, 34(4):791–804, 2011.

[13] P. Mller, G. A. Kalberer, M. Proesmans, and L. V. Gool. Realistic speech animation based on observed 3d face dynamics. *IET Vision, Image & Signal Processing*, 152(4):491–500, 2005.

[14] H. Mohammadzade and D. Hatzinakos. Expression subspace projection for face recognition from single sample per person, 2012. to appear in IEEE Affective Computing.

[15] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR*, 2005.

[16] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionary learning for sparse representation modeling. *Proceeding of the IEEE*, 98(6):1045–1057, 2010.

[17] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE SP*, 58(3):1553–1564, 2010.

[18] S. Shan, B. Cao, W. Gao, and D. Zhao. Extended fisherface for face recognition from a single example image per person. In *ISCAS*, 2002.

[19] Y. Su, S. Shan, X. Chen, and W. Gao. Adaptive generic learning for face recognition from a single sample per person. In *CVPR*, 2010.

[20] A. H. T. Ahonen and M. Pietikainen. Face recognition with local binary patterns. In *ECCV*, 2004.

[21] X. Tan, S. Chen, Z. Zhou, and F. Zhang. Recognizing partially occluded expression variant faces from single training image per person with som and soft k-nn ensemble. *IEEE NN*, 16(4):875–886, 2005.

[22] X. Tan, S. Chen, Z. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39:1725–1745, 2006.

[23] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Subspace learning from image gradient orientations. *IEEE PAMI*, 34(12):2454–2466, 2012.

[24] W. T. Y. W. Deng and Y. Zhang. Group sparse optimization by alternating direction method. Technical Report 11-06, Rice University, 2011.

[25] J. Wang, K. Plataniotis, J. Lu, and A. Venetsanopoulos. On solving the face recognition problem with one training sample per subject. *Pattern recognition*, 39:1746–1762, 2006.

[26] L. Wolf, T. Hassner, and Y. Taigman. Effective face recognition by combining multiple descriptors and learned background statistics. *IEEE PAMI*, 33(10):1978–1990, 2011.

[27] J. Wright, A. Y. Yang, S. S. Sastry, and Y. Ma. Robust face recogntion via sparse representation. *IEEE PAMI*, 31(2):210–227, 2009.

[28] Q. Yin, X. O. Tang, and J. Sun. An associate-predict modelfor face recognition. In *CVPR*, 2011.

[29] D. Zhang, S. Chen, and Z. Zhou. A new face recognition method based on svd perturbation for single example image per person. *Applied Mathematics and Computation*, 163(2):895–907, 2005.

[30] W. Zhao, R. Chellppa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Survey*, 35(4):399–458, 2003.