Uniform Generic Representation for Single Sample Face Recognition

YUHUA DING, FAN LIU (Member, IEEE), ZHENMIN TANG, AND TAO ZHANG

1School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2College of Computer and Information, Hohai University, Nanjing 210098, China
3School of Information Engineering, Yangzhou University, Yangzhou 225127, China
4Key Laboratory of Ministry of Education for Coastal Disaster and Protection, Hohai University, Nanjing 210098, China

Corresponding author: Fan Liu (fanliu@hhu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602150, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191298, and in part by the Key Laboratory of Ministry of Education for Coastal Disaster and Protection, Hohai University, under Grant 201905.

Abstract In this article, we propose a uniform generic representation (UGR) method to solve the single sample per person (SSPP) problem in face recognition, which aims to find consistency between the global and local generic representations. For the local generic representation, we require the probe patches of the same image to be constructed respectively by the corresponding patches of the same gallery image and the intra-class variation dictionaries. Therefore, the probe patches’ coefficients, corresponding to patch gallery dictionaries, should be similar to each other. For the global generic representation, the probe image’s coefficient, corresponding to the gallery dictionary, should be similar to those of its probe patches. In order to meet the two requirements, we combine local generic representation with global generic representation in soft form. We obtain the representation coefficients by solving a simple quadratic optimization problem. UGR has been evaluated on Extended Yale B, AR, CMU-PIE, and LFW databases. Experimental results show the robustness and effectiveness of our method to illumination, expression, occlusion, time variation, and pose.

Index Terms Consistency, face recognition, single sample per person, uniform generic representation.

I. INTRODUCTION

In the past few decades, face recognition (FR) has been one of the most popular research topics in computer vision and pattern recognition due to its potential application. Researchers have carried out extensive research on FR and made remarkable progress [1]. FR technology under controllable conditions has achieved satisfactory results. However, under uncontrollable circumstances, FR is still challenging due to the influence of illumination, expression, posture, occlusion, and other factors. The direct way to solve these problems is to increase training samples. But in some real scenes, such as access control, e-passport, identity card verification, judicial confirmation, etc., usually only one training sample can be obtained, which is the so-called single sample per person (SSPP) problem [2]. In that case, though some famous FR methods [3]–[5] can still be applied, they suffer from serious performance drop. And linear discriminant analysis (LDA)-based methods [6], [7] even fail to work since they are all designed assuming that multiple samples per person are available. The absence of facial variation information makes it more challenging to fulfill the FR task.

Recently, many specially designed FR methods have been proposed to solve the SSPP problem, such as singular value decomposition (SVD)-based perturbation algorithms [8], [9] and randomized sampling-based algorithm [10]. The three methods generate multiple virtual images based on the original image. They then extract the features by applying LDA, which makes the learned features redundant, because substantial correlation exists among the generated images. And other methods like discriminative multi-manifold analysis (DMMA) [11], sparse discriminative multi-manifold embedding (SDMME) [12], and robust heterogeneous discriminative analysis (RHDA) [13] obtain multiple samples for each subject by dividing each gallery image into a collection of patches. They treat each subject as a manifold and learn multiple discriminative feature spaces. These methods can work well when the probe image does not contain extreme variation...
compared with the gallery image, which is unrealistic. All of the above approaches have the same problem of being unable to handle facial variations.

To deal with the lack of the intra-class variation information, an additional generic training dataset is introduced. Deng et al. [14] proposed a prototype plus variation (P + V) model to cope with the large variation between the probe and gallery images. Zhu et al. [15] proposed a local generic representation (LGR) by using correntropy to minimize the loss since they found that the representation residuals are non-Gaussian distributed. Gao et al. [16] presented the idea that the patches of a probe image should be constructed by the patches of the corresponding gallery image. Thus, they proposed a regularized patch-based representation (RPR) by imposing a group sparsity constraint on the reconstruction coefficients. The size of the original variation dictionary from the generic training dataset is usually very large, which will result in expensive computational cost when solving the optimization problem. Therefore, the authors of [16]–[19] proposed to learn a variation dictionary, respectively. Benefiting from the external variation information, these generic learning-based methods boost the performance of FR with SSPP significantly. However, these methods only focus on a single scale of the image. In fact, images or patches on different scales can provide complementary information, which is neglected by existing methods.

In this article, we propose a novel method to solve the SSPP problem, named uniform generic representation (UGR), which aims to unify the local representation and global representation. Like the common local generic representation, each image is divided into overlapped patches. And then, each probe patch is represented by the patch gallery dictionary and patch variation dictionary at the corresponding location. Ideally, the probe patches of the same image should be constructed by the patches from the corresponding gallery image, i.e., the representation coefficients of the probe patches of the same image, corresponding to patch gallery dictionary, should be equal to each other, no matter what the size of the patch is. And from the view of global generic representation, the representation coefficient of the probe image, corresponding to gallery dictionary, should be equal to those of the patches of the same probe image in the local-representation case, which is illustrated in Fig. 1. Given this, we propose solving the optimization problems of local and global representations together to find the uniform coefficient. However, in reality, the constraint of finding the uniform coefficient is so strong as to make the non-discriminative or corrupted patches mislead the discriminative ones. Therefore, we relax the constraint by making the representation coefficients of patches close to the uniform coefficient instead of equal. The proposed representation can be regarded as a soft combination of local and global representations. In our method, both local and global representations can give full play to their advantages, and the consistency between local and global representations can be explored. After solving the optimization problem, the strategy of majority voting is adopted, i.e., classify each patch of an image and then decide the image’s label by the voting of all the patches’ classification results. Experiments on four public face databases are performed to demonstrate the effectiveness of the proposed method.

Our contributions are summarized as follows.

1) Global representation and local representation have their own advantages. The global representation has good robustness for recognizing those indiscriminative regions such as forehead and cheek, while the local representation can mitigate the effect of facial variations. In this article, we explore how to use global representation to guide local representation. We combine local generic representation with global generic representation in soft form. The consistency between local generic representation and global generic representation is explored for the first time. The soft combination also maintains local diversity.

2) Existing methods for dealing with the SSPP problem extract either local or global features. In our method, we make use of both local and global features. We evaluate our method by multiple means.

We organize the rest of this article as follows. In Section II, we introduce the related work. In Section III, we describe the proposed uniform generic representation. Section IV presents the experimental results and discussion. We conclude the paper in Section V.

II. RELATED WORK

When designing a FR system with SSPP, we must deal with three key issues: 1) The dimension of data is much larger than the number of training samples. 2) It is unable to calculate the within-class scatter matrix, and the obtained between-class scatter matrix becomes inaccurate. 3) It is difficult to describe the probe sample with unknown variations.
For the first issue, some scholars increase training samples by generating extra ones, which can also alleviate the second issue. Authors of [8], [9] generate multiple training samples for each person by applying SVD-based perturbation, respectively. Li et al. [10] proposed to get multiple sub-images from a single sample by randomized down-sampling. These methods of generating virtual samples do not gain information except from the original image. Some scholars tend to apply feature extraction to decrease the dimension of the image and meanwhile keep discriminative information. Wu and Zhou [20] proposed a projection-combined principal component analysis (E(PCA)) method. And Chen et al. [21] further proposed enhanced projection-combined principal component analysis (E(PCA)) method. The two methods only focus on mining more global information. In [22], the authors proposed to enhance the recognition ability by using the local descriptors (Gabor wavelets [23] and local binary patterns [24]). Zhou et al. [25] and Liu et al. [26] presented two bag-of-words feature-based methods to address the SSPP problem. The three methods [22, 25, 26] show that the local descriptors increase robustness to facial variation. Other scholars proposed to decrease the dimension of data by image partition. The methods using image partition are called patch-based methods, such as patch-based sparse representation for classification (PSRC) [5], patch-based collaborative representation for classification (PCRC) [27], and patch-based n-nearest classifier (PNN) [28]. In addition, Liu et al. [29] proposed to build a subspace for each patch and then explore the relationships between subspaces. Patch-based methods utilize local information, and most of them predict the image’s label based on majority voting. Therefore, the patch-based methods can mitigate the effect of facial variations.

For the second issue, one approach is to generate extra training samples for each individual, which is discussed above. Another approach is to divide each gallery image into patches and treat them as training samples of each class. Based on this conception, Chen et al. [30] proposed a block-based Fisher linear discriminative analysis (BlockFLDA) method. Lu et al. [11] treated each person as a manifold and formulated FR as a multi-manifold matching problem. Yan et al. [31] further proposed a multi-feature multi-manifold learning method. Zhang et al. [12] proposed to learn discriminative features by embedding two sparse graphs into multi-manifold learning. Combining the advantages of [11] and [12], Pang et al. [13] proposed to conduct discriminant analysis by introducing a graph-based Fisher-like criterion. Besides the two approaches mentioned here, there is another approach of dealing with the second issue. Chu et al. [32] proposed to get extra training samples by mirroring the right half of a face to the left and perform feature extraction at a half-face level.

For the third issue, the usual solution is to introduce an additional generic training dataset since the facial variations of different people share much similarity. For example, Deng et al. [14] proposed an extended SRC (ESRC) method by coding the difference between the gallery and probe samples with a generic training dataset. Su et al. [2] proposed a generic discriminant model to infer the variation information of the gallery dataset. Yang et al. [33] proposed to learn a sparse variation dictionary, which is adaptive to the gallery dataset. Ji et al. [34] proposed to construct an adaptive probabilistic label matrix for each probe sample. Deng et al. [22] proposed to embed equidistant prototypes into the linear regression model and map the intra-class differences calculated from the generic training dataset to zero vectors. The above five methods get the variation information in five different ways and have good robustness for recognizing those indiscriminative regions such as forehead and cheek as they all treat the entire image as a feature vector. There are also other methods that incorporate the generic variation information into patch-based methods, such as LGR [15] and RPR [16]. They can achieve better performance by taking advantage of the generic training dataset and image partition. Besides introducing a generic training dataset, Li et al. [35] proposed to capture the variation information by preserving the variations of each probe sample onto training samples.

It is generally known that deep learning-based methods have achieved great breakthroughs in the FR field. In view of the success of deep learning, some researchers proposed to utilize a deep model to deal with the SSPP problem. For instance, Gao et al. [36] proposed a stacked supervised auto-encoder (SSAE) by enforcing the faces with variations to be mapped with the corresponding canonical face. AbdElmaksoud et al. [37] proposed to augment the training data by synthesizing 3D faces from 2D face photos, and then train different networks to recognize faces. Tu et al. [38] proposed a 3D face modeling module to extract invariant features and generate 2D images with different variations. Hong et al. [39] proposed to jointly perform feature extraction, domain adaptation, and classification by applying a deep architecture with domain-adversarial training. Yang et al. [40] proposed to use convolutional neural networks (CNN) to extract the local adaptive features. Zeng et al. [41] proposed to utilize generated virtual samples to fine-tune a well-trained deep CNN model. Deep learning-based methods can dramatically improve the performance on the LFW (Labeled Faces in the Wild) database. However, they usually require tremendous auxiliary data to train the neural network, which will lead to high calculation cost, and it is not easy to tune the parameters.

III. UNIFORM GENERIC REPRESENTATION

In this section, we will first introduce the global and local generic representations and then propose the uniform generic representation. We also give the solution of the proposed optimization problem.

A. GLOBAL GENERIC REPRESENTATION

Under the SSPP condition, like [14], we assume that the facial intra-class variations of different people are similar, and the intra-class variant basis can be acquired from a sufficient number of generic faces. Based on this assumption, we use
additional generic training dataset to construct the intra-class variation dictionary. For clarity, we denote the single neutral images in the generic dataset by \( G = [g_1, \cdots, g_L] \in \mathbb{R}^{d \times L} \), where \( g_i \) corresponds to the neutral image from the \( i \)-th subject, and we denote the variation images by \( G^t \), where superscript \( t \) represents the type of variation. \( G^t \) is a counterpart of \( G \) with the variation type \( t \). Suppose that there are \( M \) types of variations in the generic dataset, then we construct the intra-class variation dictionary \( D \) by using the difference between \( G^t \) and \( G \):

\[
D = [G^1 - G, \cdots, G^M - G] \in \mathbb{R}^{d \times LM}
\]

(1)

We are given a gallery dataset \( \{a_1, \cdots, a_K\} \) with \( K \) classes, where \( a_k \) is the only single gallery image of the \( k \)-th subject. The gallery dataset is used as the prototype dictionary, denoted by \( A = [a_1, \cdots, a_K] \in \mathbb{R}^{d \times K} \). Then, the probe image \( x \) can be reconstructed by utilizing the prototype dictionary \( A \) and the intra-class variation dictionary \( D \), which is formulated as follows:

\[
x = A z + D s + e,
\]

(2)

where \( z \) and \( s \) are the reconstruction coefficients, and \( e \) is the residual as follows.

For multiple probe images, (2) can be rewritten as follows:

\[
X = AZ + DS + E,
\]

(3)

where \( X = [x_1, \cdots, x_N] \in \mathbb{R}^{d \times N} \) is the probe image matrix, \( Z = [z_1, \cdots, z_N] \in \mathbb{R}^{K \times N} \) and \( S = [s_1, \cdots, s_N] \in \mathbb{R}^{LM \times N} \) are the reconstruction coefficient matrices, and \( E \) is the residual matrix. The reconstruction coefficients can be obtained by solving the following optimization problem:

\[
\begin{align*}
\min_{Z,S} & \quad \|E\|_F^2 + \lambda_1\|Z\|_F^2 + \beta_1\|S\|_F^2 \\
\text{s.t.} & \quad X = AZ + DS + E
\end{align*}
\]

(4)

where \( \| \cdot \|_F \) denotes the Frobenius norm of a matrix, \( \lambda_1 > 0 \) and \( \beta_1 > 0 \) are the weighting parameters.

### B. LOCAL GENERIC REPRESENTATION

The local generic representation is a patch-based version of the global generic representation. We divide the prototype dictionary \( A \) into \( C \) overlapped patches, denoted by \( \{A^1, \cdots, A^C\} \), where \( A^i = [a^i_1, \cdots, a^i_{c_i}] \), and the superscript indicates the serial number of the patch. The probe images \( X \) and the intra-class variation dictionary \( D \) are divided respectively into \( \{X^1, \cdots, X^C\} \) and \( \{D^1, \cdots, D^C\} \) in the same way.\n
\[
X^i = A^i Z^i + D^i S^i + E^i, \quad \forall i
\]

(5)

where \( Z^i = [z^i_1, \cdots, z^i_{c_i}] \in \mathbb{R}^{K \times N} \) and \( S^i = [s^i_1, \cdots, s^i_{c_i}] \in \mathbb{R}^{LM \times N} \) are the reconstruction coefficients of \( X^i \) over \( A^i \) and \( D^i \), respectively, and \( E^i \) is the residual matrix. We can obtain the reconstruction coefficients by solving the following optimization problem:

\[
\begin{align*}
\min_{Z^i,S^i} & \quad \sum_i \|E^i\|_F^2 + \lambda_2\|Z^i\|_F^2 + \beta_2\|S^i\|_F^2 \\
\text{s.t.} & \quad X^i = A^i Z^i + D^i S^i + E^i, \quad \forall i
\end{align*}
\]

(6)

where \( \lambda_2 > 0 \) and \( \beta_2 > 0 \) are the weighting parameters.

### C. UNIFORM GENERIC REPRESENTATION

In Section III-A and III-B, we have introduced global and local generic representations. The global representation has good robustness for recognizing those indiscriminative regions such as forehead and cheek, while the local representation can mitigate the effect of facial variations. We aim to integrate the advantages of both the global representation and the local representation.

From the perspective of the local representation, regardless of the variation dictionary, the patches of the same probe image should be constructed respectively by the corresponding patches from the right gallery image, i.e., ideally, \( z^i_j = z^i_{j'} = \cdots = z^i_J = [0,0,\cdots,1,\cdots,0]^T \), where \( 1 \leq j \leq N \). From the perspective of the global representation, the probe image should be constructed by the corresponding gallery image, which means that \( z_J = z_{J'} = \cdots = z_J = [0,0,\cdots,1,\cdots,0]^T \). The above idea can be formulated as the following representation:

\[
X = AZ + DS + E, \quad X^i = A^i Z^i + D^i S^i + E^i, \quad \forall i.
\]

(7)

However, in practice, some face patches like cheek and forehead are not very discriminative and thus may be constructed by the wrong gallery patches. And even if the discriminative patches are constructed mainly by the corresponding gallery patches, the values of the representation coefficients \( z^i_j(i = 1, \cdots, C) \) are not necessarily equal to each other. Considering the two points, the constraint of directly finding the uniform coefficient \( z_J \) is unfit. Therefore, we relax the constraint by making the representation coefficients of the probe patches close to the uniform coefficient instead of equal. The proposed representation is as follows:

\[
X = AZ + DS + E, \quad X^i = A^i(Z + H^i) + D^i S^i + E^i, \quad \forall i.
\]

(8)

\( Z + H^i \) will be close to \( Z \) as long as we make \( H^i \) small. We can get the reconstruction coefficients by solving the following optimization problem:

\[
\begin{align*}
\min_{Z,S,H^i,S^i} & \quad \|E\|_F^2 + \lambda\|Z\|_F^2 + \beta\|S\|_F^2 \\
& + \sum_i \|E^i\|_F^2 + \beta\|S^i\|_F^2 + \gamma\|H^i\|_F^2 \\
\text{s.t.} & \quad X = AZ + DS + E, \quad X^i = A^i(Z + H^i) + D^i S^i + E^i, \quad \forall i
\end{align*}
\]

(9)

where \( H^i = [h^i_1, \cdots, h^i_{c_i}] \in \mathbb{R}^{K \times N} \) and \( \lambda > 0, \beta > 0, \) and \( \gamma > 0 \) are the weighting parameters. The regularization term \( \gamma\|H^i\|_F^2 \) balances the relationship between global and local representations. The proposed representation can be seen as the combination of (3) and (5) since the obtained \( Z \) satisfies (3) while the obtained \( Z + H^i \) satisfies (5), which means that the proposed representation can ensure the global consistency as well as the local diversity. In our proposed method, we can make use of both global and local features. The available information for our proposed representation is more than that for (3) or (5).

To improve the robustness of the representation and increase the classification power, we extract the eight nearest
neighbor patches of the probe and gallery patches. For a pixel on the margin of an image, we first use mirror transform and then extract its neighbors. The extracted neighbors also help to cope with the local deformation (e.g., misalignment). Corresponding to the local representation, we also extract the eight neighbors of the probe and gallery samples for the global representation.

Since images should be represented by the base elements of the class that they belong to, we propose a simple global representation. Corresponding to the local representation, we also extract its neighbors. The extracted neighbors also help to cope with the local deformation (e.g., misalignment).

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate the proposed algorithm using the Extended Yale B [42], AR [43], CMU-PIE [44], and LFW [45] databases. For the global generic representation, we resize the face images to $30 \times 30$. For the local generic representation, we resize the face images to $80 \times 80$ and fix the patch size to $20 \times 20$, and the interval $\kappa$ between the centers of two adjacent patches is 10 pixel. We set $\lambda = 3, \beta = 0.01$, and $\gamma = 0.01$ in the experiments.

We compare our method with several popular methods, including adaptive generic learning (AGL) [2], FLDA-single [9], LRA*-GL [22], SRC [5], PSRC [5], PCRC [27], PNN [28], BlockFLDA [30], ESR [14], sparse variation dictionary learning (SVDL) [33], LGR [15], JCR-ACF [40], and discriminative probabilistic latent semantic analysis (DpLSA) [25]. The results of DpLSA are cited from [25]. For the other methods, source codes are provided by the authors, and experiments are carried out according to our configuration. To make a fair comparison, we use the gray value of images as the feature for all the methods. We also evaluate these methods with the convolutional features of the LFW database provided by Yang [40]. All the experiments are performed on a 2.9GHz laptop with i5-9400 CPU and 24G RAM.

A. EXTENDED YALE B DATABASE

The Extended Yale B face database contains 2432 images of 38 individuals taken under 64 lighting conditions. Its extreme lighting conditions make it a challenging task for most face recognition methods. We show some sample images in Fig. 2.

The first 30 subjects are chosen for evaluation, while the other 8 subjects are used as the generic dataset. We use the frontal faces under normal illumination condition as the gallery dataset and the face images under other illumination conditions as the probe dataset. We list the recognition rates of all the methods and their computation times of recognizing one probe image in Table 1. One can see that the recognition rate of our method reaches 91.5%, ranking first. Although FLDA-single and LRA*-GL take the least time, their recognition rates are not satisfactory. The topic model DpLSA achieves the third-highest recognition rate by extracting bag-of-words features. Patch-based methods like

$^{15}$ means that LRA-GL is conducted in a parameter-free manner, which is proposed in the original paper.

FIGURE 2. Sample images from Extended Yale B database.
TABLE 1. Recognition Rates (%) and Computation Time (s) on Extended Yale B Database.

| Method | AGL [2] | FLDA-single [9] | LRA*-GL [22] | SRC [5] | PSRC [5] | PCRC [27] | PNN [28] |
|--------|---------|----------------|--------------|--------|---------|---------|---------|
| Accuracy | 60.1 | 59.1 | 68.2 | 41.5 | 78.9 | 76.9 | 69.2 |
| Time (s) | 0.0036 | 0.0022 | 0.0022 | 0.0027 | 0.0368 | 0.0048 | 0.0086 |

FIGURE 3. Sample images from AR database.

TABLE 2. Recognition Rates (%) on AR Database (session 1).

| Method | illumination | expression | disguise | illumination+disguise | average |
|--------|--------------|------------|----------|-----------------------|---------|
| AGL [2] | 86.3 | 75.0 | 54.4 | 47.8 | 65.3 |
| FLDA-single [9] | 85.8 | 83.8 | 38.8 | 32.8 | 59.8 |
| LRA*-GL [22] | 96.7 | 77.5 | 85.6 | 72.2 | 81.9 |
| SRC [5] | 75.4 | 85.8 | 53.8 | 22.8 | 56.9 |
| PSRC [5] | 90.4 | 87.5 | 96.3 | 78.8 | 86.8 |
| PCRC [27] | 95.4 | 87.5 | 95.0 | 80.0 | 88.2 |
| PNN [28] | 85.4 | 86.7 | 88.8 | 72.2 | 81.9 |
| BlockFLDA [30] | 72.9 | 50.4 | 60.0 | 45.6 | 56.0 |
| ESRC [14] | 98.8 | 93.8 | 77.5 | 75.3 | 86.2 |
| SVDL [33] | 97.9 | 93.3 | 81.3 | 75.6 | 86.6 |
| LGR [15] | 99.2 | 97.9 | 98.1 | 96.6 | 97.8 |
| JCR-ACF [40] | 98.3 | 96.7 | 96.9 | 93.8 | 96.2 |
| DpLSA [25] | 100 | 100 | 100 | 99.8 | 99.9 |
| UGR | 100 | 97.5 | 98.8 | 97.8 | 98.4 |

PSRC, PCRC, and PNN achieve higher recognition rates than generic-learning based methods like AGL, LRA*-GL, and SVDL, which means that image partition can help a lot under extreme lighting conditions. By taking advantage of image partition and the generic training dataset, LGR and JCR-ACF perform relatively well. But they are inferior to our method. Compared with LGR and JCR-ACF, the recognition rate of our method has increased by 5.7% and 1.3%, respectively. This is because our method benefits from the global consistency besides image partition and the generic training dataset. Moreover, our method runs much faster than LGR and JCR-ACF.

B. AR DATABASE

The AR face database contains over 4,000 color face images of 126 individuals, including frontal faces with different lighting conditions, facial expressions, and occlusions. Each individual has 26 pictures taken in two separate sessions. Fig. 3 shows some sample images from AR database.

Following the experimental setup in [14], we carry out experiments with a subset of 100 subjects. The first 80 subjects from session 1 and session 2 are used for evaluation, and the other 20 subjects from session 1 are used as the generic dataset. The images with neutral expression and under normal illumination condition from session 1 are used as the gallery samples. Table 2 and Table 3 show respectively experimental results on two sessions.

It can be seen that DpLSA achieves the highest recognition rates on session 1, reaching nearly 100%. Furthermore, on session 2, DpLSA also gets the first place. The average recognition rate of our method is lower than that of DpLSA, but DpLSA uses bag-of-words features, while our method only uses raw pixel features. Even so, our method achieves a higher recognition rate in the case of illumination, which is consistent with the experimental results on Extended Yale...
TABLE 3. Recognition Rates (%) on AR Database (session 2).

| Method      | illumination | expression | disguise | illumination-disguise | average |
|-------------|--------------|------------|----------|------------------------|---------|
| AGL [2]     | 55.4         | 44.2       | 31.3     | 26.6                   | 39.0    |
| FLDA-single [9] | 47.1         | 53.3       | 18.1     | 17.5                   | 34.0    |
| LRA*-GL [22]  | 85.4         | 65.4       | 61.3     | 50.9                   | 64.9    |
| SRC [5]      | 45.8         | 70.4       | 25.0     | 11.9                   | 37.2    |
| PSRC [5]     | 82.1         | 70.4       | 81.9     | 59.1                   | 71.5    |
| PCRC [27]    | 87.1         | 69.2       | 83.1     | 63.4                   | 74.1    |
| PNN [28]     | 75.0         | 74.6       | 71.3     | 50.9                   | 66.3    |
| BlockFLDA [30]| 56.7         | 35.8       | 45.0     | 29.7                   | 40.5    |
| ESRC [14]    | 87.5         | 80.4       | 56.3     | 47.8                   | 67.3    |
| SVDL [33]    | 84.6         | 80.4       | 59.4     | 50.6                   | 68.0    |
| LGR [15]     | 97.1         | 84.6       | 93.8     | 86.9                   | 90.0    |
| JCR-ACF [40] | 96.3         | 82.9       | 91.3     | 81.3                   | 87.1    |
| DpLSA [25]   | 97.7         | 97.0       | 100      | 97.0                   | 97.7    |
| UGR          | 99.6         | 87.9       | 93.1     | 89.7                   | 92.3    |

FIGURE 4. Sample images from CMU-PIE database.

B database. Although the generic dataset is not used, the performance of PSRC and PCRC is better than that of ESRC and SVDL on two sessions. This is because patch-based methods can deal with occlusion well. By exploiting the variation information and local features, LGR, JCR-ACF, and the proposed UGR outperform PSRC and PCRC on two sessions. And UGR also utilizes global features, so it works best among all competitive methods of using raw pixel features.

C. CMU-PIE DATABASE

The CMU pose, illumination, and expression (CMU-PIE) database contains more than 40,000 facial images of 68 individuals. For each individual, the images are taken across 13 different poses, under 43 different illumination conditions, and with four different expressions. We show some sample images in Fig. 4.

We evaluate our method using the first 48 subjects and build the generic dataset using the remaining 20 subjects. The neutral face images taken with the frontal pose (C27) are used as the gallery samples, and the remaining images with the poses C05, C07, C09, C27, and C29 are used as the probe samples. The results of different methods are listed in Table 4 from which we can see that the proposed UGR performs best in all competitive methods. For the poses C05 and C29, our method achieves the best results, but the recognition rates are only about 70%. The reason may be that the poses (looking left and right) cause the misalignment issue, and the reflection of light increases the difficulty of recognition. For the poses C07, C09, and C29, UGR achieves significantly better performance than the other methods, which demonstrates that our method is more robust to pose change than other methods.

D. LFW DATABASE

The LFW database contains over 13,000 face images of 5,749 individuals taken under an unconstrained environment. LFW-a is a version of LFW after alignment using...
Y. Ding et al.: UGR for Single Sample Face Recognition

FIGURE 5. Sample images from LFW database.

TABLE 5. Recognition Rates (%) on LFW Database.

| Method          | AGL [2] | FLDA-single [9] | LRA*—GL [22] | SRC [5] | PSRC [5] | PCRC [27] | PNN [28] |
|-----------------|---------|-----------------|---------------|---------|---------|----------|---------|
| Accuracy        | 42.7    | 15.7            | 27.9          | 23.8    | 36.6    | 32.6     | 28.4    |
| Method          | BlockFLDA [30] | ESRC [14] | SVDL [33] | LGR [15] | JCR-ACF [40] | DpLSA [25] | UGR     |
| Accuracy        | 11.2    | 43.0            | 38.7          | 46.2    | 42.9    | 37.6     | 46.8    |

TABLE 6. Recognition Rates (%) on LFW Database with the Convolutional Features.

| Method          | AGL [2] | FLDA-single [9] | LRA*—GL [22] | SRC [5] | PSRC [5] | PCRC [27] | PNN [28] |
|-----------------|---------|-----------------|---------------|---------|---------|----------|---------|
| Accuracy        | 69.8    | -               | 67.6          | 80.6    | 77.2    | 76.8     | 75.6    |
| Method          | BlockFLDA [30] | ESRC [14] | SVDL [33] | LGR [15] | JCR-ACF [40] | DpLSA [25] | UGR     |
| Accuracy        | 75.6    | 85.1            | -             | 84.4    | 86.0    | -        | 87.3    |

TABLE 7. Comparing with Deep Learning-based Methods on LFW Database.

| Method          | Accuracy(%) |
|-----------------|-------------|
| DeepID [47]     | 70.7        |
| VGG-Face [48]   | 84.7        |
| SGL+LightenedCNN [19] | 93.1      |
| JCR-ACF [40]    | 86.0        |
| UGR             | 87.3        |

commercial software tool [46]. 158 subjects with no less than 10 samples are gathered from LFW-a. And we collected 10 face images for each subject. Some sample images are shown in Fig. 5.

The first 50 subjects are chosen for evaluation, while the rest are used as the generic dataset. Because in this database, no frontal neutral face images exist, we select one image in sequence as the gallery sample and use the remaining nine images as probe samples. Thus ten experiments are performed. We report the average recognition rate. For the generic learning-based methods, we use each person’s mean face as the neutral image in the generic dataset.

The performance of different methods is presented in Table 5 from which one can see that the recognition rates of all the methods are a little low. This is because the gray features cannot represent human face well in unconstrained environment. Still, the proposed UGR outperforms all the other methods. In [40], JCR-ACF achieves exciting performance on the LFW database by taking advantage of the adaptive deep convolutional features. In this article, we also evaluate our method with the convolutional features of the LFW database provided by Yang [40]. Because the convolutional features are extracted in patches, we concatenate the patch features of one face as its global features. The experimental results are listed in Table 6. It can be seen that the results in Table 6 have a dramatic improvement compared with those in Table 5, which indicates that the convolutional features have good distinctiveness.

We further compare our method using the convolutional features with four deep learning-based methods, i.e., DeepID [47], JCR-ACF [40], VGG-Face [48], and SGL+Lightened CNN [19]. Since the implementation requires tremendous auxiliary data and many tips for fine-tuning parameters, we cite the existing results of VGG-Face and SGL+Lightened CNN in [19]. The authors of [19] declare that they take the protocol in JCR-ACF [40]. The results of DeepID and JCR-ACF are reported in accordance with [40]. We also take the same protocol and use the convolutional features provided by [40]. The performance is shown in Table 7. It can be seen that our method outperforms the other methods except SGL+LightenedCNN. SGL achieves the highest accuracy by employing LightenedCNN features. However, SGL is solved by alternating optimization, while our method has a closed-form solution. It is easy to implement our method.

E. EVALUATION OF THE PROPOSED REPRESENTATION

1) PARAMETER EVALUATION

In this section, we analyze the influences of $\lambda$, $\beta$ and $\gamma$ to the performance of our method. Experiments are carried out on Extended Yale B and AR (session 1 + session 2) databases. We tune $\beta$ and $\gamma$ within the range of $\{0.005, 0.01, 0.05, 0.1, 0.5\}$ and $\lambda$ within the range of $\{1, 2, 3, 4\}$. The performance of UGR under different parameter combinations is shown in Fig. 6. We observe that the performance tends to deteriorate when $\beta$ and $\gamma$ are too large. On Extended Yale B database, the performance of UGR reaches best when $\beta = 0.01$ and $\gamma = 0.005$ or 0.01. On AR database, the performance of UGR is stable for the changes of $\lambda$ and $\beta$ when $\gamma$ is fixed. And UGR works best when $\gamma = 0.005$ or 0.01. For the sake of simplicity, we fix $\lambda = 3$, $\beta = 0.01$, and $\gamma = 0.01$. 

VOLUME 8, 2020
In this section, we measure the magnitude of $Z$ with respect to $H^i$. We use the following equation to calculate the magnitude:

$$\text{Magnitude} = \frac{1}{C} \sum \text{mean}(\frac{|Z|}{|H^i|}),$$

(18)

where $C$ is the number of patches, \text{mean}(\cdot) calculates the mean value of all elements of a matrix, and the division is performed by element. Equation (18) reflects the magnitude relationship between the elements of $Z$ and $H^i$. We carry out experiments on AR database. The results are shown in Table 8. We can see that in all cases, the value of $Z$ is more than five times that of $H^i$. This means that $Z + H^i$ is close to $Z$.

3) NUMBER OF NEAREST NEIGHBOR PATCHES

As mentioned in the text, adding neighbor patches can improve the robustness of the representation and increase the classification power. In this section, we will verify this and analyze the impact of the number of extracted nearest neighbor patches on the performance of our method. A center patch has eight nearest neighbor patches, as shown in Fig. 7. We carry out experiments with different numbers of extracted nearest neighbor patches on AR (session 1 + session 2) database. Six different settings are listed in Table 9. For S1, we add all eight nearest neighbor patches to the classification experiment. For S2, we use six nearest neighbor patches of two combinations. For S3, S4, and S5, we use 2, 4, and 6 combinations, respectively. And for S6, we do not use the neighbor patches. For each setting, we carry out experiments with the listed combinations and report the average recognition rate in Table 10. We can see that the recognition rate decreases as the number of neighbor patches decreases. The recognition rate is only 82.6% without using neighbor patches, which is 12.8% lower than that of using eight neighbor patches. The results show the importance of using neighbor patches.

4) SCALE EVALUATION

In this section, we discuss the impact of scales on recognition performance. We carry out experiments with different scales on Extended Yale B and AR (session 1 + session 2) databases. For the global generic representation, we resize the face images to $\{20 \times 20, 30 \times 30, 40 \times 40, 50 \times 50\}$. For the local generic representation, we fix the interval $\kappa$ and choose the patch size within the range of $\{10 \times 10, 15 \times 15, 20 \times 20, 25 \times 25, 30 \times 30\}$. The performance of our method under the varied parameters on Extended Yale B and AR databases are shown in Fig. 6.
different combinations of global size and patch size is shown in Fig. 8. One can see that when the patch size is fixed to $20 \times 20$, the performance is the highest regardless of the global size. On Extended Yale B database, the performance with the global size of $20 \times 20$ and the patch size of $20 \times 20$ is higher than others. On AR database, the performance with the global size of $30 \times 30$ and the patch size of $20 \times 20$ is higher than others. The reason for this difference may be that Extended Yale B database contains extreme lighting conditions which make the obtained discriminating information contaminated, and images of different global sizes contain different amount of contaminated information. For Extended Yale B database, the global size of $20 \times 20$ may be more suitable. To make the parameters of all experiments consistent, we fix the global size to $30 \times 30$ and the patch size to $20 \times 20$.

5) NUMBER OF LOCAL REGIONS

In this section, we evaluate the impact of the number of local regions on recognition performance. The global size and the patch size are fixed, and we get different numbers of local regions by changing the interval $\kappa$. Table 11 shows the recognition rates of our method with different numbers of local regions. It can be noticed that the recognition rates of our method decrease as the number of local regions reduces. But referring to the results of other methods in Table 1, Table 2, and Table 3, our method is still very competitive.

V. CONCLUSION

In this article, we integrate the global and local generic representations into a uniform framework to address the SSPP problem. The consistency between global and local generic representations is explored. In the proposed method, the whole image and its patches are employed simultaneously. Therefore, the available information for our proposed method is more than that for alone global or local generic representations. Experiments on the Extended Yale B, AR, CMU-PIE, and LFW databases show that our method works well for the SSPP problem. Especially when dealing with illumination variation, our method using raw pixel features is superior to DpLSA using bag-of-words features. The proposed representation is scalable. In future work, we can combine global representation with multiple local representations based on different patch sizes. The idea of consistency explored in this work will continue. In this work, we put forward a quadratic optimization problem in response to the proposed representation. When multiple representations are combined, the quadratic optimization problem will become complicated. Therefore, how to put forward the optimization problem is worth thinking about. It will be discussed in the future.

REFERENCES

[1] M. Sharif, F. Naz, M. Yasmin, M. A. Shahid, and A. Rehman, “Face recognition: A survey,” J. Eng. Sci. Technol. Rev., vol. 10, no. 2, pp. 166–177, 2017.
[2] Y. Su, S. Shan, X. Chen, and W. Gao, “Adaptive generic learning for face recognition from a single sample per person,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2699–2706.
[3] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” Chemometrics Intell. Lab. Syst., vol. 2, nos. 1–3, pp. 37–52, 1987.
X. He and P. Niyogi, “Locality preserving projections,” in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 153–160.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 210–227, Feb. 2009.

P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, Jul. 1997.

J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Face recognition using LDA-based algorithms,” IEEE Trans. Neural Netw., vol. 14, no. 1, pp. 195–200, Jan. 2003.

D. Zhang, S. Chen, and Z.-H. Zhou, “A new face recognition method based on SVD perturbation for single example image per person,” Appl. Math. Comput., vol. 163, no. 2, pp. 895–907, Apr. 2005.

Q.-X. Gao, L. Zhang, and D. Zhang, “Face recognition using FLDA with single training image per person,” Appl. Math. Comput., vol. 205, no. 2, pp. 726–734, Nov. 2008.

J. Li, J. Pan, and S. Chu, “Face recognition from a single image per person using common subfaces method,” in Proc. Int. Symp. Neural Netw. Springer, 2007, pp. 905–912.

J. Lu, Y. P. Tan, and G. Wang, “Discriminative multimaniifold analysis for face recognition from a single training sample per person,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 39–51, Jan. 2013.

P. Zhang, X. You, W. Ou, C. L. Philip Chen, and Y.-M. Cheung, “Sparse discriminative multi-manifold embedding for one-sample face identification,” Pattern Recognit., vol. 52, pp. 249–259, Apr. 2016.

M. Pang, Y.-M. Cheung, B. Wang, and R. Liu, “Robust heterogeneous discriminative analysis for face recognition with single sample per person,” Pattern Recognit., vol. 89, pp. 91–107, May 2019.

W. Deng, J. Hu, and J. Guo, “Extended SRC: Undersampled face recognition via intraclass variant dictionary,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1864–1870, Jan. 2012.

T. Khadhraoui, F. Benzarti, and H. Amiri, “Gabor-feature-based local sparse representation for single sample per person,” Pattern Recognit., vol. 7, pp. 30573–30587, 2019.

Y. Ding received the B.S. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, China, in 2011, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, pattern recognition, and deep learning.
FAN LIU (Member, IEEE) received the B.S. degree in network engineering and the Ph.D. degree from the Nanjing University of Science and Technology (NUST), in June 2009 and January 2015, respectively. From February 2014 to May 2014, he worked as an Intern with Microsoft Research Asia. From September 2008 to December 2008, he studied at Ajou University of South Korea as an Exchange Student. He is currently a Lecturer with Hohai University. He also serves as a Reviewer of Information Sciences, Neurocomputing, KSII Transactions on Internet and Information Systems, and Pattern Analysis and Application. His research interests include computer vision, image processing, pattern recognition and deep learning.

ZHENMIN TANG received the B.S. degree from Harbin Engineering University (HEU), in January 1982, and the M.S. and Ph.D. degrees from the Nanjing University of Science and Technology (NUST), in January 1988 and March 2003, respectively. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include image processing, pattern recognition, artificial intelligence, and expert systems. He has over 60 journal and conference papers in these areas.

TAO ZHANG received the B.S. and Ph.D. degrees from the School of Computer Science and Engineering, Nanjing University of Science and Technology, China, in 2010 and 2018, respectively. He is currently an Assistant Professor with the School of Information Engineering, Yangzhou University, China. His research interests include subspace learning, multi-view learning, and deep learning.

* * *

VOLUME 8, 2020