Local Generic Representation for Face Recognition with Single Sample per Person

Pengfei Zhu[1], Meng Yang[2], Lei Zhang[1⋆], and Il-Yong Lee[3,4]

[1]Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong
[2]Computer Vision Institute, School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[3]LG Electronics Institute of Technology, Korea
[4]Dept. of Computer Science, Yonsei University, Korea

Abstract. Face recognition with single sample per person (SSPP) is a very challenging task because in such a scenario it is difficult to predict the facial variations of a query sample by the gallery samples. Considering the fact that different parts of human faces have different importance to face recognition, and the fact that the intra-class facial variations can be shared across different subjects, we propose a local generic representation (LGR) based framework for face recognition with SSPP. A local gallery dictionary is built by extracting the neighboring patches from the gallery dataset, while an intra-class variation dictionary is built by using an external generic dataset to predict the possible facial variations (e.g., illuminations, pose, expressions and disguises). LGR minimizes the total representation residual of the query sample over the local gallery dictionary and the generic variation dictionary, and it uses correntropy to measure the representation residual of each patch. Half-quadratic analysis is adopted to solve the optimization problem. LGR takes the advantages of patch based local representation and generic variation representation, showing leading performance in face recognition with SSPP.

1 Introduction

Face recognition (FR) is a very active topic in computer vision research because of its wide range of applications, including access control, video surveillance, social network, photo management, criminal investigation, etc [1]. Though FR has been studied for many years, it is still a challenging task due to the many types of large face variations, e.g., pose, expressions, illuminations, corruption, occlusion and disguises. Furthermore, in applications such as smart cards, law enforcement, etc., we may have only one template sample of each subject, resulting in the single sample per person (SSPP) problem [2]. SSPP makes FR much more difficult because we have little information from the gallery set to predict the variations in the query face image [3].

Since the intra-class variations cannot be well estimated in the SSPP problem, the traditional discriminative subspace learning based FR methods can fail

⋆ Corresponding author. Email: cslzhang@comp.polyu.edu.hk

to work. In addition, since the number of samples per class is so small, the robustness of extracted features and the generalization ability of learned classifiers can be much reduced. To alleviate these difficulties of FR with SSPP, researchers have proposed to generate virtual samples of each subject, extract more discriminative features, and learn the facial variations from external data, etc. Generally speaking, the existing FR methods for SSPP can be categorized into three groups: virtual sample generation, generic learning and patch/block based methods.

Virtual sample generation aims to estimate the intra-class face variations by simulating extra samples for each subject. Virtual samples can be generated by perturbation-based approaches [4], geometric transform and photometric changes [5], SVD decomposition [6] and 3D methods [7], etc. With the virtual samples, intra-class scatter can be calculated to make Fisher linear discriminant analysis feasible in the scenario of SSPP [4][5][6]. Although virtual samples are helpful to FR with SSPP, they are highly correlated with the original face images and cannot be considered as independent samples for feature extraction. Therefore, there may exist much redundancy in the learned discriminative feature subspace [4][8].

Considering the similarity of face images across subjects, a generic training set can be used to compensate for the shortage of samples in FR. On one hand, the face variation information in the generic training set can be used to learn a projection matrix to extract discriminative features [9][10][11][12]. In [9] and [12], discriminative pose-invariant and expression-invariant projection matrices are learned by using a collected generic training set for pose-invariant and expression-invariant FR tasks, respectively. On the other hand, the abundant intra-class variations in the generic training set are very useful to more accurately represent a query face with unknown variations [13][3][14]. The sparse representation based classification (SRC) [15] represents a query face as a sparse linear combination of training samples from all classes. SRC shows interesting FR results; however, its performance will deteriorate significantly when the number of training samples of each class is very small because in such cases the variation space of each subject cannot be well spanned. The extended SRC (ESRC) [13] constructs an intra-class variation dictionary to represent the changes between the gallery and query images. In the case of SSPP, Yang et al. [3] learned a sparse variation dictionary by taking the relationship between the gallery set and the external generic set into account. The so-called sparse variation dictionary learning (SVDL) scheme shows state-of-the-art performance in FR with SSPP. However, SVDL ignores the distinctiveness of different parts of human faces.

Patch/block based methods [16][8][17][18] [19] partition each face image into several patches/blocks, and then perform feature extraction and classification on them. First, patches can be viewed as independent samples for feature extraction [16][8]. In [16], the patches of each subject are considered as the samples of this class and then the within-class scatter matrix can be computed. In [8], the patches of each subject are considered to form a manifold and a projection matrix is learned by maximizing the manifold margin. Second, a weak classifier can be

obtained from each patch, and then the classifiers on all patches can be combined to output the final decision (i.e., a strong classifier) [17][18]. In [17], the nearest neighbor classifier (NNC) is used for classification on each patch, and a kernel plurality method is proposed to combine the decisions on all patches. In [18], the collaborative representation based classifier (CRC) [20] is applied to each patch, and the majority voting is used for decision combination. Although the patch based methods in [17] and [18] significantly improve the FR performance compared with the original NNC and CRC classifiers, respectively, they do not solve the problem of lacking facial variations in the gallery set.

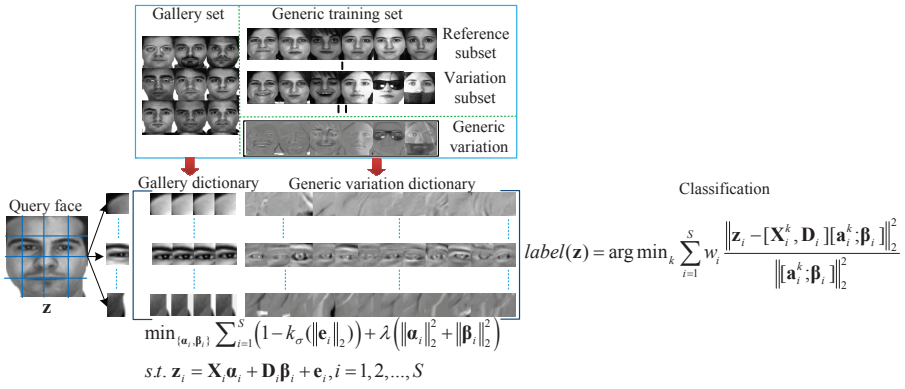

Fig. 1. Framework of local generic representation based classification. Gallery set is composed of the training face images. Generic training set includes reference subset and variation subset, while reference subset is composed of the neutral face images or the mean faces of each subject, and variation subset is composed of face images with different facial variations.

In this paper, we propose a local generic representation (LGR) based scheme for FR with SSPP, whose framework is illustrated in Fig. 1. The training samples in the gallery set are used to build a gallery dictionary. To introduce the face intra-class variation information that is lacked in the gallery set, a generic training set, which contains a reference subset and several variation subsets, is collected. A generic variation dictionary is then constructed as the difference between the reference subset and the variation subsets. Considering the different importance of different facial parts in FR, we adopt a local representation approach, i.e., each patch of the query sample is represented by the patch gallery dictionary and patch variation dictionary at the corresponding location. LGR aims to minimize the total representation residual of all patches. Since the residuals are non-Gaussian distributed, we use correntropy to measure the loss in minimization. The half-quadratic optimization technique is used to solve the optimization problem. Finally, the classification is performed based on the overall representation residual of the query sample by each class. The experimental results on benchmark face databases, including Extended Yale B [21], CMU Multi-PIE [22], AR [23] and LFW [24], show that LGR outperforms many state-of-the art methods for FR with SSPP.

2 Local generic representation

2.1 Generic representation

In FR with SSPP, we have a gallery set $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_k, ..., \boldsymbol{x}_K] \in \mathbb{R}^{d \times K}$, where $\boldsymbol{x}_k \in \mathbb{R}^d$ is the only single gallery sample of class k, $k = 1, 2, ..., K$. Given a query sample $\boldsymbol{z} \in \mathbb{R}^d$, representation based classifiers such as SRC [15] represent it over the gallery set $\boldsymbol{X}$ as:

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{e} \tag{1}$$

If the gallery set has many training samples for each subject, most of the facial variations in the query sample can be synthesized by the multiple samples from the same class, and consequently correct classification can be made via comparing the representation residual of each class. For FR with SSPP, unfortunately, there is only one training sample per subject, and the variations (e.g., illumination, pose, expression, etc.) in $\boldsymbol{z}$ cannot be well represented by the single same-class sample in $\boldsymbol{X}$. Thus, the representation residual of $\boldsymbol{z}$ can be big, and $\boldsymbol{z}$ can be wrongly represented by samples from other classes, leading to misclassification of $\boldsymbol{z}$. Fig. 2(a) shows an example. The query image has some illumination change compared with the single gallery sample of its class. We use the SRC model to solve the representation in Eq. (1), i.e., $\min_{\boldsymbol{\alpha}} \|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$. One can see from Fig. 2(a) that the synthesized image $\boldsymbol{X}\boldsymbol{\alpha}$ does not overcome the problem of illumination change, and the illumination change is put forward into the representation residual $\boldsymbol{e}$. Such a representation will cause trouble in the classification stage.

Fig. 2. Sparse representation versus generic representation.

Considering that the intra-class facial variations caused by illumination, pose, and expression changes and disguise can be shared across subjects, an external generic training set which consists of enough face images with various types of variations can be adopted to construct an intra-class variation dictionary [13][3]. Suppose that we have collected a generic training set $\boldsymbol{G} = [\boldsymbol{G}^r, \boldsymbol{G}^v]$, where $\boldsymbol{G}^r$ and $\boldsymbol{G}^v$ are the reference subset and variation subset, respectively. The reference subset $\boldsymbol{G}^r \in \mathbb{R}^{d \times n}$ is composed of neutral face images or the mean faces of each subject. The variation subset $\boldsymbol{G}^v$ involves M possible facial variations: $\boldsymbol{G}^v = [\boldsymbol{G}_1^v, ..., \boldsymbol{G}_m^v, ..., \boldsymbol{G}_M^v]$, where $\boldsymbol{G}_m^v$ is the subset of the m^{th} variation, $m = 1, 2, ..., M$. In [3], a sparse variation dictionary is learned from $\boldsymbol{G}$. In our work, we simply construct an intra-class variation dictionary, denoted by $\boldsymbol{D}$, by using the difference between $\boldsymbol{G}^r$ and $\boldsymbol{G}^v$:

$$\boldsymbol{D} = [\boldsymbol{G}_1^v - \boldsymbol{G}^r, ..., \boldsymbol{G}_m^v - \boldsymbol{G}^r, ..., \boldsymbol{G}_M^v - \boldsymbol{G}^r] \in \mathbb{R}^{d \times nM} \tag{2}$$

We then propose to represent the query sample $\boldsymbol{z}$ over the gallery set $\boldsymbol{X}$ and the generic variation dictionary $\boldsymbol{D}$ simultaneously:

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{e} \tag{3}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the representation vectors of $\boldsymbol{z}$ over $\boldsymbol{X}$ and $\boldsymbol{D}$, respectively, and $\boldsymbol{e}$ is the representation residual. We call the representation in Eq. (3) generic representation, which uses a generic intra-class variation dictionary $\boldsymbol{D}$ to account for the variations in the query sample. Fig. 2(b) shows the generic representation of the query sample in Fig. 2(a). We use the following model to solve Eq. (3): $\min_{\{\boldsymbol{\alpha},\boldsymbol{\beta}\}} \|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{D}\boldsymbol{\beta}\|_2^2 + \lambda(\|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{\beta}\|_1)$. One can clearly see that the illumination change in the query sample is well encoded by the generic variation dictionary $\boldsymbol{D}$, and the residual $\boldsymbol{e}$ has much lower energy ($\|\boldsymbol{e}\|_2^2$=0.0049) than the residual in Fig. 2(a) ($\|\boldsymbol{e}\|_2^2$=0.0502).

2.2 Patch based local generic representation

Different parts (e.g., eye, mouth, nose, cheek) of human faces exhibit distinct structures, and they have different importance in identifying the identity of a face. Taking this fact into account, we propose to localize the representation model in Eq. (3) and present a patch based local generic representation scheme.

We partition the query sample $\boldsymbol{z}$ into S (overlapped) patches and denote these patches as $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_S\}$. Correspondingly, the gallery dictionary $\boldsymbol{X}$ and the generic variation dictionary $\boldsymbol{D}$ can be partitioned as $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_S\}$ and $\{\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_S\}$, respectively. For each local patch $\boldsymbol{z}_i, i = 1, 2, .., S$, its associated local gallery dictionary and local variation dictionary are $\boldsymbol{X}_i$ and $\boldsymbol{D}_i$, respectively. To increase the representation power of local gallery dictionaries and better address the local deformation (e.g., misalignment) of a patch, we extract the neighborhood patches at location i from each gallery sample, and add them to $\boldsymbol{X}_i$. Such a sample expansion of local gallery dictionaries can improve much the stability and robustness of local representation [18]. In our implementation, the 8 closet neighboring patches to the underlying patch at location i are extracted. With $\boldsymbol{X}_i$ and $\boldsymbol{D}_i$, we can represent each local patch $\boldsymbol{z}_i$ as:

$$\boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\alpha}_i + \boldsymbol{D}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, i = 1, 2, ..., S \tag{4}$$

where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are the representation vectors of $\boldsymbol{z}_i$ over $\boldsymbol{X}_i$ and $\boldsymbol{D}_i$, respectively, and $\boldsymbol{e}_i$ is the representation residual.

Clearly, in order to find meaningful solutions of vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$, appropriate loss function should be defined on the representation residual $\boldsymbol{e}_i$ and appropriate regularization can be imposed on $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. Denote by $l(\|\boldsymbol{e}_i\|_2)$ the loss function defined on the l_2-norm of $\boldsymbol{e}_i$ and denote by $R(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ some regularizer imposed on the representation coefficients. We consider the following optimization problem to solve $\{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}$:

$$\begin{array}{l} \min_{\{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i\}} \sum_{i=1}^{S} l(\|\boldsymbol{e}_i\|_2) + \lambda R(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i) \\ s.t.\ \boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{a}_i + \boldsymbol{D}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i,\ i = 1, 2, ..., S \end{array} \tag{5}$$

The problem now turns to how to define the loss function $l(\|\boldsymbol{e}_i\|_2)$ and regularizer $R(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$.

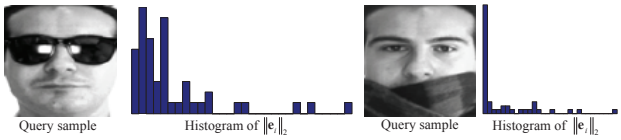

Fig. 3. The histogram of $\|\boldsymbol{e}_i\|_2, i = 1, 2, ..., S$, for two query samples.

Let $e_i = \|\boldsymbol{e}_i\|_2$. Due to the special structure of human face, the different patches will have very different representation residuals e_i. We solve $\{\hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\beta}}_i\} = \min_{\{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}} \|\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\alpha}_i + \boldsymbol{D}_i\boldsymbol{\beta}_i\|_2^2 + \lambda(\|\boldsymbol{\alpha}_i\|_2^2 + \|\boldsymbol{\beta}_i\|_2^2)$ and then calculate $e_i = \left\|\boldsymbol{z}_i - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_i + \boldsymbol{D}_i\hat{\boldsymbol{\beta}}_i\right\|_2$. Fig. 3 illustrates the distribution for e_i for two query face images. One can see that the distribution of e_i is highly non-Gaussian. The widely used l_2-norm loss function relies highly on the Gaussianity assumption of the data [25] and hence it is not suitable to measure such non-Gaussian distributed residual. In [26], the concept of correntropy is proposed to measure the loss of non-Gaussian data. A correntropy induced metric (CIM) for residual e_i is defined as [26]:

$$\mathrm{CIM}(e_i) = (k_\sigma(0) - k_\sigma(e_i))^{1/2} \tag{6}$$

where $k_\sigma(\cdot)$ is a kernel function. The Gaussian kernel function $k_\sigma(x) = \exp(-x^2/2\sigma^2)$ is widely used with good performance [26] [25]. The robustness of CIM to non-Gaussian residual/noise has been verified in signal processing [27], feature selection [28], and FR [29]. Hence, we adopt correntropy to model the representation residual of different patches.

For the regularizer $R(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$, we define it as the l_2-norm of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. It has been shown that the l_2-norm regularization on representation coefficients can lead to similar classification performance to l_1-norm regularization but with much less computational cost [20]. Finally, the proposed local generic representation (LGR) model becomes:

$$\begin{aligned} &\min_{\{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}} \textstyle\sum_{i=1}^{S} (1 - k_\sigma(\|\boldsymbol{e}_i\|_2)) + \lambda \left(\|\boldsymbol{\alpha}_i\|_2^2 + \|\boldsymbol{\beta}_i\|_2^2\right) \\ &s.t. \boldsymbol{z}_i = \boldsymbol{X}_i\boldsymbol{\alpha}_i + \boldsymbol{D}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, i = 1, 2, ..., S \end{aligned} \tag{7}$$

3 Optimization and classification

3.1 Half-quadratic optimization

The minimization problem in Eq. (7) can be solved by half-quadratic optimization [27]. If a function $\phi(x)$ satisfies the following conditions [27]: (a) $x \to \phi(x)$ is convex on $\mathbb{R}$; (b) $x \to \phi(\sqrt{x})$ is concave on $\mathbb{R}_+$; (c) $\phi(x) = \phi(-x), x \in \mathbb{R}$; (d) $x \to \phi(x)$ is C^1 on $\mathbb{R}$; (e) $\phi^{''}(0^+) > 0$; (f) $\lim_{x\to\infty}\phi(x)/\|x\|_2^2 = 0$, there exists a dual function φ such that

$$\phi(x) = \inf_{w\in\mathbb{R}} \left\{\tfrac{1}{2}wx^2 + \varphi(w)\right\} \tag{8}$$

where w is determined by the minimizer function $\delta(\cdot)$ with respect to $\phi(\cdot)$. $\delta(\cdot)$ admits an explicit form under certain restrictive assumptions [27]:

$$w = \begin{cases} \delta(t) = \phi^{''}(0^{+}), & \text{if } t = 0 \\ \phi^{''}(t)/t, & \text{if } t \neq 0 \end{cases} \tag{9}$$

Obviously, $\phi_\sigma(x) = 1 - k_\sigma(x) = 1 - \exp(-x^2/2\sigma^2)$ satisfies all the conditions from (a) to (f). Then the problem in Eq. (7) can be equivalently written as the following augmented minimization problem:

$$\min_{\boldsymbol{A},\boldsymbol{w}} \sum_{i=1}^{S} \left(\tfrac{1}{2} w_i \left\| \boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{D}_i\boldsymbol{\beta}_i \right\|_2^2 + \varphi(w_i) \right) + \lambda \left\| \boldsymbol{A} \right\|_2^2 \tag{10}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_S]$ with $\boldsymbol{a}_i = [\boldsymbol{\alpha}_i; \boldsymbol{\beta}_i]$, and $\boldsymbol{w} = [w_1, w_2, ..., w_S]$.

According to the half-quadratic analysis [27], Eq. (10) can be easily minimized by updating $\boldsymbol{A}$ and $\boldsymbol{w}$ alternatively, and there is no need to have an explicit form of the dual function $\varphi(w_i)$. When $\boldsymbol{w}$ is fixed, $\boldsymbol{A}$ can be solved by

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \sum_{i=1}^{S} \left(w_i \left\| \boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{D}_i\boldsymbol{\beta}_i \right\|_2^2 \right) + \lambda \left\| \boldsymbol{A} \right\|_F^2 \tag{11}$$

Clearly, the above minimization is a least square regression problem, and we have the closed-form solution of each $\{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}$:

$$[\hat{\boldsymbol{\alpha}}_i; \hat{\boldsymbol{\beta}}_i] = w_i (w_i [\boldsymbol{X}_i, \boldsymbol{D}_i]^T [\boldsymbol{X}_i, \boldsymbol{D}_i] + \lambda \boldsymbol{I})^{-1} [\boldsymbol{X}_i, \boldsymbol{D}_i]^T \boldsymbol{z}_i \tag{12}$$

When $\boldsymbol{A}$ is fixed, the weights $\boldsymbol{w}$ can be updated as

$$\hat{w}_i = \frac{1}{\sigma^2} \exp(- \left\| \boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{D}_i\beta_i \right\|_2^2 / 2\sigma^2) \tag{13}$$

The weight w_i corresponds to the i^{th} patch, and it is used to control the portion of $\left\| \boldsymbol{e}_i \right\|_2$ in the whole energy of Eq. (10). If the representation residual of a patch is big (e.g., caused by sunglasses, scarf and/or other large variations), the corresponding weight w_i will become small, and consequently the effect of this patch in the overall representation will be suppressed.

3.2 LGR based classification

After the optimal solutions of $\boldsymbol{A}$ and $\boldsymbol{w}$ are resolved by the half-quadratic optimization in Section 3.1, an LGR based classification scheme can be proposed to determine the class label of query face $\boldsymbol{z}$. Let $\boldsymbol{X}_i = \left[\boldsymbol{X}_i^1, ..., \boldsymbol{X}_i^k, ..., \boldsymbol{X}_i^K\right]$, where $\boldsymbol{X}_i^k$ is sub-gallery dictionary associated with class k. Accordingly, the representation vector $\boldsymbol{\alpha}_i$ can be written as $\boldsymbol{\alpha}_i = \left[\boldsymbol{\alpha}_i^1; ...; \boldsymbol{\alpha}_i^k; ...; \alpha_i^K\right]$, where $\boldsymbol{\alpha}_i^k$ is the coefficients vector associated with class k. By using the class-specific sub-gallery dictionary $\boldsymbol{X}_i^k$ and the generic variation dictionary $\boldsymbol{D}_i$, we can calculate the representation residual of each patch $\boldsymbol{z}_i$ by each class k. Then the sum of the weighted residual (by w_i) over all patches can be calculated. Our classification principle is to check which class can lead to the minimal residual over all patches. Specifically, the classification rule of query face $\boldsymbol{z}$ is as follows:

$$label(\boldsymbol{z}) = \arg\min_k \sum_{i=1}^{S} w_i \left\| \boldsymbol{z}_i - [\boldsymbol{X}_i^k, \boldsymbol{D}_i][\boldsymbol{a}_i^k; \boldsymbol{\beta}_i] \right\|_2^2 / \left\| [\boldsymbol{a}_i^k; \boldsymbol{\beta}_i] \right\|_2^2 \quad (14)$$

Note that in Eq. (14), we also use the l_2-norm of $[\boldsymbol{a}_i^k; \boldsymbol{\beta}_i]$ to adjust the residual of patch i by class k. $1/\left\| [\boldsymbol{a}_i^k; \boldsymbol{\beta}_i] \right\|_2^2$ can be considered as a "class weight". If class k has a larger $\left\| [\boldsymbol{a}_i^k; \boldsymbol{\beta}_i] \right\|_2^2$, it means that the query patch is more similar to the gallery patch of class k, and thus a smaller weight should be assigned to weaken the representation residual by this class. The query sample $\boldsymbol{z}$ is classified to the class which has the minimal weighted representation residual over all patches. The algorithm of LGR based classification is summarized in Table 1.

Table 1. The algorithm of local generic representation (LGR) based classification.

Input: The query sample $\boldsymbol{z}$, gallery set $\boldsymbol{X}$, reference subset $\boldsymbol{G}^r$, variation subset $\boldsymbol{G}^v$ and regularization parameter λ.
Output: The class label of $\boldsymbol{z}$
1: Initialize $\boldsymbol{w} = [1, 1, ..., 1]$;
2: Caculate $\boldsymbol{D} = [\boldsymbol{G}_1^v - \boldsymbol{G}^r, \boldsymbol{G}_2^v - \boldsymbol{G}^r, ..., \boldsymbol{G}_m^v - \boldsymbol{G}^r]$.
3: Partition $\boldsymbol{z}$, $\boldsymbol{X}$ and $\boldsymbol{D}$ into patches.
4: While convergence
5: Update $\boldsymbol{A}$ by Eq. (11);
6: Update $\boldsymbol{w}$ by Eq. (13);
7: End
8: Output the class label of sample $\boldsymbol{z}$ by Eq.(14).

3.3 Convergence and complexity

According to half-quadratic optimization [27], the objective function in Eq. (10) is non-increasing under the update rules in Eq.(11) and Eq. (13). Therefore, our algorithm is guaranteed to converge based on the theory of half-quadratic optimization [27]. In Fig.4, the convergence curve of LGR on the AR database [23] is shown (please refer to section 4.4 for the details of experiment setting). We can see that the LGR algorithm converges after 5 iterations.

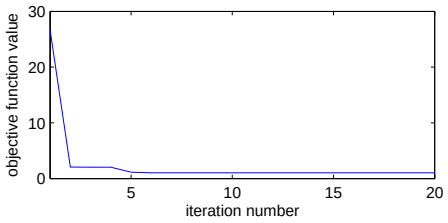

Fig. 4. The convergence curve of LGR on the AR database.

The main computational cost of LGR is spent on solving the least square regression problem in Eq. (11), whose time complexity is $O(S(n_d^3 + n_d^2 d_p))$, where S is the number of patches, n_d is the total number of patches in $[\boldsymbol{X}_i, \boldsymbol{D}_i]$ and d_p is the feature dimension of patches. Denote by T the total number of iteration in our algorithm, the time complexity of LGR is $O(TS(n_d^3 + n_d^2 d_p))$.

4 Experimental analysis

We test the performance of LGR on four benchmark face databases, including three face databases in controlled environment, i.e., Extended Yale B [21], large-scale CMU Multi-PIE [22], and AR [23], and one face database in uncontrolled environment, i.e., Labeled Faces in the Wild (LFW) database [24]. Extended Yale B database contains illumination variations; AR database contains illumination and expression variations and disguises; Multi-PIE database contains pose, illumination and expression variations; LFW reflects the variations in real-world applications. We compare the proposed LGR method with the following eleven methods:

- Baseline methods: nearest neighbor classifier (NNC) [30], support vector machines (SVM) [31], sparse representation based classifiers (SRC) [15] and collaborative representation based classifiers (CRC) [20];
- Generic learning methods: adaptive generic learning (AGL) [32], extended SRC (ESRC) [13] and sparse variation dictionary learning (SVDL) [3];
- Patch/block based methods: Block linear discriminative analysis (BlockLDA) [16], patch based NN (PNN) [17], patch based CRC (PCRC) [18], and discriminative multi-manifold analysis (DMMA) [8].

Note that the generic learning method SVDL learns a sparse variation dictionary from the generic training set. The proposed LGR also belongs to the generic learning methods; however, we use the raw face difference images as the dictionary rather than learning a dictionary with some objective function. Among the competing methods, we implement NN and DMMA; the code of SVM is from [33]; and the codes of all the other methods are obtained from the original authors.

4.1 Parameter setting

In all the experiments, the face images are resized to 80×80 (using the Matlab function "resize.m"). For patch/block based methods including BlockLDA, PNN, PCRC, DMMA, and the proposed LGR, the patch size is fixed as 20×20 and the overlap between neighboring patches is 10 pixels. That is, the query sample is partitioned into S=49 patches.

Apart from the setting of patch size and patch number, there are only two parameters to set in the proposed LGR. The first is the regularization parameter λ in Eq. (6). We fix it as λ= 0.001 in all our experiments. Another is the scale parameter σ of the kernel function $k_\sigma(x)$. Based on our experimental experience, if the representation residual is big, a large value of σ could be set to make the representation more robust. Therefore, we adaptively set σ as the average representation residual after solving the coefficients $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ in the first iteration of our algorithm; that is, $\sigma = \sqrt{\frac{1}{2S}\sum_{i=1}^{S} \|\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{D}_i\boldsymbol{\beta}_i\|_2^2}$.

For the competing algorithms, we tune their parameters for the best results. In particular, for SVDL we follow the parameter setting in [3]. The three parameters λ_1, λ_2, λ_3 are set as 0.001, 0.01, 0.0001, respectively, and the number of

dictionary atoms is set as 400 in the initialization. For SRC, CRC and PCRC, the optimal regularization parameter λ is chosen from $\{0.0005, 0.001, 0.005, 0.01\}$. As BlockLDA and AGL are sensitive to the feature dimension, the best result of different feature dimensions is reported.

4.2 Extended Yale B database

The Extended Yale B face database [21] contains 38 human subjects and 2,414 face images with 64 illumination conditions. The frontal faces with light source directions at 0 degree azimuth (A+000) and at 0 degree elevation (E+00) are used as the gallery set, and the face images under other illumination conditions are used as the query set. We use the face images of the first 30 subjects to form the gallery and query sets, and use the face images of the other 8 subjects as the generic set.

Table 2 lists the recognition rates by different methods. By combining the decisions of different patches, the PCRC method achieves much higher recognition rate than the baseline methods. The generic learning based method SVDL achieves the second highest recognition rate by learning a dictionary that consists of different illumination variations. By exploiting the advantages of both patch based local representation and generic variation information, the proposed LGR method achieves the highest recognition accuracy.

Table 2. Recognition rate (%) on Extended Yale B database.

Method	NNC[30]	SVM[31]	SRC[15]	CRC[20]	BlockLDA[16]	AGL[32]
Accuracy	46.5	41.4	49.2	51.2	49.2	59.5
Method	DMMA[8]	PNN[17]	PCRC[18]	ESRC[13]	SVDL[3]	LGR
Accuracy	61.7	67.5	77.8	67.9	85.0	**86.6**

4.3 CMU Multi-PIE database

The Multi-PIE database [22] contains a total of more than 750,000 images from 337 individuals, captured under 15 viewpoints and 19 illumination conditions in four recording sessions. The face images of the first 100 subjects in session 1 are used for the gallery set and the other 149 subjects are used as generic set. Following the experiment setting in [3], in the generic training set, the frontal images with illumination 7 and neutral expression are used as the reference subset and the face images with different variations in Session 1 are used as the variation subset.

Illumination variations In this experiment, we test the performance of LGR under different illuminations. The frontal face images with neutral expression from session 2, session 3 and session 4 are used as the query set, respectively. The recognition rates on Multi-PIE with illumination variations are listed in Table 3. LGR shows superior performance to all the other competing methods. Compared with SVDL, which achieves the second highest accuracy, the recognition rate is improved by 2.7%, 3.0% and 4.0% on session 2, session 3 and

session 4, respectively. Compared with PCRC, the recognition rate is improved by about 15%. The performance of SRC and CRC is very poor because with only one gallery face image per person, the query image cannot be well represented.

Table 3. Recognition accuracy (%) on Multi-PIE with illumination variations.

Method	Session 2	Session 3	Session 4
NNC[30]	44.3	40	43.8
SVM[31]	43.6	40.5	40.1
SRC[15]	51.9	46.5	50.6
CRC[20]	52.8	47.4	50.5
BlockLDA[16]	68.2	60.4	65.1
AGL[32]	84.5	79.6	78.5
DMMA[8]	64.1	56.6	60.1
PNN[17]	65.1	55.6	60.8
PCRC[18]	83.7	72.7	77.7
ESRC[13]	92.6	84.6	87.6
SVDL[3]	94.2	87.5	90.4
LGR	**96.9**	**90.5**	**94.4**

Expression and illumination variations We then test the robustness of the proposed LGR method to face images with both expression and illumination variations. The query set includes the frontal face images with smile expression in session 1 (Smile-S1), smile expression in session 3 (Smile-S3) and surprise expression (Surprise-S2). Table 4 presents the recognition results in this experiment. Clearly, LGR outperforms all the other methods. SVDL still works the second best, but it lags behind LGR by 1.8%, 5.6% and 21.7% for Smile-S1, Smile-S3 and Surprise-S2, respectively.

Table 4. Recognition accuracy (%) on Multi-PIE with expression and illumination variations.

Method	Smile-S1	Smile-S3	Surprise-S2
NNC[30]	46.8	29.1	18.3
SVM[31]	46.8	29.1	18.3
SRC[15]	50.1	28.1	21.1
CRC[20]	50	29.7	22.4
BlockLDA[16]	49.5	30	26.2
AGL[32]	85.2	39.5	31.5
DMMA[8]	58.5	33.4	23
PNN[17]	53.1	31.1	31.4
PCRC[18]	74.9	44.1	44.9
ESRC[13]	82	50.8	49.9
SVDL[3]	88.9	59.6	52.8
LGR	**90.7**	**65.2**	**74.5**

Pose, expression and illumination variations In this experiment, there are pose, expression and illumination variations in the query set simultaneously. We select the face images with pose 05_0 in Session 2 (P1), pose 04_1 in Session 3

(P2), and pose 04_1 and smile expression in Session 3 (P3) as the query set. Some face images from the gallery and query set are illustrated in Fig. 5.

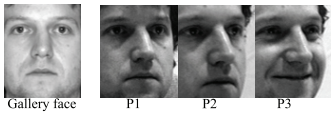

Fig. 5. Images of Multi-PIE database with pose, expression and illumination variations.

Table 5 lists the recognition rate of all methods. LGR achieves the highest accuracy on all the three query sets. Because of the large variations caused by pose, expression and illumination variations, the FR rates in this experiment are relatively lower than the experimental results in Table 3 and Table 4. The patch based methods such as PCRC do not work well because they are sensitive to pose variation. The generic learning methods, including AGL, ESRC, SVDL and the proposed LGR, outperform the other methods since they can exploit the variation information from the external generic training set. LGR consistently exhibits better results than SVDL, which still works the second best.

Table 5. Recognition accuracy (%) on Multi-PIE with pose, expression and illumination variations.

Method	P1	P2	P3
NNC[30]	25.7	8.8	11.9
SVM[31]	25.7	8.8	11.9
SRC[15]	23.9	6.1	10.1
CRC[20]	24.9	5.4	9.0
BlockLDA[16]	29.5	13.2	15.8
AGL[32]	66.4	25.5	24.0
DMMA[8]	28.2	5.5	12.1
PNN[17]	35.3	11.8	13.5
PCRC[18]	37.3	8.0	10.2
ESRC[13]	63.8	31.9	27.0
SVDL[3]	76.0	37.9	33.5
LGR	**79.1**	**39.5**	**36.3**

4.4 AR face database

The AR face database [23] contains about 4,000 color face images of 126 people, which consists of the frontal faces with different facial expressions, illuminations and disguises. There are two sessions and each session has 13 face images per subject. Following the SSPP experiment setting in [13], a subset with face images of 50 males and 50 females is selected. The first 80 subjects from sessions 1 are used for the gallery and query set while the other 20 subjects are used as the generic training set. We also use the face images from session 2 as the query set to test the FR performance. There are different variations, including illumination, expression, and disguise (scarf and sunglass) in this experiment.

The experimental results on session 1 and session 2 are shown in Table 6 and Table 7, respectively. LGR exhibits significantly better performance than all the

other methods on both sessions. In particular, on session 2 LGR outperforms SVDL by 16.4%, 10.8%, 32.5% and 34.7% under different variations. Note that in this experiment the performance of patch based methods such as PCRC is very competitive. This is because the disguises (i.e., scarf and sunglass) can be well dealt with by patch/block based methods. Therefore, PCRC can achieve higher recognition rate than the global representation based SVDL though it does not learn any variation information from a generic dataset. The proposed LGR utilizes both local presentation and generic information, leading to very promising performance for the task of FR with SSPP.

Table 6. Recognition accuracy (%) on AR face database (session1).

Method	illumination	expression	disguise	illumination+disguise
NNC[30]	70	79.2	39.4	23.5
SVM[31]	55.8	90.4	43.1	29.4
SRC[15]	80.8	85.4	55.6	25.3
CRC[20]	80.5	80.4	58.1	23.8
BlockLDA[16]	75.3	81.4	65.4	53.5
AGL[32]	93.3	77.9	70.0	53.8
DMMA[8]	92.1	81.4	46.9	30.9
PNN[17]	84.6	86.7	90.0	72.5
PCRC[18]	95.0	86.7	95.6	81.3
ESRC[13]	99.6	85.0	83.1	68.6
SVDL[3]	98.3	86.3	86.3	79.4
LGR	**100**	**97.9**	**98.8**	**96.3**

Table 7. Recognition accuracy (%) on AR face database (session2).

Method	illumination	expression	disguise	illumination+disguise
NNC[30]	41.7	58.8	26.3	12.8
SVM[31]	40.0	58.8	26.9	14.4
SRC[15]	55.8	68.8	29.4	12.8
CRC[20]	55.8	69.6	35.0	13.5
BlockLDA[16]	54.7	61.2	31.9	21.0
AGL[32]	70.8	55.8	40.6	30.7
DMMA[8]	77.9	61.7	28.1	21.9
PNN[17]	77.5	73.8	71.9	52.8
PCRC[18]	88.8	71.7	81.8	63.1
ESRC[13]	87.9	70.4	59.4	45.0
SVDL[3]	87.1	74.2	61.3	54.1
LGR	**97.5**	**85.0**	**93.8**	**88.8**

4.5 LFW database

The LFW database [24] contains images of 5,749 different individuals in unconstrained environment. LFW-a is a version of LFW after alignment using commercial face alignment software [34]. Following the experiment setting in [18] and [3], a subset of 158 subjects with more than 10 images per person is collected. Each face image is cropped to 120×120 and then resized to 80×80. One

can see that although face alignment has been conducted, the variations in this database is still very large compared with the face databases in the controlled environment. Face images of the first 50 subjects are selected to form the gallery and query sets, while the face images of the remaining subjects are used to build the generic training set. Since there are no frontal neutral face images in this database, the mean face of each person is used to form the reference subset in the generic set.

The face recognition rates of different methods are listed in Table 8. Because of the challenging face variations in uncontrolled environment, no method achieves very high accuracy in this experiment. Nonetheless, LGR still works the best among all competing methods. The patch based method PCRC works better than the global representation based CRC, which is similar to what we observed in the experiments of previous sections. SVDL again achieves the second highest recognition rate, demonstrating that the face variation information learned from other subjects is indeed helpful to improve the robustness of FR with SSPP, no matter in controlled or uncontrolled environment.

Table 8. Recognition accuracy (%) on LFW database.

Method	NNC[30]	SVM[31]	SRC[15]	CRC[20]	BlockLDA[16]	AGL[32]
Accuracy	12.2	11.6	20.4	19.8	16.4	19.2
Method	DMMA[8]	PNN[17]	PCRC[18]	ESRC[13]	SVDL[3]	LGR
Accuracy	17.8	17.6	24.2	27.3	28.6	**30.4**

5 Conclusions

We proposed a local generic representation (LGR) based approach for the challenging task of face recognition with single sample per person (SSPP). LGR utilizes the advantages of both patch based local representation and generic learning. A generic intra-class variation dictionary was constructed from a generic dataset, and it can well compensate for the face variations lacked in the SSPP gallery set. A patch gallery dictionary was built by using the gallery samples, which can more accurately represent the different parts of face images. Considering that the distribution of representation residual of different patches is highly non-Gaussian, a correntropy based metric was adopted to measure the loss of each patch so that the importance of different patches in face recognition can be more robustly evaluated. As a result, LGR can adaptively suppress the role of patches with large variations. The extensive experimental results on four benchmark face databases showed that LGR always achieves higher face recognition rate than the state-of-the-art SSPP methods used in competition.

Acknowledgement. This work was (partially) supported by LG Electronics Co., Ltd.

References

1. Jain, A.K., Li, S.Z.: Handbook of face recognition. Springer (2005)
2. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Recognition **39** (2006) 1725–1745
3. Yang, M., Gool, L.V., , Zhang, L.: Sparse variation dictionary learning for face recognition with a single training sample per person. In: Proc. 14th IEEE International Conf. Computer Vision (ICCV). (2013) in press.
4. Martínez, A.M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24** (2002) 748–763
5. Shan, S., Cao, B., Gao, W., Zhao, D.: Extended fisherface for face recognition from a single example image per person. In: Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on. Volume 2., IEEE (2002) II–81
6. Gao, Q.x., Zhang, L., Zhang, D.: Face recognition using flda with single training image per person. Applied Mathematics and Computation **205** (2008) 726–734
7. Vetter, T.: Synthesis of novel views from a single face image. International Journal of Computer Vision **28** (1998) 103–116
8. Lu, J., Tan, Y.P., Wang, G.: Discriminative multimanifold analysis for face recognition from a single training sample per person. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35** (2013) 39–51
9. Kim, T.K., Kittler, J.: Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27** (2005) 318–327
10. Wang, J., Plataniotis, K.N., Lu, J., Venetsanopoulos, A.N.: On solving the face recognition problem with one training sample per subject. Pattern recognition **39** (2006) 1746–1762
11. Kan, M., Shan, S., Su, Y., Xu, D., Chen, X.: Adaptive discriminant learning for face recognition. Pattern Recognition **46** (2013) 2497–2509
12. Mohammadzade, H., Hatzinakos, D.: Projection into expression subspaces for face recognition from single sample per person. Affective Computing, IEEE Transactions on **4** (2013) 69–82
13. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34** (2012) 1864–1870
14. Huang, D.A., Wang, Y.C.F.: With one look: robust face recognition using single sample per person. In: Proceedings of the 21st ACM international conference on Multimedia, ACM (2013) 601–604
15. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31** (2009) 210–227
16. Chen, S., Liu, J., Zhou, Z.: Making flda applicable to face recognition with one sample per person. Pattern recognition **37** (2004) 1553–1555
17. Kumar, R., Banerjee, A., Vemuri, B.C., Pfister, H.: Maximizing all margins: Pushing face recognition with kernel plurality. In: Computer Vision (ICCV), 2011 IEEE International Conference on. (2011) 2375 –2382
18. Zhu, P., Zhang, L., Hu, Q., Shiu, S.C.: Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In: Computer Vision–ECCV 2012. Springer (2012) 822–835

19. Kumar, R., Banerjee, A., Vemuri, B.C.: Volterrafaces: Discriminant analysis using volterra kernels. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 150–155
20. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: Int. Conf. on Comput. Vis. (2011)
21. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23** (2001) 643–660
22. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28** (2010) 807–813
23. Martinez, A.: The ar face database. CVC Technical Report **24** (1998)
24. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)
25. Lu, C., Tang, J., Lin, M., Lin, L., Yan, S., Lin, Z.: Correntropy induced l2 graph for robust subspace clustering. In: Proc. 14th IEEE International Conf. Computer Vision (ICCV). (2013) in press.
26. Liu, W., Pokharel, P.P., Príncipe, J.C.: Correntropy: properties and applications in non-gaussian signal processing. Signal Processing, IEEE Transactions on **55** (2007) 5286–5298
27. Nikolova, M., Ng, M.K.: Analysis of half-quadratic minimization methods for signal and image recovery. SIAM Journal on Scientific computing **27** (2005) 937–966
28. He, R., Tan, T., Wang, L., Zheng, W.S.: $l_{2,1}$-regularized correntropy for robust feature selection. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2504–2511
29. He, R., Zheng, W.S., Tan, T., Sun, Z.: Half-quadratic-based iterative minimization for robust sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **36** (2014) 261–275
30. Cover, T., Hart, P.: Nearest neighbor pattern classification. Information Theory, IEEE Transactions on **13** (1967) 21–27
31. Cortes, C., Vapnik, V.: Support vector machine. Machine learning **20** (1995) 273–297
32. Su, Y., Shan, S., Chen, X., Gao, W.: Adaptive generic learning for face recognition from a single sample per person. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2699–2706
33. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011) 27
34. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Computer Vision–ACCV 2009. Springer (2010) 88–97