Int J Comput Vis
DOI 10.1007/s11263-014-0750-4

Neither Global Nor Local: Regularized Patch-Based Representation for Single Sample Per Person Face Recognition

Shenghua Gao · Kui Jia · Liansheng Zhuang · Yi Ma

Received: 8 May 2013 / Accepted: 3 July 2014

Abstract This paper presents a regularized patch-based representation for single sample per person face recognition. We represent each image by a collection of patches and seek their sparse representations under the gallery images patches and intra-class variance dictionaries at the same time. For the reconstruction coefficients of all the patches from the same image, by imposing a group sparsity constraint on the reconstruction coefficients corresponding to the patches from the gallery images, and by imposing a sparsity constraint on the reconstruction coefficients corresponding to the intra-class variance dictionaries, our formulation harvests the advantages of both patch-based image representation and global image representation, i.e. our method overcomes the side effect of those patches which are severely corrupted by the variances in face recognition, while enforcing those less discriminative patches to be constructed by the gallery patches from the right person. Moreover, instead of using the manually designed intra-class variance dictionaries, we propose to learn the intra-class variance dictionaries which not only greatly accelerate the prediction of the probe images but also improve the face recognition accuracy in the single sample per person scenario. Experimental results on the AR, Extended Yale B, CMU-PIE, and LFW datasets show that our method outperforms sparse coding related face recognition methods as well as some other specially designed single sample per person face representation methods, and achieves the best performance. These encouraging results demonstrate the effectiveness of regularized patch-based face representation for single sample per person face recognition.

Keywords Single sample per person · Regularized patch-based representation · Group sparsity · Intra-class variance dictionary

Communicated by Yi Ma.

S. Gao (✉) · Y. Ma
School of Information Science and Technology, ShanghaiTech University, Shanghai, China
e-mail: gaoshh@shanghaitech.edu.cn

Y. Ma
e-mail: mayi@shanghaitech.edu.cn

K. Jia
Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Macau, China
e-mail: kuijia@gmail.com

L. Zhuang
CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Hefei, China
e-mail: lszhuang@ustc.edu.cn

1 Introduction

Face recognition (FR) is a classical and important problem in both computer vision and pattern classification because of its potential applications in security, video surveillance, human-computer interface, etc. In real applications, single sample per person (SSPP) face recognition is more realistic and more important because of the limitations on the availability of training photos for persons to be identified in many application scenarios, for example, passport identification and gate ID identification. Furthermore, variations between the test faces (probe images) and their training faces (gallery images) in illumination, expression, occlusion, etc., usually exist and make SSPP face recognition even more challenging.

It is very important for SSPP to use the right image representation. An effective image representation should be able to overcome the effect of variances in expression, illumination, pose, occlusion, etc. Lots of work (Deng et al. 2012;

Published online: 26 July 2014

Zhu et al. 2012; Su et al. 2010; Lu et al. 2011) has been done to explore different image representations for more effective SSPP. These methods can be roughly categorized as global image representation and patch-based/local representation. Global image representation represents each image by one feature vector (Lee et al. 2006; Su et al. 2010). On one hand, global image representation is robust to the recognition of those regions which are not very discriminative, like cheek, forehead, etc., in SSPP. On the other hand, global image representation can be easily affected by those regions that are severely corrupted by variances in illumination, occlusion, expression, etc., Patch-based/local representation divides each image into a collection of patches, representing each patch with one feature vector (Zhu et al. 2012; Gottumukkal and Asari 2004), and it predicts the label of the image based on the labels predicted by all the patches. Therefore a patch-based representation can easily avoid the effect of severely corrupted noninformative regions, but may suffer from non-discriminative regions. After representing each image with a global feature or a collection of local features corresponding to all the patches, a classification technique, like nearest neighbor (NN), sparse representation based classification (SRC) (Wright et al. 2009), collaborative representation based classification (CRC) (Zhang and Feng 2011), or patch-based CRC (PCRC) (Zhu et al. 2012) can be used to predict the label of each image or each patch. Recently, Deng et al. (2012) propose an extended sparse representation for classification (ESRC) which extends SRC to the SSPP case. By introducing an intra-class variance dictionary that characterizes possible variances of the probe images, such a global representation based ESRC method shows promising results in SSPP.

Interestingly, the advantages of global image representation and those of patch-based representation are rather complementary. To harvest the advantages of both patch-based representation and global image representation, and to overcome their disadvantages, we propose a regularized patch-based representation (RPR) for face recognition in the SSPP setting. Specifically, in our method, each face image is represented by a collection of patches. For the patches from the same image, we simultaneously seek the sparse representations of them with respect to their gallery patches and their corresponding intra-class variance dictionaries, with some additional constraints. Here the gallery patches are the patches from the gallery images, while the intra-class variance dictionaries are learned to capture intra-class variations. For reconstructing coefficients for all the image's patches, we impose a group sparsity constraint on the coefficients associated with the gallery patches, and impose a sparsity constraint on the coefficients associated with the intra-class variance dictionaries. Such structured sparsity constraints allow the identification of non-discriminative patches to be predicted by the discriminative ones, and they also allow a small number of severely corrupted regions to be represented (sparsely) by their intra-class variance dictionaries. In this way, such a regularized patch-based representation elegantly integrates the advantages of both global and patch-based image representations while avoiding their respective shortcomings.

The main contributions of this paper can be summarized as follows. (i) We propose a regularized patch-based image representation for SSPP face recognition. By seeking the sparse representations for all the patches of the same image simultaneously, our regularized patch-based image representation integrates the advantages of both the global image representation and the patch-based image representation. (ii) To improve the computational efficiency of our regularized patch-based image representation, we propose to learn the intra-class variance dictionaries automatically from data other than using the manually designed dictionaries. (ii) We evaluate the effect of different regularizers on the reconstruction coefficients and validates the effective of our regularized patch-based image representation in the case of using intra-class variance dictionary. Experimental results show that in additional to the improvement in the computational efficiency, our learnt dictionaries also boost the recognition accuracy.

The rest of the paper is organized as follows. In Sect. 2, work related to our SSPP face recognition is reviewed. In Sect. 3, we first briefly revisit the ESRC based face recognition, after which we propose the formulation of our regularized patch-based image representation as well as its optimization. In Sect. 4, we introduce the details of learning the intra-class variance dictionary. We experimentally evaluate our proposed method in Sect. 5, and we conclude our work in Sect. 6.

2 Related Work

General face recognition is comprised of two subproblems: face representation and pattern classification. For an effective and efficient face representation, subspace analysis methods are usually adopted among which Eigenfaces (Turk and Pentland 1991) and Fisherfaces (Belhumeur et al. 1997) are two representative methods. Eigenfaces are generated by performing principle component analysis (PCA) on a large set of face data. Then each image can be represented by these Eigenfaces. Fisherfaces projects face data onto a subspace which maximizes the between-class distance and minimizes the within-class distance by using Fisher linear discriminative analysis (FLDA). After face representation, a pattern classification technique is used to to predict the label for any given probe image. For face recognition, nearest neighbor (NN) and nearest subspace (NS) are two typical classifiers. NN predicts the label of the test face based on the label of the nearest training sample. NS represents the test face with

the instances from each class sequentially, and assigns it to the class with the minimum reconstruction error. Breaking from NN and NS, Wright et al. (2009) propose to use all the training samples from all the subjects to sparsely represent the test sample, and to assign the label of the class with the minimum reconstruction error to the test sample. Such a sparse representation based face classification (SRC) technique achieves considerable success for the face recognition with enough training samples. Aside from sparse representation, Zhang and Feng (2011) argue that collaborative representation for face classification (CRC) is the key for the success of using SRC. By replacing the ℓ_1 norm constraint on the reconstruction coefficients with the squared ℓ_2 norm constraint, CRC greatly accelerates the computational speed. Moreover, Gao et al. (2013) also propose to use kernel sparse representation for face recognition by mapping features to a high dimensional reproducing kernel Hilbert space (RKHS), which further improves the recognition accuracy.

However, previous image representation methods and collaborative representation based methods (e.g. CRC and SRC) only work well under the condition that there are sufficient training samples for each subject so that the intra-class variances can be covered by these training samples. Compared with general face recognition, the key problem in SSPP face recognition is that there is only one training sample for each person, hereby affecting both the image representation and classification. Specifically,

- Though unsupervised image representation methods, like PCA (Turk and Pentland 1991), 2DPCA (Yang et al. 2004), and Kernel PCA (Kim et al. 2002), can still be directly used, they may easily suffer from variances in occlusion, expression, illumination, etc. that usually accompany with the probe images but cannot be estimated by only one gallery image per subject. To make such unsupervised methods more suitable for SSPP, projection-combined principal component analysis ($(PC)^2A$) (Wu and Zhou 2002), enhanced $(PC)^2A$ ($E(PC)^2A$) (Chen et al. 2004b), etc., are proposed, and the idea behind these methods is to perform PCA with the help of virtual faces, like first order projection (Wu and Zhou 2002), or second order projection (Chen et al. 2004b). These methods have demonstrated good performance on some simple data, but they cannot handle real SSPP data where intra-class variance is more significant and cannot be estimated from virtual samples. Different from these methods, an uniform pursuit algorithm is proposed by Deng et al. (2010) and this uniform pursuit algorithm improves the SSPP by discriminating similar looking faces with the help of an additional dataset, and such uniform pursuit algorithm can be viewed as an extension of PCA by taking advantage of local neighborhood information.
- FLDA (Belhumeur et al. 1997) based image representation is impossible to be directly used in the SSPP scenario because the within-class variance cannot be estimated by only one gallery image for each subject. To make FLDA feasible in SSPP, many efforts have been made, and they can roughly be categorized as virtual samples based methods and generic training sets based methods. Virtual samples based methods generate the images with the same category label via a small perturbation (Martinez 2002), some kind of transformation (Shan et al. 2002), an SVD decomposition on the original images (Gao et al. 2008), or by dividing the whole image into small sub-images (patches) with the same size (Chen et al. 2004a). Generic training set based methods (Su et al. 2010; Kim et al. 2005) introduce a separate dataset which includes the possible variances in, for example, illumination, pose, expression, and occlusion. Then the within-class (and between-class) variance of each gallery image is estimated with the help of this generic dataset.
- SSPP greatly restricts the classification accuracy of the NN classifier, NS classifier, SRC and CRC. To solve this problem, some researchers Zhu et al. (2012), Tan et al. (2005), Kumar et al. (2011), and Lu et al. (2011) propose to divide each image into many sub-images (patches), and perform classification for these sub-images by either using NN, CRC, or some other classifier. Then the classification results of all the sub-images can be aggregated to make the final decision, where a majority voting strategy is usually used. As previously stated, SRC and CRC depend on having enough training samples that include all possible variance of each subject, therefore SSPP setting greatly restricts their performance. To make SRC suitable in SSPP, Deng et al. (2012) successfully extended SRC by introducing an intra-class variance dictionary which characterizes the variances in illumination, occlusion, and expression. Such extended SRC (ESRC) achieves good performance for SSPP. But in ESRC, the intra-class variance dictionary is manually designed and is usually large, thus greatly reducing its computational efficiency. Moreover, ESRC is based on a global image representation that may suffer from image regions that are severely corrupted by variances in SSPP. Recently, Yang et al. (2013) and Deng et al. (2013) further improve ESRC with better intra-class variance dictionary but these methods are still based on the global image representation.

3 Regularized Patch-Based Presentation with Intra-class Variance Dictionaries

In this section, we first briefly review the ESRC based face recognition. Then we introduce the formulation of our regu-

larized patch-based image representation as well as its optimization.

3.1 A Revisit of ESRC for Face Recognition

3.1.1 ESRC for SSPP

To extend SRC to the SSPP task, Deng et al. (2012) proposed an ESRC which manually constructs an intra-class variance dictionary from an additional collection of faces which cover the possible intra-class variance for each class (following Su et al. (2010), we also name this dataset as the generic dataset). Then a given probe image can be reconstructed by using the single training sample which has the same class label with this probe image and the intra-class variance dictionary. Mathematically, we denote the training set as $A = [A_1, \ldots, A_N]$ where $A_i \in \mathbb{R}^d$ only contains one training sample from the ith class, and we denote the intra-class variance dictionary as D which includes the possible intra-class variances. Therefore, in ESCR, the reconstruction of the test sample can be written as $y = A\alpha + D\beta + e$. Similar to the SRC, the reconstruction coefficients in ESRC are also sparse. So ESRC is formulated as follows:

$$\min_{\alpha,\beta} \quad \|\alpha\|_1 + \|\beta\|_1 \quad \text{s.t. } \|y - A\alpha - D\beta\|_2 \leq \epsilon. \tag{1}$$

After getting the sparse reconstruction coefficients, ESRC also predicts the label of the test sample in the same way as SRC, i.e. it assigns the test sample to the class with the minimum reconstruction error.

$$\text{Label}(y) = \arg\min_i \|y - A_i\alpha_i - D\beta\|_2. \tag{2}$$

Here α_i is the reconstruction coefficient corresponding to the single gallery image of the ith class (A_i).

3.2 Formulation of Regularized Patch-Based Representation

To represent variances in illumination, expression, occlusion, etc., we also adopt intra-class variance dictionaries in our regularized patch-based representation. Given a probe image y, we divide it into N patches, and each patch is characterized as a column feature vector Y_i. Then $Y = [Y_1, Y_2, \ldots, Y_N]$ is used to represent the image y. We also divide all the gallery images into patches in the same way, and name these patches as gallery patches. For the patch Y_i, we denote the gallery patches used for reconstructing it as $\mathcal{A}_i$, and denote its corresponding intra-class variance dictionary as $\mathcal{D}_i$. Here we assume that all the images are well aligned, therefore $\mathcal{A}_i$ is constructed by collecting the patches with the same coordinate of Y_i from all the gallery images. With the help of $\mathcal{A}_i$ and $\mathcal{D}_i$, the patch Y_i can be reconstructed as follows:

$$Y_i = \mathcal{A}_i X_i + \mathcal{D}_i S_i + E_i, \quad \forall i. \tag{3}$$

Firstly, in addition to the intra-class variance dictionaries, the patches corresponding to the same probe image should ideally be constructed by the gallery patches from the same gallery image that the probe image belongs to. However, some faces patches are not very discriminative, for example, the cheek patch in Fig. 1. If we directly impose the sparsity constraint on the reconstruction coefficients corresponding to the gallery patches [X_i in Eq. (3)] as what SRC does, this cheek patch may also be well reconstructed by gallery patch from other persons, which would mislead the recognition of the probe. We denote the reconstruction coefficients corresponding to the gallery patches as $X = [X_1, X_2, \ldots, X_N]$. Therefore, it is desirable that all the non-zero coefficients only appear at the place which corresponding to the person these patches belonging to (please refer to Fig. 1), which results in a row-wise sparse structure on X. Hence, to avoid the misclassification of those less discriminative patches, we impose a group sparsity constraint on X. Secondly, we also assume that variances of the patches between the test sample and the training sample are caused by a limited number of variations. For example, for the patches around eyes, variances in appearance may be caused by wearing glasses, and for the patches around lips, variances may caused by illumination or expression. Therefore, for each patch, the coefficients associated with the intra-class variance dictionary should be sparse. Thirdly, it is also desirable that the reconstruction error should be as small as possible. Based on the above discussions, we arrive at the following optimization problem:

$$\begin{aligned} \min_{X,S,E} \quad & \|E\|_F^2 + \lambda\|S\|_1 + \gamma\|X\|_{2,1} \\ \text{s.t. } & Y_i = \mathcal{A}_i X_i + \mathcal{D}_i S_i + E_i, \quad \forall i. \end{aligned} \tag{4}$$

Here $\|X\|_{2,1} = \sum_m \sqrt{\sum_n X_{mn}^2}$ promotes X to be row-wise sparse, and $E = [E_1, \ldots, E_N]$ corresponds to the reconstruction error. The group sparsity requirement for X inherits the advantages of global image representation. Meanwhile by allowing certain patches to be reconstructed only by the intra-class variance dictionaries, the effect of the severely corrupted patches can be minimized. We name this image representation method as **regularized patch-based representation (RPR)**. We illustrate the idea of such regularized patch-based representation for SSPP face recognition in Fig. 1. After getting the reconstruction coefficients of each patch, we can predict the subject label of each patch based on the minimum reconstruction error criteria used in SRC and ESRC, and we can predict the label of the probe image based on the majority voting of all the patches.

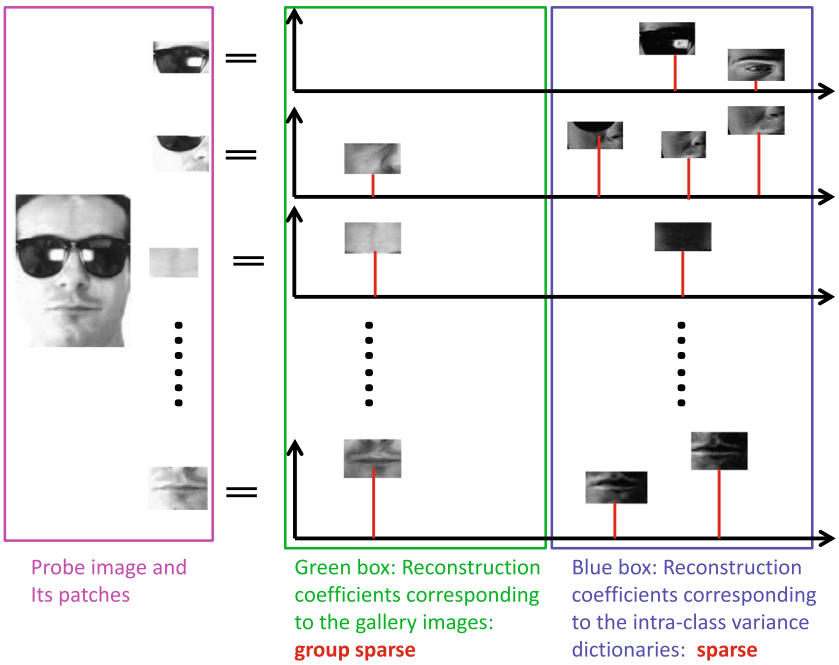

Fig. 1 An illustration of our regularized patch-based image representation for SSPP face recognition. Our method can overcome the effect of noninformative regions severely corrupted by variances like occlusion, while enforcing non-discriminative regions to be reconstructed by the regions from the right person

3.3 Optimization

To solve problem (4), we adopt the commonly used augmented Lagrange multiplier (ALM) method. Specifically, we first convert Eq. (4) to the following problem:

$$\begin{aligned} \min_{X,S,E,G,Z} \quad & \|E\|_F^2 + \lambda\|Z\|_1 + \gamma\|G\|_{2,1} \\ \text{s.t. } & Y_i = \mathcal{A}_i X_i + \mathcal{D}_i S_i + E_i, \quad \forall i \\ & G = X, Z = S. \end{aligned} \tag{5}$$

Based on the ALM method, problem (5) can be further converted to the following unconstraint problem:

$$\begin{aligned} \mathcal{L} = & \|E\|_F^2 + \lambda\|Z\|_1 + \gamma\|G\|_{2,1} \\ & + \mathrm{tr}(H_1^T(G - X)) + \mathrm{tr}(H_2^T(Z - S)) \\ & + \sum_i \mathrm{tr}(J_i^T(Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i)) \\ & + \frac{\mu}{2}\Big(\|G - X\|_F^2 + \|Z - S\|_F^2 \\ & + \sum_i \|Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i\|_F^2\Big). \end{aligned} \tag{6}$$

Here $\mathrm{tr}(\cdot)$ is the trace of a matrix, and $\mu > 0$ is a penalty parameter. Then we alternatively update each unknown variable with the rest of variables fixed. We list the details for solving Eq. (6) in Algorithm 1. Please note that the objectives in steps 1–5 in Algorithm 1 have closed-form solutions. Specifically, for step 1, the objective is in the form (Yang et al. 2009) $\min_V \frac{1}{2}\|X - V\|_F^2 + \mu\|V\|_{2,1}$, and its solution is

$$V(i,:) = \begin{cases} \frac{\|X(i,:)\|_2 - \mu}{\|X(i,:)\|_2} X(i,:), & \text{if } \|X(i,:)\|_2 > \mu, \\ 0, & \text{otherwise.} \end{cases}$$

For step 2, the objective is in the form $\min_V \frac{1}{2}\|X - V\|_F^2 + \mu\|V\|_1$, and it can be solved with the singular value thresholding (SVT) operator (Cai et al. 2008). To get the solutions in steps 3–5, we just set the derivative of the objective function with the corresponding variable to be 0, then we can get the closed-form solutions easily. Because we have the closed-form solution for each subproblem, the optimization in Algorithm 1 is efficient.

4 Designing Intra-class Variance Dictionary

4.1 Manually Designed Intra-class Variance Dictionary

The intra-class variance dictionary plays an extremely important role in removing the effect of troublesome intra-class variances during the recognition of the probe images. Because of the limitation in the number of gallery images (one training image for each person), a **generic dataset**, which contains all possible intra-class variances in, expression, occlusion, illumination, etc., can be used to construct the intra-class variance dictionary. In the following sections, we name the images which are in the generic set and are used to simulate the probe images in the evaluation set as the *reference images*, and name the patches from the reference images as *reference patches*. We also name the rest of faces which contains some variations compared with the reference

Algorithm 1 ALM algorithm for optimizing Eq. (6)

Input: Local patches Y_i, gallery patches $\mathcal{A}_i$, intra-class variance dictionaries $\mathcal{D}_i$, λ and γ.

Output: Reconstruction coefficients X and S.

Initialize: Initialize X and S with the results of ESRC (Following Yuan et al. (2012), we initialize them with the results of ESRC, and we also find that such an initialization strategy achieves better performance than an initialization with 0.). $G = X$, $Z = S$, $E = 0$, $H_1 = 0$, $H_2 = 0$, $J_i = 0, \forall i$, $\mu = 0.5$, $\mu_{max} = 10^{10}$, $\rho = 1.1$ and $\varepsilon = 10^{-7}$.

while not converged **do**

1: Fixing the other variables and update G by

$$G = \arg\min \frac{\gamma}{\mu}\|G\|_{2,1} + \frac{1}{2}\|G - (X - H_1/\mu)\|_F^2.$$

2: Fixing the other variables and update Z by

$$Z = \arg\min \frac{\lambda}{\mu}\|Z\|_1 + \frac{1}{2}\|Z - (S - H_2/\mu)\|_F^2.$$

3: Fixing the other variables and update X by

$$X = \arg\min \mathrm{tr}(H_1^T(G-X)) + \sum_i \mathrm{tr}(J_i^T(Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i)) + \frac{\mu}{2}\left(\|G - X\|_F^2 + \sum_i \|Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i\|_F^2\right).$$

4: Fixing the other variables and update S by

$$S = \arg\min \mathrm{tr}(H_2^T(Z-S)) + \sum_i \mathrm{tr}(J_i^T(Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i)) + \frac{\mu}{2}\left(\|Z - S\|_F^2 + \sum_i \|Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i\|_F^2\right).$$

5: Fixing the other variables and update E by

$$E = \arg\min \|E\|_F^2 + \sum_i \mathrm{tr}(J_i^T(Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i)) + \frac{\mu}{2}\sum_i \|Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i\|_F^2.$$

6: Update the multipliers

$H_1 = H_1 + \mu(G - X)$.
$H_2 = H_2 + \mu(Z - S)$.
$J_i = J_i + \mu(Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i),\ \forall i$

7: Update the parameter μ by $\mu = min(\rho\mu, \mu_{max})$.

8: Check the convergence conditions:
$\|Y_i - \mathcal{A}_i X_i - \mathcal{D}_i S_i - E_i\|_2 < \varepsilon, \forall i$, $\|Z - S\|_\infty < \varepsilon$ and $\|G - X\|_\infty < \varepsilon$.

end while

images as *variation images*, and the patches from these variation images as *variation patch*.

In ESRC (Deng et al. 2012), such an intra-class variance dictionary is generated from one of the following four ways: the difference from a natural image, the difference from the class centroid, pairwise difference, and the original generic samples themselves. Experimental results show that the difference based methods achieve better performance than generic samples themselves (Deng et al. 2012). Hence for our regularized patch-based image representation, we can also manually design an intra-class variance dictionary for each patch. Specifically, for each reference patch at certain location, its intra-class variance dictionary can be generated as follows. We collect all the variation patches which have the same coordinates with the reference patch, and pair them with the reference patch. Then the differences of all these pairs make up the intra-class variance dictionaries.

4.2 Learning the Intra-class Variance Dictionary

One issue for manually designed intra-class variance dictionaries is that their size is usually very large. Therefore such large intra-class variance dictionaries inevitably lead to expensive computational cost when optimizing regularized patch-based face recognition. Therefore one question naturally arises: can we learn some smaller intra-class variance dictionaries without decreasing the performance compared with manually designed intra-class dictionaries? To this end, we propose an intra-class variance dictionary learning strategy. Interestingly, in addition to improving the computational efficiency, the learnt intra-class variance dictionaries also achieve better performance than those based on manually designed dictionaries. One possible reason for the better performance of the learnt dictionaries is that the learnt intra-class variance dictionaries not only include the variances in the generic dataset but also may cover certain variances that do not appear in the generic dataset but appear in the evaluation dataset. Specifically, the manually designed dictionaries are obtained by directly subtracting the variation images in the generic dataset from their respective reference images or the class centroid, therefore these intra-class variance dictionaries are somewhat person-specific, i.e., they are probably restricted by the shape of faces and the positions of the nose and mouth/eye corners of the persons in the generic dataset. Given a probe image in the evaluation dataset, its variances can be well removed probably only under the condition that there is a person in the generic dataset having a similar face shape (if the variances are related to the face shape), or having similar mouth/eye corner positions (if the variances are related to the mouth/eye corners). But for our learnt dictionaries, we require the variances for each person to be reconstructed by using several atoms in our learnt dictionaries. Therefore our dictionary learning method may decompose each variance into several components which are shared by different persons. As a result, the learnt dictionaries are less restricted by person-specific properties, like face shape, the positions of nose and mouth/eye corners, etc., and thus the atoms in these learnt dictionaries may reconstruct the possible variances which appear only in the evaluation dataset.

For the sake of an easy explanation and simple notation, next we will explain how to learn the intra-class variance dictionary for the global image representation. For raw pixel features, the intra-class variance dictionary for each patch can be generated by dividing the global intra-class variance dictionary into local patches based on the coordinates of the patches

of the reference images. Please also note that this intra-class variance dictionary learning method also applies to learning intra-class variance dictionary for each patch regardless of the feature used for patch characterization.

Suppose there are K subjects in the generic dataset, and denote the reference image corresponding to the cth subject in the generic dataset as $\mathcal{A}_c^G$, then all the reference images make up the matrix $\mathcal{A}^G = [\mathcal{A}_1^G, \ldots, \mathcal{A}_K^G]$. Denote all the variation images belonging to the cth subject as $\mathcal{G}_c$. Then we use the following formulation to learn the intra-class variance dictionary.

$$\min_{\mathcal{D},\alpha_c,\beta_c} \sum_{c=1}^{K} \left(\|\mathcal{G}_c - \mathcal{A}_c^G\alpha_c - \mathcal{D}\beta_c\|_F^2 + \nu \left\| \begin{bmatrix} \alpha_c \\ \beta_c \end{bmatrix} \right\|_1 \right) + \eta \|\mathcal{D}^T \mathcal{A}^G\|_F^2$$
$$\text{s.t. } \|\mathcal{D}(:, j)\|_2 = 1, \quad \forall j. \tag{7}$$

Remarks (i) The first two terms mean that each variation image should be sparsely reconstructed by its reference image and the intra-class variance dictionary. (ii) The term $\|\mathcal{D}^T\mathcal{A}^G\|_F^2$ means that the intra-class variance dictionary should be incoherent with the reference image of each subject. Without this term, the $\mathcal{D}$ may itself be enough for the reconstruction of the variation images ($\alpha_c = 0$), which is undesirable because we want the intra-class variance dictionary to only cover the possible variances in SSPP. (iii) The constraint on each column of $\mathcal{D}$ is used to get rid of the trivial solution.

To avoid the effect of differences in the number of variation images used for dictionary learning on different datasets during the experiments, we set $\eta = \eta_0 \times$ the number of columns in $\mathcal{A}^G$. By default, we set $\eta_0 = 10^{-5}$ and $\nu = 10^{-2}$ in all the experiments. Needless to say, our intra-class variance dictionary learning can be plugged into the ESRC (Deng et al. 2012) framework, and it will be shown that our learnt intra-class variance dictionary both accelerates the sparse coding optimization and improves its prediction accuracy (please refer to Sect. 5.3).

4.3 Optimization

The objective function in Eq. (7) is not convex, but it is convex for the reconstruction coefficients $\begin{bmatrix} \alpha_c \\ \beta_c \end{bmatrix}$ when the intra-class variance dictionary $\mathcal{D}$ is fixed, and vice versa. Following Lee et al. (2006), Kong and Wang (2012), we alternatively optimize the reconstruction coefficients and the intra-class variance dictionary. When the intra-class variance dictionary $\mathcal{D}$ is fixed, it is the standard Lasso/sparse coding formulation, and we use the feature-sign search algorithm (Lee et al. 2006) to infer the reconstruction coefficients.[1] When the reconstruction coefficients are fixed, we arrive at the following problem:

Algorithm 2 Learning intra-class variance dictionary

Input: The variation images in the generic dataset $\mathcal{G}_c$; The reference images in the generic dataset $\mathcal{A}_c$ ($c = 1, \ldots, K$); ν; η
Initialize $\mathcal{D}$ by randomly selecting some atoms from the manually designed intra-class variance dictionary.
repeat
  for $i = 1$ **to** K **do**
    Infer the sparse codes $\begin{bmatrix} \alpha_c \\ \beta_c \end{bmatrix}$ by optimizing (7) with $\mathcal{D}$ fixed using the feature-sign search algorithm;
  end for
  for $i = 1$ **to** the number of atoms in $\mathcal{D}$ **do**
    Update each $\mathcal{D}(:, j)$ with Eq. (9);
    Normalize the ℓ_2 norm of $\mathcal{D}(:, j)$ to 1.
  end for
until stopping criteria is reached.
Output: The Intra-Class Variance Dictionary: $\mathcal{D}$.

$$\min_{\mathcal{D}} \quad \|R - \mathcal{D}\beta\|_F^2 + \eta\|\mathcal{D}^T\mathcal{A}^G\|_F^2$$
$$\text{s.t. } \|\mathcal{D}(:, j)\|_2 = 1, \quad \forall j \tag{8}$$

where $R_c = \mathcal{G}_c - \mathcal{A}_c^G\alpha_c$, $R = [R_1, \ldots, R_K]$, $\beta = [\beta_1^T, \ldots, \beta_K^T]^T$. Following the work of Kong and Wang (2012), we alternatively update each column of $\mathcal{D}$. All the rest of the columns are fixed while updating a given column. By setting the derivative of Eq. (8) w.r.t. $\mathcal{D}(:, j)$ to be 0, we get

$$\mathcal{D}(:, j) = (\eta(\mathcal{A}^G)(\mathcal{A}^G)^T + \|\beta(j, :)\|^2\mathbf{I})^{-1}\hat{R}_j\beta(i, :)^T. \tag{9}$$

Here $\hat{R}_j = R - \sum_{i,i\neq j} \mathcal{D}(:, i)\beta(i, :)$. Then we normalize $\mathcal{D}(:, j)$ to make its ℓ_2 norm equal to 1.

Since both the update of the reconstruction coefficients and the update the codebook reduce the objective, the solution of the problem in Eq. (7) converges after several iterations. In Fig. 2 we empirically show the changes of the objective value with respect to the number of iterations in alternative updating the coefficients and codebook on the Extended Yale B and the CMU-PIE datasets. We can see that the solution converges very fast.

5 Experiments

In this section, we evaluate our proposed method on the AR, Extended Yale B, CMU-PIE, and LFW datasets (some samples of the images in the datasets are shown in Fig. 3). Moreover, the effect of different parameters in our formulation will also be empirically evaluated.

[1] Feature-sign search can be regarded as a variant of LARS (Efron et al. 2004) and the LARS method demonstrates good performance for solving sparse coding for face recognition (Wright et al. 2009).

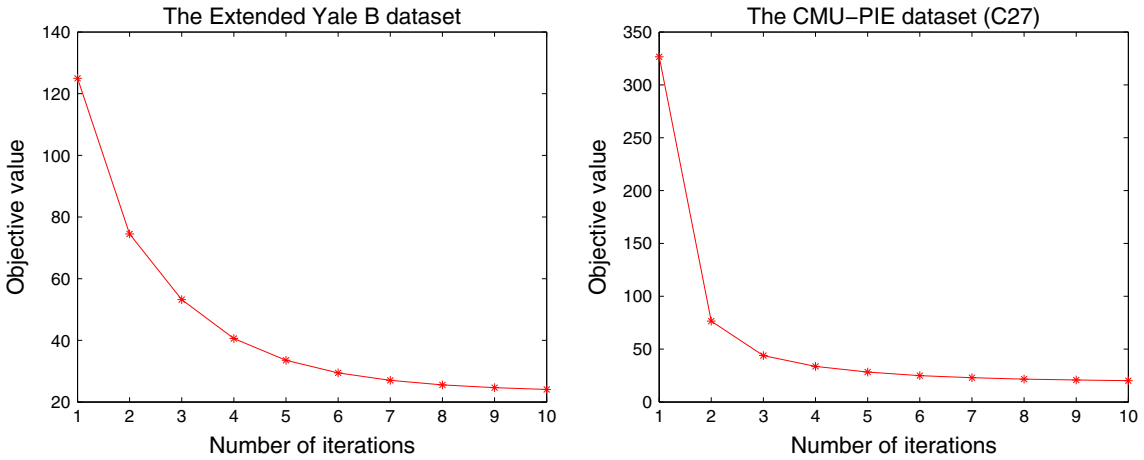

Fig. 2 The change of the objective value with respect to the number of iterations on the Extended Yale B and the CMU-PIE datasets

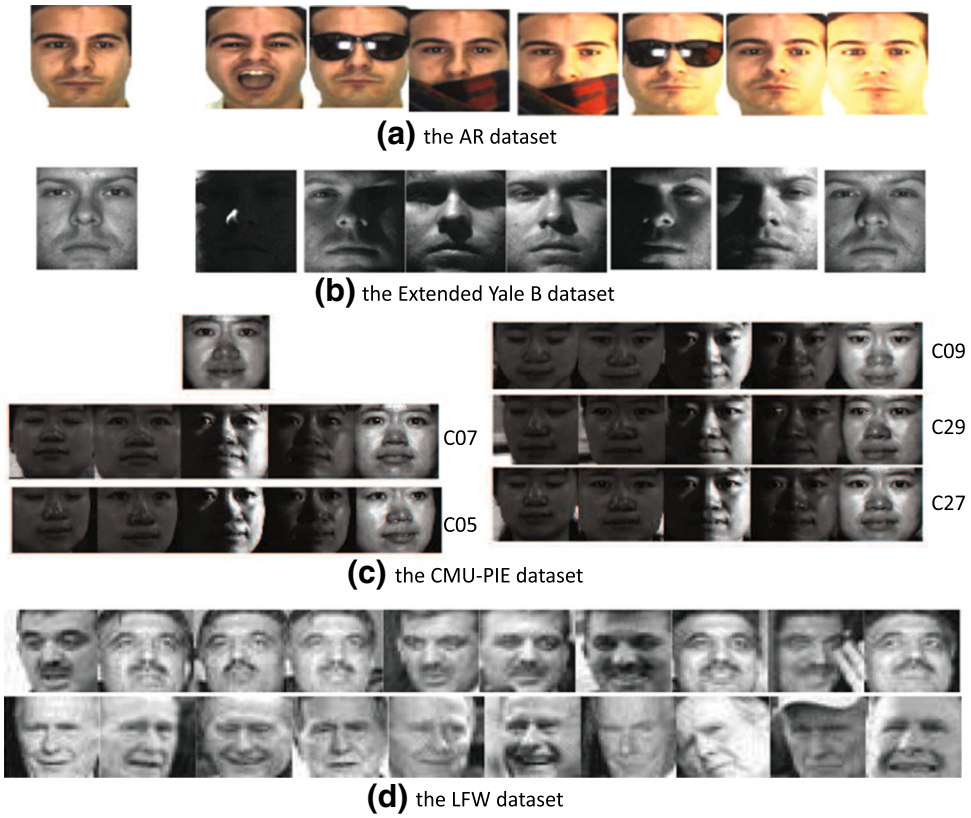

Fig. 3 **a–c** Samples of the gallery image (the first one) and probe images (the rest of the faces) in the AR, Extended Yale B, and CMU-PIE datasets. **d** Some sample images in the LFW dataset. The variances in illumination, pose, occlusion, and expression between the gallery image and probe make single sample per person face recognition extremely challenging

5.1 Experimental Setup

This paper focuses on proposing a regularized patch-based image representation for single sample per person face recognition. To make the comparison fair among all methods and to remove the effect of features, following patch-based collaborative representation for classification (PCRC) (Zhu et al. 2012), unless otherwise stated, we use the intensity of the pixels as the features for all the methods, and all the images are resized to 32×32 pixels on all the datasets. To make a fair comparison and to avoid the effect of patch size on the final performance, we also keep the same setting with PCRC, i.e., the patch size is set to be 8×8 pixels, and the distance between two patch centers is 4, so the patches are overlapped with each other. All the features are normalized with their ℓ_2 norm being 1. In the following experiments, we set $\lambda = 0.001$ and $\gamma = 0.05$ in the formulation of the regularized patch-based representation.

We compare our work with the following work because of their close relationships with our method: (1) SRC (Wright et al. 2009), (2) CRC (Zhang and Feng 2011), (3) ESRC (Deng et al. 2012), (4) PCRC (Zhu et al. 2012), and (5) Sparse variational dictionary learning (SVDL) (Yang et al. 2013).[2] Moreover, we also compare our proposed method with other existing works which are specially designed for solving SSPP problems, including (6) Block LDA (Chen et al. 2004a), (7) SVD based LDA (Gao et al. 2008), (8) AGL (Su et al. 2010), (9) $(PC)^2A$ (Wu and Zhou 2002), and (10) $E(PC)^2A$ (Chen et al. 2004b). Besides the above mentioned baseline methods, we also designed the following two baseline methods: (11) Patch based SRC (PSRC). Similar to PCRC, we use SRC for each patch and majority voting is used for the final label prediction. (12) Fisherfaces (Belhumeur et al. 1997). We learn the projection matrix from the generic set, and apply the learnt projection matrix to the evaluation data. For a fair comparison, the same features are used for both global image representation based methods and patch-based methods.

5.2 Evaluation on Different Datasets

5.2.1 Evaluation with the AR Dataset

The AR dataset (Martinez and Benavente 1998) contains over 4,000 frontal faces which are taken from 126 subjects (70 men and 56 women) in two different sessions, and these images contain variances in occlusion (sunglasses and scarves), expression (neutral expression, smile, angry, and scream), and illumination. Some images contain both occlusion and illumination variances. Following the work of ESRC (Deng et al. 2012), 20 subjects from session 1 are used as the generic dataset for learning the intra-class variance dictionaries, and the rest of 80 subjects, also from session 1, are used for evaluation in our experiments. The size of the intra-class variance dictionaries for all the patches is fixed to be 120. The frontal faces taken under normal lighting conditions and neutral expressions are used as the gallery images, and all the rest of 12 images for each subject are used as the probe images. We list the performance of different methods in Table 1. Results show that our method outperforms other image representation methods that are specially designed for SSPP, including PCRC, PSRC, and ESRC, and it achieves the best performance under all the cases.

Table 1 Performance comparison between different methods on the AR dataset (%)

Method	Subset 1	Subset 2	Subset 3	Subset 4
$(PC)^2A$	60.83	81.25	35.00	15.94
$E(PC)^2A$	60.00	81.25	35.00	15.94
FLDA-Block	69.17	53.33	51.25	37.50
FLDA-SVD	66.25	77.92	55.00	27.50
AGL	74.17	77.50	48.13	38.75
Fisherfaces	71.25	73.33	48.75	38.44
CRC	75.83	80.00	50.62	20.31
PCRC	92.92	91.25	96.88	85.94
SRC	77.08	80.42	50.62	21.56
PSRC	94.17	90.83	96.88	87.50
ESRC	96.67	80.83	84.38	68.44
SVDL	98.33	85.83	88.12	75.56
Ours	99.58	95.83	98.75	93.13

Images in subset 1–4 correspond to changes in illumination, expression, disguise, and illumination+disguise, respectively

Specifically, for the variance in illumination (subset 1), ESRC, PSRC, SVDL, and PCRC have already achieved relatively good performance. But our method still outperforms these methods. For the variances in expression and occlusion (subset 2–4), the global image representation based ESRC performs very poorly, but the patch-based PCRC and PSRC perform relatively well. The reason may be that the informative regions for face recognition like the eyes and lips are affected by expression and occlusion. Therefore global image representation is severely handicapped when these regions vary; but the patch-based method can overcome the side effect of variances in these regions. Therefore, PCRC and PSRC achieve better performance than ESRC. But compared with PCRC and PSRC, our method overcomes the effect of those non-discriminative patches like the forehead and the cheeks, and it further improves face recognition accuracy. For example, for the images in subset 4 which contain variances in both illumination and occlusion, the recognition accuracy of ESRC, PSRC, SVDL, and PCRC is 68.44, 87.50, 76.56, and 85.94 %, respectively, but our regularized patch-based representation achieves a rate of 93.13 %. Therefore the improvement of our method over the existing work is very significant in the presence of the variances in expression and occlusion. These results demonstrate the effectiveness of our regularized patch-based face representation for SSPP face recognition.

Compared with SSPP when only a single variance (subset 1–3) is present, multiple variances (subset 4) decrease the

[2] To learn the variation dictionary with SVDL, all subjects in the generic set should have images for a given type of variation. For LFW, the number of the variation type is unknown, and it is also impossible to find the images with the same type of variation. Therefore it is impossible to learn the dictionary with SVDL on this dataset. It is worth noting that to make SVDL applicable to LFW dataset, Yang et al. (2013) used the data from CMU-MultiPIE dataset as the generic set to learn the dictionary, but such setting is different from ours. For fair comparison, the performance of SVDL under such setting is not included in our paper. For the Extended Yale B and CMU-PIE datasets, some persons in the generic set don't have images for some type of variations. If we remove these persons, it would be unfair to compare SVDL with our method and other baseline methods which use the generic set. Therefore we only include the results of SVDL on the AR dataset.

Table 2 Performance comparison between different methods on the Extended Yale B dataset (%)

Method	S1	S2	S3
$(PC)^2A$	37.38	39.24	39.24
$E(PC)^2A$	37.04	38.89	38.89
FLDA-Block	61.58	60.41	60.41
FLDA-SVD	43.33	41.53	41.53
AGL	58.26	60.23	58.59
Fisherfaces	56.93	62.08	59.61
CRC	45.06	47.44	47.44
PCRC	74.57	71.87	71.87
SRC	42.04	41.45	41.45
PSRC	69.38	69.22	69.22
ESRC	68.00	70.90	69.58
Ours	84.52	86.68	84.92

Table 3 Performance comparison between different methods on the CMU-PIE dataset (%)

Method	C27	C05	C07	C09	C29
$(PC)^2A$	26.09	22.49	19.64	24.05	20.75
$E(PC)^2A$	25.87	22.24	19.64	23.78	20.75
FLDA-Block	64.67	23.21	24.19	34.03	25.95
FLDA-SVD	37.33	18.75	23.59	28.04	17.71
AGL	69.23	40.43	44.39	48.61	48.87
Fisherfaces	66.14	46.56	42.97	48.61	49.05
CRC	51.98	26.96	35.77	42.36	30.38
PCRC	76.58	44.69	57.18	59.20	41.15
SRC	51.46	29.08	35.16	42.01	31.25
PSRC	76.45	41.84	57.35	60.50	41.84
ESRC	81.83	67.35	62.4	70.23	65.28
Ours	88.79	67.82	70.67	76.74	67.45

performance. The possible reason for such a phenomenon is that that for faces with multiple variances, the intra-class variance dictionaries are more complex. But due to the restriction in the size of the generic set, we cannot have enough samples to learn all the possible variances, therefore restricting the performance of the SSPP face recognition.

Moreover, Table 1 also shows that patch based representation achieves better performance than global image representation for both SRC and CRC, which validates the necessity of patch-based representation in SSPP. It is need to mention that though the Fisherfaces learns a projection matrix by minimizing the intra-class variance and maximizing the inter-class variance, but the projection is learnt from the generic set in which the subjects doesn't overlap with the subjects in the evaluation set. Therefore the learnt matrix probably won't make the subjects in the evaluation set separable very well. In contrast to our designed Fisherfaces baseline, AGL (Su et al. 2010) approximates the intra-class variance for the subjects in the evaluation set by leveraging the generic set, which makes it learn a better FLD projection matrix to increase the separability for subjects in the evaluation dataset. Therefore AGL usually achieves better performance than the designed Fisherfaces baseline on AR, and similar phenomenons can also be observed on Extended Yale B, CMU-PIE and LFW in Tables 2, 3, and 4, respectively.

Table 4 Performance comparison between different methods on the LFW dataset (%)

Method	S1	S2
$(PC)^2A$	8.36 ± 0.81	8.36 ± 0.81
$E(PC)^2A$	8.40 ± 1.41	8.40 ± 1.41
FLDA-Block	4.47 ± 0.87	4.47 ± 0.87
FLDA-SVD	6.99 ± 1.15	6.99 ± 1.15
AGL	13.33 ± 2.26	14.31 ± 0.86
Fisherfaces	12.69 ± 1.34	11.89 ± 2.16
CRC	14.57 ± 1.78	14.57 ± 1.78
PCRC	26.62 ± 1.60	26.62 ± 1.60
SRC	7.40 ± 1.35	7.40 ± 1.35
PSRC	28.26 ± 3.43	28.26 ± 3.43
ESRC	24.89 ± 1.87	26.54 ± 1.16
Ours	30.25 ± 1.54	31.39 ± 1.74

5.2.2 Evaluation with the Extended Yale B Dataset

The Extended Yale B dataset (Georghiades et al. 2001) contains 38 categories, and 2,414 frontal-face images with severe changes in illumination. For each subject, we use the frontal face whose light source direction with respect to the camera axis is at 0° azimuth ('A+000') and at 0° elevation ('E+00') as the gallery image, and we use the images with other lighting conditions as probe images. We try three different settings on this dataset. (i) Setting S1. We take 15 subjects as the generic dataset to learn the intra-class variance dictionaries and use the rest of 23 subjects for evaluation. (ii) Setting S2. We take 20 subjects as the generic dataset to learn the intra-class variance dictionaries and use the remaining 18 subjects for evaluation. (iii) Setting S3: We take 15 subjects as the generic dataset, and use 18 subjects for evaluation. In S3, the generic dataset is the same as that in S1, and the evaluation set is the same as that in S2. The size of the intra-class variance dictionaries in S1–S3 are all fixed to be 240. The performance of different methods under different settings is listed in Table 2. We can see that our method achieves the best performance in all of the cases.[3]

[3] $(PC)^2A$, $E(PC)^2A$, FLDA-Block, FLDA-SVD, CRC, PCRC, PSRC, and SRC don't use the generic dataset, so the performance of these methods under S2 and S3 is the same.

Compared with the illumination conditions in the AR dataset, the illumination on the Extended Yale B is more severe and more complex, and sometimes entire faces are almost totally covered by shadow (please refer to Fig. 3), therefore the performance of PCRC and PSRC is also very poor, let alone the global representation based ESRC which is not robust to faces with many severely corrupted regions. Compared with the PCRC which achieves the best performance of all the existing work, the improvement of our method is more than 10 % in all the cases. Moreover, the comparison between S2 and S3 shows that more subjects in the generic dataset helps in learning better intra-class variance dictionaries, and thus boosts the performance of SSPP, which is consistent with the observations in ESRC (Deng et al. 2012).

5.2.3 Evaluation with the CMU-PIE Dataset

The CMU pose, illumination, and expression (CMU-PIE) dataset (Sim et al. 2002) contains 41,368 images of 68 subjects. For each subject, the images are taken under 13 different poses, four different illumination conditions, and four different expressions. We use 20 subjects as the generic dataset to learn the intra-class variance dictionaries, and use all the remaining 48 subjects for evaluation. For each subject, we use the face images taken with the frontal pose (C27), neutral expression, and normal lighting condition as the gallery images, and we use the remaining images with the poses C27, C29, C07, C05, C09 as probe images. We learn different intra-class variance dictionaries for different poses, with the size for the intra-class variance dictionary under all the poses fixed to be 240. The performance of different methods on this dataset is given in Table 3, and our method again achieves the best performance under all the cases. Compared with the frontal face pose (C27), we notice that as the pose changes, the performance of SSPP drops significantly.

In this dataset, each image usually contains multiple variances. Therefore it is more challenging than both AR and Extended Yale B. Interestingly we find that for the frontal pose (C27), faces looking up (C07), and faces looking down (C09), the improvement of our method over ESRC is usually around 7 %, but for the faces looking left (C29) and looking right (C05), the improvement of our method over ESRC is not that significant. The reason may be that we assume that all the images are well aligned, and we collect the patches of probe images by following the same coordinates with the gallery images. But for faces looking left and looking right, usually some parts of the face, like the cheek, are occluded (Fig. 3). These occluded parts cause the misalignment issue, therefore affecting the performance of our regularized patch-based representation. These occluded parts explain the better performance of C07 (looking up) and C09 (looking down) compared to C05 (looking right) and C29 (looking left).

5.2.4 Evaluation with the LFW Dataset

The Labeled Faces in the Wild (LFW) dataset (Huang et al. 2007) contains images of 5,749 individuals taken under an unconstrained setting.[4] LFW-a is a subset of the LFW dataset, and the images in LFW-a have been aligned with a commercial software tool (Wolf et al. 2009). The faces acquired under the unconstrained setting and inaccurate alignment make the LFW data extremely challenging for face verification,[5] let alone face recognition in the SSPP setting. However, our goal here is not to design a full fledged face recognition system. Rather, we want to compare under the same alignment and feature conditions, which representation and classification methods are more appropriate. Following the work of Zhu et al. (2012), we use the LFW-a for evaluation. On this dataset, we conduct experiments under two different settings. (i) Setting S1: Following the work of Zhu et al. (2012), we only use the persons with no less than 10 photos as evaluation data and generic set. There are 158 persons under such setting, among which 78 subjects are used as the generic dataset for learning the intra-class variance dictionaries, and 80 subjects are used for evaluation. The size of the intra-class variance dictionaries is fixed to 390 under such setting, which is the half of the size of the intra-class variance dictionary used in ESRC. (ii) Setting S2: Besides the 78 subjects used in the S1, we also add the 1,522 subjects (4,840 images) which contains 2–9 images per person to the generic set. Therefore the total subjects in the generic set is 1,600 under S2. As the manually designed intra-class variance dictionary is very large (around 5 k), we fix the size of learnt intra-class variance dictionary to be 500. The evaluation data in S2 are the same as that in S1. To learn the intra-class variance dictionaries in S1 and S2, we use the mean face of each person as the reference image in the generic dataset. The reason of using the mean face as the reference image in the generic dataset is that the variances of the face in this dataset are very significant, and misalignment also frequently appears. Meanwhile we observed qualitatively that the mean face looks like the frontal face. Following the work of Zhu et al. (2012), we also randomly choose one image as the gallery image for each subject, and use nine images as the probe images for evaluation. The performance of different methods based on 10 independent experiments on this dataset is listed in Table 4.

[4] The LFW dataset is usually used for the face verification problem.

[5] In order to obtain better face verification and recognition systems for such datasets, typically one needs to use more sophisticated alignment methods for more complicated face shape models.

The results show that our method achieves the best performance, and it outperforms PSRC which achieves the best performance of all the existing work, by about 2 and 3 % under S1 and S2, respectively. Please note that the SSPP face recognition on the LFW dataset is extremely challenging because of the variances in illumination, pose, expression, and occlusion, as well as the grossly simplified alignment used (Fig. 3). Moreover, the gallery image used for evaluation for each person is randomly selected, and it is usually a nonfrontal face with lots of intra-class variances for each subject. Therefore the performance of all the methods on this dataset is very poor. As aforementioned, we notice that the mean face looks like a frontal face, and may overcome the variances in pose, expression, illumination and occlusion. Thus we manage to use the mean face of nine images from each subject as the gallery image, and use one image which is not overlapped with the images for generating the mean face for evaluation. The accuracy of our method is 58.62 ± 3.33 % under S1. Therefore a good gallery image is very important for SSPP face recognition. Further, by comparing with the performance of AGL, ESRC, and our RPR under S1 and S2, we can see that a larger generic set helps improve the recognition accuracy because a larger generic set helps characterize more possible intra-class variances.

Compared with the performance of our method on other datasets, the poorer performance on LFW also suggests that good alignment is indispensable for good performance of the SSPP recognition task. Nevertheless, our experiments clearly show that the regularized part based representation should still hold advantages over other methods even if a more complicated deformation model is used for more careful face alignment.

5.2.5 Evaluation on C27 (CMU-PIE) with the Intra-class Variance Dictionary Learnt from Extended Yale B

We also conduct experiments by using heterogenous data for generic set and evaluation set. Specifically, we use the whole Extended Yale B dataset to learn the intra-class variance dictionary, and evaluate the learnt intra-class variance dictionary with C27 from CMU-PIE which is a collection of frontal faces with different illumination and expression. The intra-class variance dictionary is fixed to be 500 in this experiment. It is worth noting that the evaluation data here are the same with that used in previous experiments. We list the performance of different methods under such setting in Table 5. It can be seen that our method still achieves the best performance, which proves its effectiveness for SSPP face recognition. In spite of more subjects (38 subjects vs. 20 subjects) and images are used as generic set under such setting, it can be observed that that the accuracy under such setting is worse than that by using the intra-class variance dictionary learnt from C27. The reason may be that the Extended Yale B dataset doesn't contain the variance in expression, therefore the dictionary learnt under such setting can only compensate the illumination on C27 dataset. In real applications, we should construct a large generic dataset which covers all possible variances. Such a large generic dataset would be a great help for all the methods based on the generic dataset.

Table 5 Performance comparison on C27 CMU-PIE with different generic sets (%)

Source of generic set	Fisherfaces	AGL	ESRC	Ours
Extended Yale B (38 subjects)	54.03	50.98	71.23	86.57
C27 (20 subjects)	66.14	69.23	81.83	88.79

5.3 Evaluation of Intra-class Variance Dictionary Learning Formulation

5.3.1 Parameter Evaluation

For simplification and computational efficiency, we select the parameters in the formulation of learning intra-class variance dictionary based on the ESRC. We plot the performance of ESRC with different parameters in Eq. (7) in Fig. 4. We can see for all the parameters tested in our experiments, the performance of ESRC based on the learnt dictionary is usually better than that based on the manually designed dictionary. Therefore the proposed formulation is robust to these hyper-parameters in the dictionary learning formulation. For simplification, we fix $\eta_0 = 10^{-5}$ and $\nu = 10^{-2}$ for all the datasets.

5.3.2 Recognition Accuracy and Computational Efficiency

The biggest motivation for learning the intra-class variance dictionaries is that the manually designed dictionaries are usually very large and thus bring about expensive computational costs for the prediction of the faces to be recognized. To further demonstrate the effect of learnt dictionaries in our regularized patch-based face representation, we list the recognition accuracy as well as the computational cost of our face recognition method based on manually designed dictionaries and learnt dictionaries in Table 6.[6] We can see that our learnt dictionaries improve both the efficiency and the accuracy. Especially on the AR dataset, compared with the manually designed dictionaries whose size is 240, our learnt dictionaries whose size is 120, can speed up the prediction time by almost three times, meanwhile improving the recog-

[6] We run the Matlab implementation of these methods on a Windows Server (64bit) with a 2.13GHz CPU and 16GB RAM.

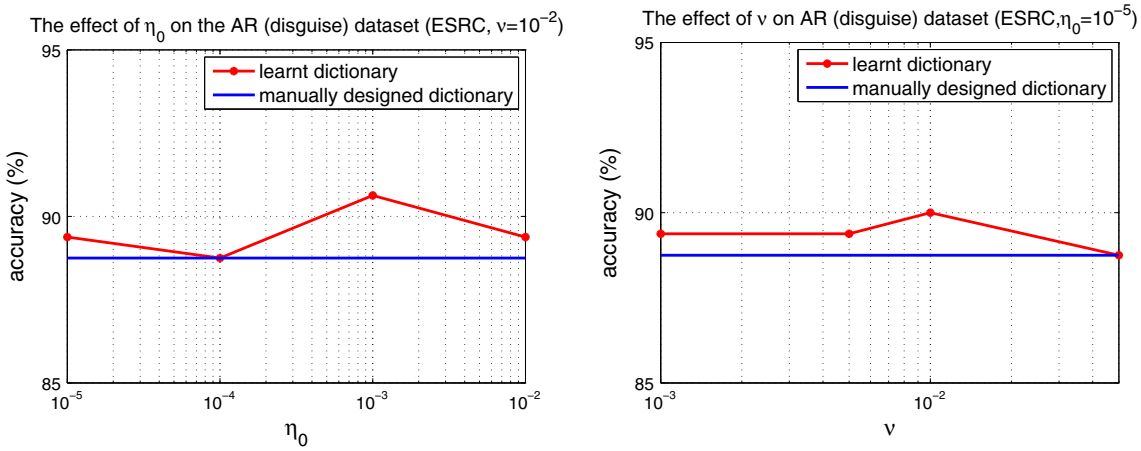

Fig. 4 The effect of different parameters on dictionary learning. For all the parameters we used, the performance based on the learnt dictionary is similar or better compared with that based on the manually designed dictionary. Moreover, our dictionary learning method is robust to η_0 and ν

Table 6 Performance comparison between the learnt dictionaries and the manually designed dictionaries on the AR and CMU-PIE dataset

Dictionary type	AR				CMU-PIE	
	Illumination	Expression	Disguise	Disguise+illumination	C07	C09
Accuracy (%)						
Designed	99.58	95.41	98.75	92.81	67.27	75.00
Learnt	99.58	95.83	98.75	93.13	70.67	76.73
Cost (s)						
Designed	21.89	22.42	24.01	19.76	39.59	19.37
Learnt	7.46	7.87	8.12	7.00	14.84	14.84

nition accuracy. For a larger generic dataset, the improvement in speed is more significant.[7]

5.3.3 Dictionary Size

In Fig. 5 we also experimentally evaluate the effect of the size of intra-class variance dictionary on the performance of our method on the AR dataset. We can see that the recognition accuracy of our method increases with the size of the intra-class variance dictionary and becomes stable when the dictionary reaches a certain size, but the computational cost increases steadily with the size of intra-class variance dictionary. In real applications, we choose the proper size of intra-class variance dictionary based on the trade-off between the accuracy and the computational cost. We also visualize the learnt intra-class variance dictionary on the AR dataset in Fig. 6. We can see that the learnt dictionary covers the possible variances in illumination, occlusion, and expression.

5.3.4 Different Dictionary Learning Strategies

For the case of using the intensity of raw pixels as feature, besides learning the intra-class variance dictionary for the global image representation first and divide it into patches which are used as the intra-class variance dictionaries, we can also learn the patch-specific dictionaries one by one. We show the performance of these two different dictionary learning strategies on the AR and CMU-PIE datasets in Fig. 7. It can be seen that the performance of these two methods is comparable on AR, but the performance based on the dictionary learnt from the global image representation is better on CMU-PIE, especially for pose C05 (looking right) and C29 (looking left). We conjecture the reason for such observation may be that learning intra-class dictionary based on the global image representation makes the learnt intra-class variance dictionaries be regularized by the structure of the face, thus they are more meaningful and more helpful in characterizing the possible intra-class variances for SSPP face recognition. It is also worth noting that there are many patches in all as the patches are overlapped, which makes the computa-

[7] For the CMU-PIE dataset, the manually designed dictionaries are very large for other poses (C27, C05, C29), therefore the prediction is extremely very expensive, and we cannot finish it on our machine in one day. Therefore we didn't report the performance based on the manually designed or learnt dictionaries under those poses. This fact further proves the importance and necessity in learning the intra-class variance dictionaries.

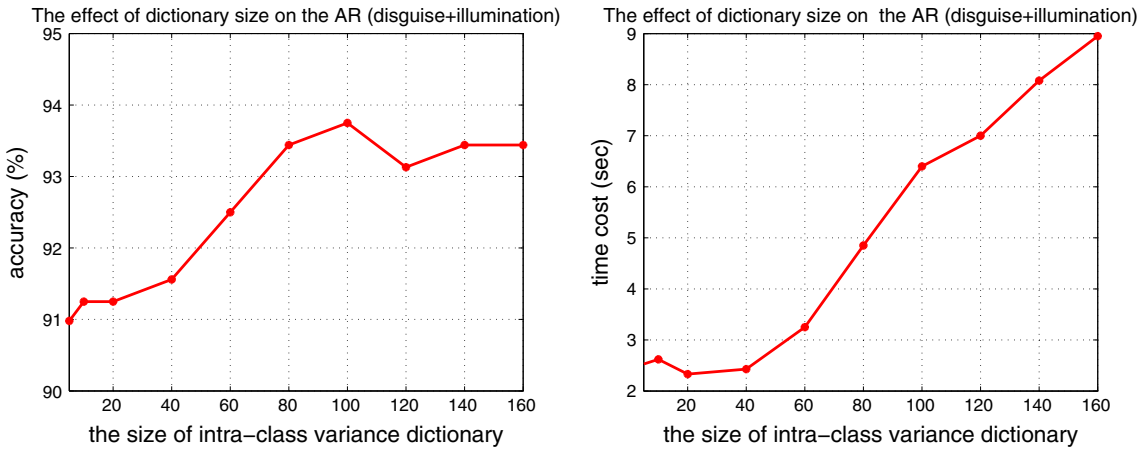

Fig. 5 The effect of intra-class variance dictionary size on the performance of our method. The recognition accuracy increases with the size of the dictionary first and gradually becomes stable when the dictionary reaches some size, but the computational cost increases steadily with the increase of the size of the dictionary

Fig. 6 An illustration of learnt intra-class variance dictionaries on the AR, Extended Yale B, and CMU-PIE datasets, respectively. We can see that the learnt dictionaries characterize the variances in illumination, occlusion and expression

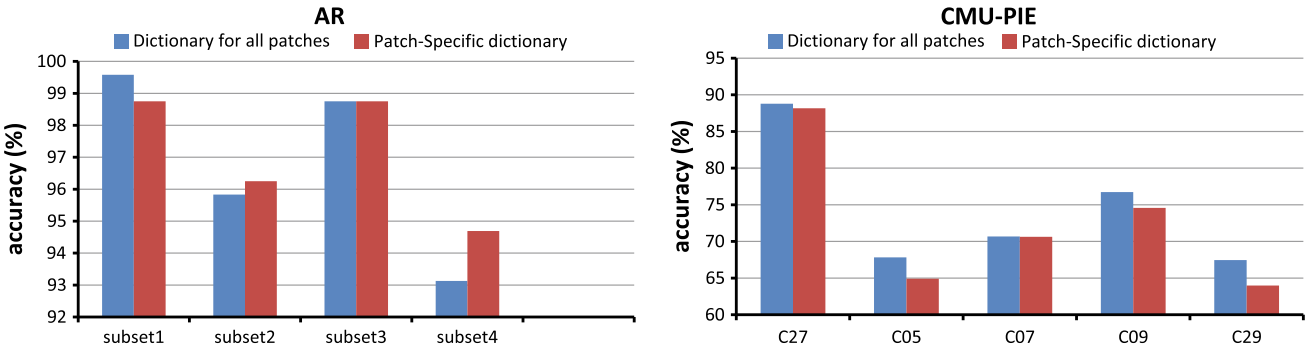

Fig. 7 The comparison between different dictionary learning strategies on AR and CMU-PIE. Images in subset 1–4 correspond to the changes in illumination, expression, disguise, and illumination+disguise, respectively

tional cost of learning patch-specific dictionary more expensive than that of learning the intra-class variance dictionary for the whole image. Therefore for the intensity based features, we learn the intra-class variance dictionary based on the global image representation and divide it into patches. But for other feature based patch characterization, like the features in Sect. 5.5, we have to learn patch-specific intra-class variance dictionaries for patches at different coordinates, separately.

5.4 Evaluation of Regularized Patch-Based Image Representation

5.4.1 Parameter Evaluation

We experimentally evaluate the effect of λ and γ on our regularized patch-based image representation on the AR dataset in Fig. 8. Here we increase λ from 0.001 to 0.05, and increase γ from 0.005 to 0.1. We can see that the

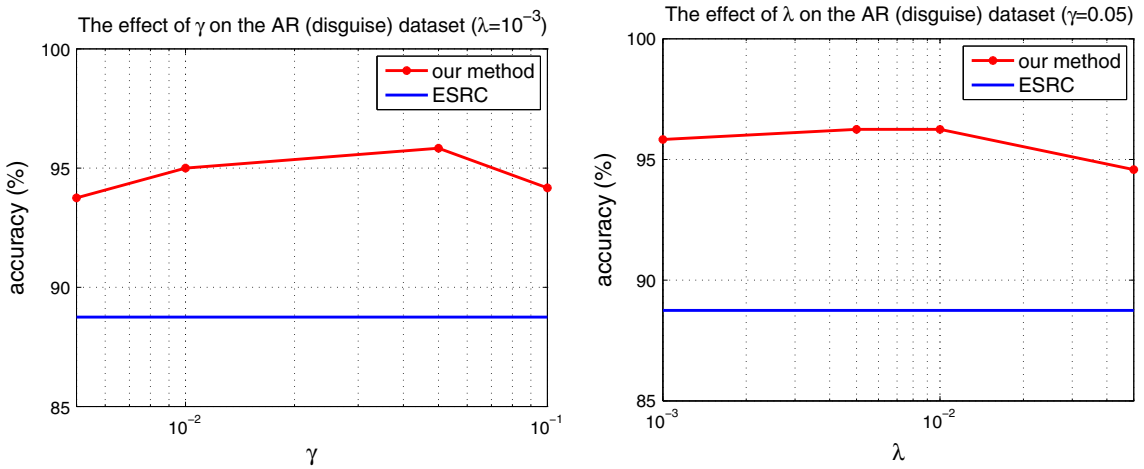

Fig. 8 The effect of different parameters in our formulation for face recognition. The performance of our method is relatively robust to λ and γ

performance of our patch-based ESRC is relatively stable for the changes of these two parameters. Moreover, for all the parameters we tested, our method always outperforms the ESRC. Hence we fix $\lambda = 0.001$ and $\gamma = 0.05$.

5.4.2 *Different Regularizers on the Reconstruction Coefficients*

To further demonstrate the effectiveness of our method, we also design some other baselines with different constraints on the reconstruction coefficients in Eq. (3). (i) We extend PCRC of Zhu et al. (2012) to the case of using the intra-class variance dictionaries by replacing the sparsity constraint and group sparsity constraint in the RPR formulation [Eq. (4)] with the squared ℓ_2 norm. We denote this method as extended PCRC (EPCRC); (ii) We propose to conduct ESRC of Deng et al. (2012) on patch level by replacing the group sparsity constraint in the RPR formulation with the sparsity constraint, and denote such method as extended PSRC (EPSRC). (iii) We set $\lambda = 0$ in the RPR formulation, which means we only enforce the small reconstruction error and the group sparsity of the reconstruction coefficients corresponding to the gallery patches. (iv) We set $\gamma = 0$ in the RPR formulation, which means we only enforce the small reconstruction error and the sparsity of the reconstruction coefficients corresponding to the intra-class variance dictionaries. (5) We switch off the constrains on the reconstruction coefficients ($\lambda = \gamma = 0$) in the RPR formulation and minimize the reconstruction error only. The comparisons between our method and these baselines on different subsets of the AR dataset are listed in Table 7. Some interesting phenomenons can be observed in Table 7.

Firstly, it can be easily seen that the performance of formulations without the constraints on the reconstruction coefficients corresponding either gallery patch dictionary, intra-class variance dictionaries, or both dictionaries, is worse than that of RPR with proper regularizers. This proves the crucialness of regularizing the reconstruction coefficients. Moreover, we also notice that only adding the constrains on coefficients corresponding to one dictionary may result even worse performance than that without any regularizers. Specifically, compared with the recognition accuracy by only minimizing the reconstruction error ($\lambda = \gamma = 0$), if only the coefficients corresponding to patches of gallery is regularized with group sparsity, the performance improves for the illumination and expression cases. This observation hints that group sparsity is crucial for the recognition where face contains locally corrupted regions or nondiscriminat regions, like are blocked by sun-glasses. If only the coefficients corresponding to the intra-class variance dictionary are regularized, the performance improves for the face recognition containing occlusions and expressions. The reason for this may be that these variances of different persons are very similar on AR. For example, the occlusions comes from sunglasses and scarf, and expressions only contain three types. Therefore these variances can be relatively easily character-

Table 7 Performance comparison with different constraints on the reconstruction coefficients on the AR dataset (%)

Method	Subset 1	Subset 2	Subset 3	Subset 4
$\lambda = 0$	95.83	81.25	66.87	54.69
$\gamma = 0$	89.17	85.00	87.50	73.75
$\lambda = \gamma = 0$	94.67	76.67	85.00	69.37
EPCRC	74.58	74.58	81.87	62.81
EPSRC	98.75	94.16	98.75	91.35
Ours	99.58	95.83	98.75	93.13

Images in subset 1–4 correspond to the changes in illumination, expression, disguise, and illumination+disguise, respectively

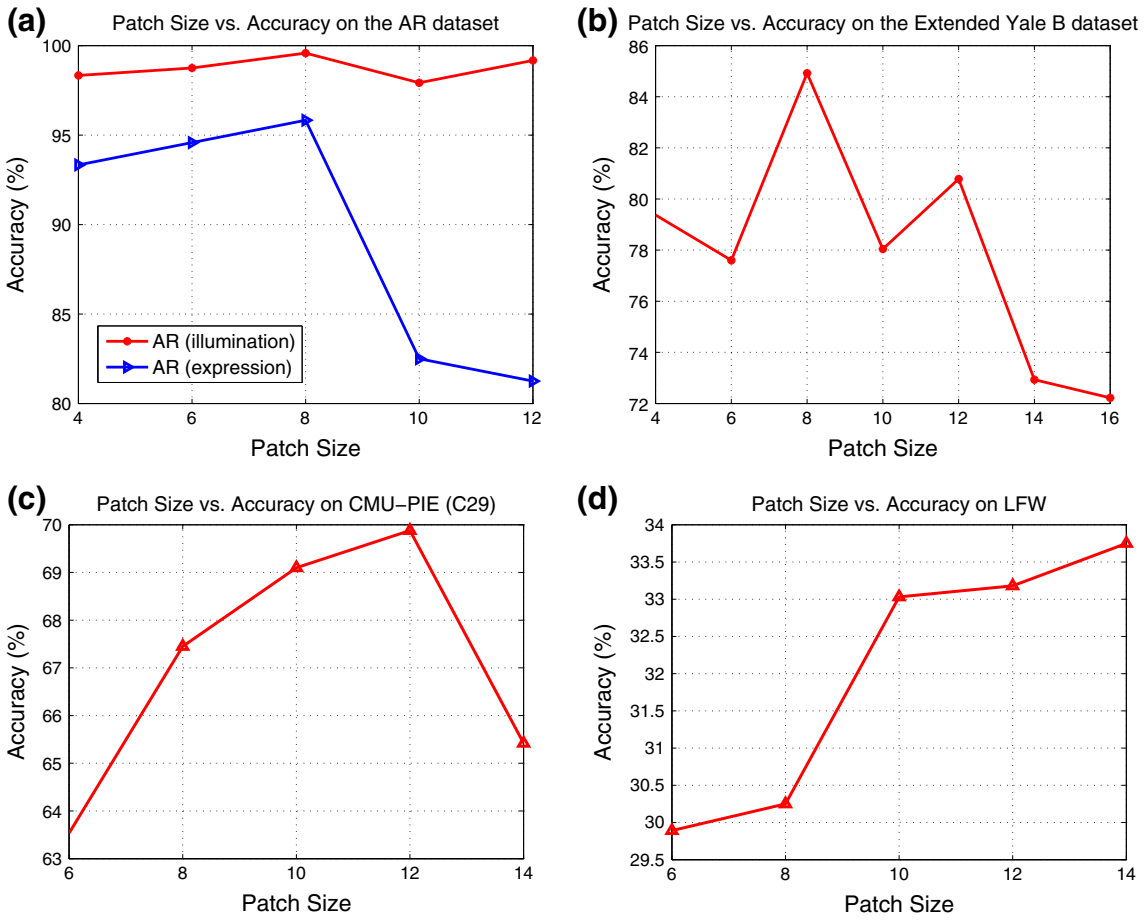

Fig. 9 The performance results using patches with different size

ized by the intra-class variance dictionary with the sparsity constraint.

Secondly, interestingly we find that with the intra-class variance dictionary, the ℓ_2 norm regularized EPCRC performs even worse than that without the intra-class variance dictionary (PCRC). It seems our observation contradict with the perseveration in Zhang and Feng (2011) work that both ℓ_1 norm and ℓ_2 norm achieve good performance for face recognition. However, actually both results are valid because the dictionary setting in these two papers are different paper. In Zhang and Feng (2011), the atoms in the dictionary are all the training faces. In our setting, only a small fraction of bases in the dictionary are the training faces, and most bases characterize the possible intra-class variances which are irrelevant to the gallery images. Densely combining these irrelevant bases and only a small fraction of bases of gallery images for the recognition of test sample would mislead the recognition. Similar phenomenon is also observed and analyzed in the work of Deng et al. (2013). Such observation validates the effectiveness of sparse representation for the SSPP with intra-class variance dictionary.

Thirdly, EPSRC achieves the same accuracy for the disguise case with our RPR, but RPR achieves better performance than EPSRC for illumination, expression, and faces containing variance in both illumination and disguise on the AR dataset. The reason for this observation is that the AR dataset is a well-controlled face recognition dataset, and for the disguise case, either the sun-glasses or the scarf blocks the face. These blocked region can be well overcome by the intra-class variance dictionary. The unblocked faces regions are still discriminative enough for face recognition. For the case of expression and illumination, the recognition can be mislead by those patches that are not very class-specific. With the help of group sparsity, the recognition of those patches can be boosted. To further verify our assumption, we compare EPSRC and our RPR on the CMU-PIE dataset where each face usually contains multiple variances. For the cases of C05 and C29 (looking left and right), the performance of EPSRC is 62.63 and 63.54 %, which is lower than our method by 5.19 and 3.91 %, respectively. These comparisons verifies the effectiveness of group sparsity regularizer on the face recognition with complex variances.

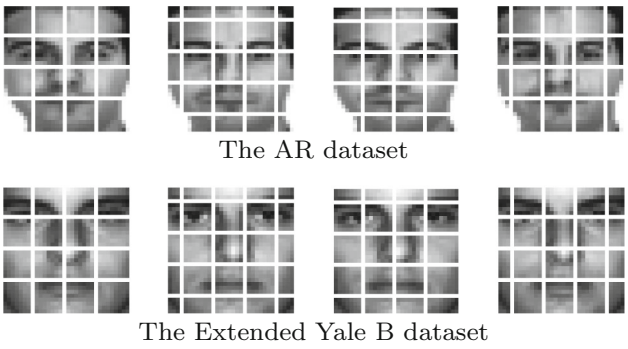

Fig. 10 An illustration of the 8×8 patches on the AR and Extended Yale B datasets. These patches usually fall over the semantically meaningful parts of the faces

5.4.3 Patch Size

We also experimentally evaluate the effect of patch sizes on the face recognition accuracy. We list the performance of using patches with different sizes on AR dataset, Extended Yale B (setting S3), CMU-PIE (C29), and LFW in Fig. 9. We can see that the performance is the best when the patch size is 8 on AR and Extended Yale B and the possible reason for this is that the 8×8 patches usually cover the semantically meaningful part of the face on these two datasets (Fig. 10), like the eyes, the nose, the lips, and these these patches are the most informative parts of the face. But on CMU-PIE and LFW, the optimal patch size is different. It is worth noting that the optimal patch size is determined by the way of aligning and cropping faces and the size of the cropped faces. As shown in Fig. 3, as the face are aligned and cropped in different methods on these datasets, though the image size are the same, the optimal patch size are still different on these datasets.

5.5 The Effect of Different Features

In addition to the intensity of the pixels (64D), we also implement RPR based on the following features. It is worth noting that for the patch-specific dictionaries are learnt for the following features. (1) Modular PCA (MPCA) (Gottumukkal and Asari 2004) (32D), which divides each image into patches and performs PCA on these patches, (2) Block LDA (Chen et al. 2004a) (32D), which treats patches from the same gallery image as instances with the same class label and performs LDA at patch level, (3) uniform patterns based LBP (Ahonen et al. 2006) (59D), and (4) Gabor feature (Liu and Wechsler 2002) (20D). We list the performance of these features on the AR dataset in Table 8. Results show that the intensity feature and MPCA feature based RPR usually achieve the best performance in terms of recognition accuracy for patch-based image representation.

Table 8 Performance comparison between different features on the AR dataset (%)

Feature type	Subset 1	Subset 2	Subset 3	Subset 4
Intensity based RPR	**99.58**	95.83	**98.75**	93.13
MPCA based RPR	99.17	**97.08**	**98.75**	**95.31**
LDA-Block based RPR	96.67	85.42	94.38	78.13
LBP based RPR	81.67	95.00	97.50	74.38
Gabor based RPR	95.42	95.42	95.00	89.69

Images in subset 1–4 correspond to the changes in illumination, expression, disguise, and illumination+disguise, respectively. The best performance is highlighted with bold font

The possible reason for the good performance of MPCA feature may be that such patch-based PCA removes some noises which may harm face recognition. Moreover MPCA feature is more computational efficient because of its lower dimensionality.[8]

5.6 Computational Complexity

The main computational cost of our RPR based SSPP comes from the two parts: (i) Learning the intra-class variance dictionary in Eq. (7); and (ii) Solving the reconstruction coefficients in the RPR formulation in Eq. (4) with ALM algorithm.

For learning the intra-class variance dictionary, its computational cost is $O(LMn^2d^3)$ where L is the number of outer loop in the Algorithm 2, M is the number of atoms in the intra-class variance dictionary, n is the number of variation images in the generic set, and d is the dimensionality of the features.[9] Empirically, for the setting S2 on the LFW dataset where the number of images in the generic dataset is more than 5 k and the atoms in the intra-class variance dictionary is 500, the dictionary can be learnt within 1 h on a workstation with four 2.80GHz Intel Xeon CPUs.

For the computational cost of solving the RPR, because each subproblem of RPR in the Algorithm 1 has a closed-form solution, we can still solve the RPR efficiently. Specifically, we list the computational costs of AGL and some SRC and CRC based methods in Table 9. Compared with ESRC, though we have more patches, the intra-class variance dictionary is smaller than that in ESRC, which helps improve the efficiency of our method. Specifically, our method is about three times slower than ESRC and two times faster than EPSRC. Our method can be further accelerated with more elaborately designed implementation.

[8] Please also note that all the results are based on the same parameters, which are designed for intensity features. Fine-tuned parameters for other features may further improve their performance.

[9] The computational cost for updating the sparse coefficients with feature-sign-search algorithm is less expensive than that of updating the intra-class variance dictionary.

Table 9 Time costs of different methods (s)

CRC	PCRC	SRC	PSRC
0.0068	0.067	0.033	0.37
AGL	ESRC	EPSRC	Ours
0.00073	0.43	2.2	1.2

6 Conclusion

In this paper, we propose a regularized patch-based representation for the single sample per person face recognition task. Our formulation harvests the advantages of both patch-based image representation and global image representation. Moreover, we also propose to learn the intra-class variance dictionary which not only accelerates the face prediction but also improves the recognition accuracy. Experimental results on the AR, Extended Yale B, CMU-PIE, and LFW datasets demonstrate the effectiveness of our proposed method.

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 2037–2041.

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Cai, J. F., Candes, E. J., & Shen, Z. (2008). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, *20*(4), 1956–1982.

Chen, S., Liu, J., & Zhou, Z. H. (2004a). Making FLDA applicable to face recognition with one sample per person. *Pattern Recognition*, *37*, 1553–1555.

Chen, S., Zhang, D., & Zhou, Z. H. (2004b). Enhanced $(PC)^2A$ for face recognition with one training image per person. *Pattern Recognition Letters*, *25*, 1173–1181.

Deng, W., Hu, J., Guo, J., Cai, W., & Feng, D. (2010). Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach. *Pattern Recognition*, *43*(5), 1748–1762.

Deng, W., Hu, J., & Guo, J. (2012). Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 1864–1870.

Deng, W., Hu, J., & Guo, J. (2013). In defense of sparsity based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2), 407–499.

Gao, S., Tsang, I. W., & Chia, L. T. (2013). Sparse representation with kernels. *IEEE Transactions on Image Processing*, *22*(2), 423–434.

Gao, Q., Zhang, L., & Zhang, D. (2008). Face recognition using FLDA with single training image per person. *Applied Mathematics and Computation*, *205*, 726–734.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 643–660.

Gottumukkal, R., & Asari, V. K. (2004). An improved face recognition technique based on modular PCA approach. *Pattern Recognition Letters*, *25*(4), 429–436.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report*. Amherst, MA: University of Massachusetts.

Kim, K. I., Jung, K., & Kim, H. J. (2002). Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, *9*, 40–42.

Kim, T. K., Kittler, J., & Kittler, J. (2005). Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 318–327.

Kong, S., & Wang, D. (2012). A dictionary learning approach for classificaiton: separating the particularity and the commonality. In *Proceedings of the European Conference on Computer Vision*.

Kumar, R., Banerjee, A., Vemuri, B. C., & Pfister, H. (2011). Maximizing all margins: Pushing face recognition with kernel plurality. In *Proceedings of the International Conference on Computer Vision*.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2006). Efficient sparse coding algorithms. In *Proceedings of the Conference on Neural Information Processing Systems*.

Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, *11*(4), 467–476.

Lu, J., Tan, Y. P., & Wang, G. (2011). Discriminative multi-manifold analysis for face recognition from a single training sample per person. In *Proceedings of the International Conference on Computer Vision* (pp. 1943–1950).

Martinez, A. M. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(6), 748–763.

Martinez, A., & Benavente, R. (1998). *The AR face database* (Vol. 24).

Shan, S., Cao, B., Gao, W., & Zhao, D. (2002). Extended fisherface for face recognition from a single example image per person. In *IEEE International Symposium on Circuits and Systems*.

Sim, T., Baker, S., & Bsat, M. (2002). The CMU pose, illumination, and expression (PIE) database. In *International Conference on Automatic Face and Gesture Recognition*.

Su, Y., Shan, S., Chen, X., & Gao, W. (2010). Adaptive generic learning for face recognition from a single sample per person. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tan, X., Chen, S., Zhou, Z. H., & Zhang, F. (2005). Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble. *IEEE Transactions on Neural Networks*, *16*, 875–886.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Wolf, L., Hassner, T., & Taigman, Y. (2009). Similarity scores based on background samples. In *Proceedings of the Asian Conference on Computer Vision*.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 210–227.

Wu, J., & Zhou, Z. H. (2002). Face recognition with one training image per person. *Pattern Recognition Letters*, *23*, 1711–1719.

Yang, M., Van Gool, L., & Zhang, L. (2013). Sparse variation dictionary learning for face recognition with a single training sample per person. In *International Conference on Computer Vision*.

Yang, J., Yin, W., Zhang, Y., & Wang, Y. (2009). A fast algorithm for edgepreserving variational multichannel image restoration. *SIAM Journal on Imaging Sciences*, *2*(2), 569–592.

Yang, J., Zhang, D., Frangi, A. F., & Yang, J. Y. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(1), 131–137.

Yuan, X. T., Liu, X., & Yan, S. (2012). Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, *21*(10), 4349–4360.

Zhang, L., & Feng, X. (2011). Sparse representation or collaborative representation: Which helps face recognition? In *Proceedings of the International Conference on Computer Vision*.

Zhu, P., Zhang, L., Hu, Q., & Shiu, S. C. (2012). Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In *Proceedings of the European Conference on Computer Vision*.