# Location-Sensitive Sparse Representation of Deep Normal Patterns for Expression-robust 3D Face Recognition

Huibin Li[1], Jian Sun[1], Liming Chen[2]
[1]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China
[2]Department of Mathematics and Informatics, Ecole Centrale de Lyon, Lyon, 69134, France
{huibinli,jiansun}@xjtu.edu.cn, liming.chen@ec-lyon.fr

## Abstract

*This paper presents a straight-forward yet efficient, and expression-robust 3D face recognition approach by exploring location sensitive sparse representation of deep normal patterns (DNP). In particular, given raw 3D facial surfaces, we first run 3D face pre-processing pipeline, including nose tip detection, face region cropping, and pose normalization. The 3D coordinates of each normalized 3D facial surface are then projected into 2D plane to generate geometry images, from which three images of facial surface normal components are estimated. Each normal image is then fed into a pre-trained deep face net to generate deep representations of facial surface normals, i.e., deep normal patterns. Considering the importance of different facial locations, we propose a location sensitive sparse representation classifier (LS-SRC) for similarity measure among deep normal patterns associated with different 3D faces. Finally, simple score-level fusion of different normal components are used for the final decision. The proposed approach achieves significantly high performance, and reporting rank-one scores of 98.01%, 97.60%, and 96.13% on the FRGC v2.0, Bosphorus, and BU-3DFE databases when only one sample per subject is used in the gallery. These experimental results reveals that the performance of 3D face recognition would be constantly improved with the aid of training deep models from massive 2D face images, which opens the door for future directions of 3D face recognition.*

## 1. Introduction

Recently, deep learning, especially Convolutional Neutral Network (CNN) have witnessed great success in the computer vision community, significantly improving the state of the art in many tasks, such as image classification, segmentation, and object detection, etc [26], [14], [42]. One of most important reason of these achievements is the availability of large-scale training data (e.g., ImageNet). Meanwhile, the *off-the-shelf* pre-trained deep CNN models have surprising and consistent good generalization ability for various visual recognition tasks [14], [42].

Due to the universality of deep CNN methods, 2D face recognition has also achieved great breakthrough by training deep face models using a huge number of 2D face images of different subjects. By combing CNN and 3D morphable model [58], [24], the challenging problem of large pose variations for 2D face recognition recognition has also achieved great progress. As shown in Table 1, the Facenet proposed by Google [46] was trained using 200 million images of eight million unique identities. However, building such a large dataset is beyond the capabilities of most research groups. Although the amount of training identities have proved can be largely reduced to several thousands (e.g., 2,622 for VGGFace) without loss the performance, there should be enough samples for each subjects for training (e.g., 2.6 million) [37]. However, collecting millions of 3D face scans from thousands of identities with large diversity is still unrealized in both industrial community and academia. As shown in Table 1, the largest publicly available 3D face dataset (i.e., FRGC v2.0) only contains 4,007 high-resolution 3D face scans of 466 subjects. The recently published Lock3DFace dataset contains 5,711 video clips of 509 identities, however, these 3D face scans were captured by a Kinect, thus with low spatial resolution and depth accuracy. Therefore, the limitation of enough training data is currently the main bottleneck of developing deep CNN methods for 3D face recognition. A promise solution is generating 3D face scans from 2D face images based on 3D face reconstruction (e.g. automated 3D Morphable Model based fitting [40]) or 3D face modeling (e.g. Robust Multilinear Model Learning [7]) algorithms.

The goal of this paper is to propose a deep CNN-based approach for 3D face recognition. Instead of training deep CNN models directly using 3D face scans, we propose to use the pre-trained deep CNN models originally learned for 2D face recognition to generate the facial descriptors of 3D face scans. In particular, we propose to use the pre-trained

978-1-5386-1124-1/17/$31.00 ©2017 IEEE

Table 1. Comparisons of 2D and 3D face datasets.

| 2D Face Dataset | Identities | Images | 3D Face Dataset | Identities | Images |
|---|---|---|---|---|---|
| LFW | 5,749 | 13,233 | FRGC v2.0 (2005) [39] | 4,66 | 4,007 |
| WDRef [11] | 2,995 | 99,773 | BU-3DFE (2006) [56] | 100 | 2,500 |
| CelebFaces [51] | 10,177 | 202,599 | Bosphorus (2008) [45] | 1,05 | 4,666 |
| VGGFace [37] | 2,622 | 2.6M | 3D-TEC (2011) [53] | 214 | 428 |
| FaceBook [52] | 4,030 | 4.4M | Florence Superface (2012) [5] | 50 | $50^v$ |
| WebFace [15] | 10,575 | 494,414 | KinectFaceDB (EURECOM) (2013) [23] | 52 | $936^v$ |
| Google [46] | 8M | 200M | Lock3DFace (2015) [57] | 5,09 | $5711^v$ |

deep CNN model to generate the deep representations of facial surface normal maps, namely deep normal patterns (DNP). The idea of using facial surface normal maps is inspired by Li *et al.* [27], [28], which claimed that encoding facial surface normals can generate more discriminative descriptors for 3D face recognition than directly encoding the 3D coordinates of facial surfaces. Based on the deep normal representations, we propose to use the sparse representation-based classifier for 3D face matching. Finally, to deal with facial expression variations, we propose to learn the importance weights of different facial parts, i.e., importance weights of deep surface normal activations at different spatial locations, resulting in location sensitive sparse representation-based classifier (LS-SRC). Experiments in Section 5 demonstrates that this straight-forward solution is significantly efficient for expression-robust 3D face recognition, and achieving state-of-the-art results on the FRGC v2.0, Bosphorus and BU-3DFE datasets. To the best of our knowledge, this is the first work attempting to use deep learning for 3D face recognition.

## 2. Related work

In the past decade, 3D face recognition methods have received widely attention [8] [48] along with the releases of different 3D face databases such as FRGC v2.0, BU-3DFE, Bosphorus, etc., see Table 1 for more information. There are two main concerns of existing 3D face recognition approaches: *1) how to produce identity-distinguishable facial shape descriptors to recognize 3D face scans from different subjects? 2) how to provide expression-invariant facial shape representations to resist the facial expression variations of 3D face scans belong to the same subject?* To meet these two requirements, a large number of 3D face recognition approaches have been proposed in the literature. According to [16], these methods can be classified into three categories: *expression deformation modeling based approaches; intrinsic surface-distance based approaches; and local region/feature based approaches.*

**Expression deformation modeling based approaches**. These methods generally assume that human facial surfaces are non-rigid and deformable surfaces, deforming according to the variations of facial expression changes. The facial surface deformation behaviors are modeled either by physical principles such as *thin-plate-spline based fitting* [31], *annotated deformable model based fitting* [25] and *deformable model based fitting* [35], or by statistical models such as *principal component analysis based subspace model* [1], *3D morphable model* [3], and *3D bilinear model* [35]. Based on these deformation models, facial expressions can be removed (3D face neutralization or decomposition), or transferred (3D face morphing), making this kind of methods robust to facial expression variations.

**Intrinsic surface-distance based approaches**. These methods generally assume that facial surfaces are compact, connected, and zero-genus manifolds embedded in 2D Euclid space. Geodesic distance, as an intrinsic surface geometry quantity, has been used to construct expression-invariant representations of 3D faces [9]. These methods assume that facial expression variations only change the extrinsic geometry of facial surface, while the intrinsic geometry is approximately invariant and thus can be contributed to the subject's identity. Furthermore, if one assume that facial surfaces are differentiable, their tangent spaces are also manifolds. Thus, elastically matching of facial curves can be accomplished by inducing elastic Riemannian metrics on the shape space of various facial curves. These ideas have been successfully proposed and developed in 3D face recognition algorithms, in which facial surfaces are described as iso-depth curves [43], iso-geodesic curves [44], iso-geodesic path curves [16], or iso-geodesic strips [4] for distance measurements. This kind of methods are generally robust to facial expression variations, while the open month problem is a main issue for practical face matching.

**Local region/feature based approaches**. Local region based approaches assume that there always exist some rigid facial parts (e.g., nose region), which are approximately invariant to facial expression deformations. This kind of approaches first divide the 3D face into a number of regions, then perform recognition (e.g. feature extraction, region selection, and matching) on each region, and finally fuse all the results. Some representative approaches are [2], [10], [18], [41], [50] [17]. Local feature based approaches are further classified into two categories: local tex-

ture descriptor-based methods and local shape descriptor-based approaches. The former ones generally treat facial geometric-maps (e.g., geometry map, normal map) as 2D texture images while the later ones treat 3D faces as free-form surfaces for feature extraction. Local texture descriptors such as log-Gabor filter [12], Haar wavelet [54], Local Binary Patterns (LBP) [54], [28] and its extension [22] have been successfully employed. More sophisticated and robust 2D face descriptors such as Multi-Directional Multi-Level Dual-Cross Patterns [13] and Local Patterns of Gabor Magnitude and Phase [55] can also been successfully explored to describe those facial geometric-maps. Local shape based descriptors such as surface normals, curvatures, and shape index values [29], [30], 3D shape context [6], and spherical face representation [32] are some representative ones. Comparing with local texture descriptors, local shape descriptors are pose-invariant and thus can handle the cases of large head pose variations.

Different from existing 3D face recognition approaches, this paper is inspired by the following two ideas. 1) Deep CNN features extracted from pre-trained deep models have strong generalization ability for different tasks. 2) As pointed out in [28], facial surface normals contain more discriminative information than facial surface coordinates, and their discriminative ability can be further enhanced by different encoding methods [27]. Similar idea has also been explored in recent algorithm utilizing facial normal vectors for dense face matching [34] and nasal patch description [17]. Combing these two ideas, we propose to extract deep CNN features from facial surface normal maps, and generate Deep Normal Patterns (DNP) for facial surface description, which can be regarded as a naturally generalization of the Multi-scale and Multi-component Local Normal Patterns [28]. To handle the expression variation issue, we propose to learn the importance weights of deep activations (i.e., filter responses) at different spatial locations, which implicitly corresponds to different facial parts. These weights are further used to formulate location-sensitive sparse representation based classifier for final 3D face recognition.

The reminder of the paper is organized as follows. Section 3 introduces the proposed Deep Normal Patterns (DNP). Section 4 describes the location-sensitive sparse representation-based classifier. In section 5, we will present the experimental settings and discuss the experimental results, and Section 6 concludes the paper.

## 3. Deep Normal Patterns

### 3.1. Facial Surface Normal Estimation

Depending on the data format of 3D face surfaces, surface normals can be estimated directly from triangular meshes, point-clouds, or the coordinates of depth images. In this paper, all facial surfaces are preprocessed and normalized into range images with the $x$, $y$, and $z$ coordinates. Thus, we employ the local plane fitting method [21] for normal estimation. Specifically, given a facial range image $\mathbf{P}$ represented by an $m \times n \times 3$ matrix:

Figure 1. Illustration of facial normal estimation: (a) the original range image, (b-d) its normal images of component $x$, $y$ and $z$ (the sample comes from the FRGC v2.0 dataset).

$$\mathbf{P} = [p_{ij}(x,y,z)]_{m\times n} = [p_{ijk}]_{m\times n\times\{x,y,z\}}, \quad (1)$$

where $p_{ij}(x,y,z) = (p_{ijx}, p_{ijy}, p_{ijz})^T, (1 \le i \le m, 1 \le j \le n, i, j \in \mathbb{Z})$ represents the 3D coordinates of the point $p_{ij}$. Let its unit normal vector matrix ($m \times n \times 3$) be

$$\mathbf{N}(\mathbf{P}) = [n(p_{ij}(x,y,z))]_{m\times n} = [n_{ijk}]_{m\times n\times\{x,y,z\}}, \quad (2)$$

where $n(p_{ij}(x,y,z)) = (n_{ijx}, n_{ijy}, n_{ijz})^T, (1 \le i \le m, 1 \le j \le n, i, j \in \mathbb{Z})$ denotes the unit normal vector of $p_{ij}$. As described in [21], the normal vector **N(P)** of range image **P** can be estimated by fitting local plane. That is to say, for each point $p_{ij} \in \mathbf{P}$, its normal vector $n(p_{ij})$ can be estimated as the normal vector of the following local fitted plane:

$$S_{ij} : n_{ijx}q_{ijx} + n_{ijy}q_{ijy} + n_{ijz}q_{ijz} = d, \quad (3)$$

where $(q_{ijx}, q_{ijy}, q_{ijz})^T$ represents any point within the local neighborhood of point $p_{ij}$ and $d = n_{ijx}p_{ijx} + n_{ijy}p_{ijy} + n_{ijz}p_{ijz}$. In this paper, a neighborhood of $5 \times 5$ window is used. To simplify, each normal component in equation (2) can be represented by an $m \times n$ matrix:

$$\mathbf{N}(\mathbf{P}) = \begin{cases} \mathbf{N}(\mathbf{X}) = [n^x_{ij}]_{m\times n}, \\ \mathbf{N}(\mathbf{Y}) = [n^y_{ij}]_{m\times n}, \\ \mathbf{N}(\mathbf{Z}) = [n^z_{ij}]_{m\times n}. \end{cases} \quad (4)$$

where $\|(n^x_{ij}, n^y_{ij}, n^z_{ij})^T\|_2 = 1$.

Figure 2 shows a normalized range image of the FRGC v2.0 database and its estimated three normal component images. It's not difficult to find that the normal images contain more informative geometric information than their corresponding range image which looks quite smooth. The shape details around the eyes and mouth regions are well highlighted in the normal images.

### 3.2. Deep Normal Patterns

To comprehensively highlight facial surface normals, we proposed to use the "vgg deep face net" to generate deep

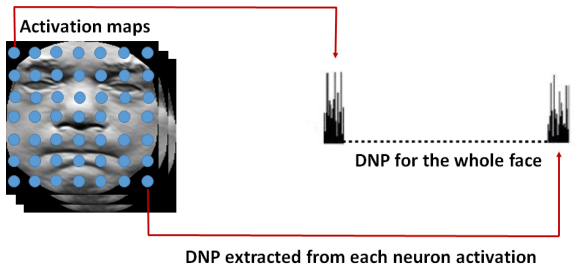


Figure 2. The proposed Deep Normal Patterns for each normal component image is generated by concatenating all the filters' responses at a certain spatial locations.

normal representations. It was trained for 2D face description and recognition by using 982,803 face images of 2,622 identities [37]. It comprises 16 learnable weight layers, 13 of which are convolutional layers and followed by 3 fully connected (FC) layers [47]. Each weight layer contains a linear operator followed by one or more non-linearities such as ReLU and max pooling. The input to this deep network is a color face image of size $214 \times 214 \times 3$. The outputs of the first two FC layers are 4,096 dimensional and the last FC layer has either 2,622 or 1,024 dimensions, depending on the face identification and verification tasks, respectively.

Given a normal component image $\vec{n} \in \mathbb{R}^{214\times214\times3}$ and a deep Convolutional Neural Network (CNN) with $L$ layers $\varphi_L \circ \cdots \circ \varphi_1$. The output of each layer $\boldsymbol{x}_l = \varphi_l \circ \cdots \circ \varphi_1(\boldsymbol{x})$ can be seen as a descriptor tensor $\boldsymbol{x}_l \in \mathbb{R}^{W_l \times H_l \times N_l}$, where $W_l$ and $H_l$ are the width and height of the tensor and $N_l$ is the number of filters. If all $\varphi_l$ are convolutional layers, this descriptor tensor can preserve the facial spatial information. By collecting the $N_l$ responses at a certain spatial location, one obtains a $N_l$ dimensional descriptor vector, which encodes a certain facial region. To achieve a global facial representation and preserve facial spatial information, a order-sensitive pooling method is used. This pooling method outputs a global descriptor, namely deep normal patterns, by simply concatenating the $N_l$ dimensional descriptor vectors according to their spatial locations (as illustrated in Fig. 2). Note that considering the trade-off between feature dimension and generalization ability, we only use the activations of the last convolutional layer (i.e., *conv5-3* of the vgg-deep-face net) to generate the deep normal patterns. Notice that other layers can also been used to generate DNP.

## 4. Location-Sensitive Sparse Representation for 3D Face Recognition

Sparse Representation-based classifier (SRC) is originally proposed for robust 2D face recognition and then extended to 3D face recognition [28]. The idea of Location-Sensitive Sparse Representation-based Classifier (LS-SRC) is with the same sprit as the weighted SRC proposed in [28], both of them inspired by considering the fact that different facial parts have different importance for 3D face recognition. For completeness, we give a short introduction of SRC and the proposed LS-SRC.

Given a gallery set with $N$ 3D face scans, each of which belongs to one subject, we define the dictionary of sparse representation model as $\boldsymbol{D} \doteq [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_N]$. Then for any probe $\boldsymbol{y}$ we have

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} + \varepsilon. \tag{5}$$

Sparse coefficients $x$ in eqn. (5) can be solved by the following $l_0$ minimization problem:

$$\min \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 \text{ s.t. } \|\boldsymbol{x}\|_0 \leq L. \tag{6}$$

where $L$ measures the sparsity of the representation coefficient. Assume $\hat{\boldsymbol{x}}$ is the minimizer of problem (6), and then the index of minimal reconstruction error vector:

$$r_i(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{D}\delta_i(\hat{\boldsymbol{x}})\|_2^2, i = 1, 2, \ldots, N. \tag{7}$$

delivers the identity of the probe $\boldsymbol{y}$, where $\delta_i$ is a characteristic function which selects the coefficient associated with the $i$-th gallery.

Now suppose there are $K$ different spatial locations for each activation map ($K = W_l \times H_l$ ), and we denote $w_k \in \mathbb{R}^{N_l}$ as the importance weight of all activation maps at a certain location. Then the DNP for each normal component image $f_i$ can be written as

$$f_i = [f_{i1}; v_{i2}; \ldots; f_{ik}; \ldots; f_{iK}],$$

where $f_{ik} \in \mathbb{R}^{N_l \times 1}$, and the dictionary $\boldsymbol{D}$ can be denoted as

$$\boldsymbol{D} = [\boldsymbol{D}_1; \boldsymbol{D}_2; \ldots; \boldsymbol{D}_k; \ldots; \boldsymbol{D}_K],$$

where $\boldsymbol{D}_k = [f_{1,k}, f_{2,k}, \ldots, f_{i,k}, \ldots, f_{n,k}]$, and a probe $y$ can be denoted as

$$\boldsymbol{y} = [\boldsymbol{y}_1; \boldsymbol{y}_2; \ldots; \boldsymbol{y}_k; \ldots; \boldsymbol{y}_K],$$

where $\boldsymbol{y}_k \in \mathbb{R}^{N_l \times 1}, k = 1, 2, \ldots, K$.

Equation (6) can then be rewritten as the following location-sensitive sparse representation model:

$$\min \sum_{k=1}^{K} w_k \|\boldsymbol{y}_k - \boldsymbol{D}_k x\|_2^2 \text{ s.t. } \|\boldsymbol{x}\|_0 \leq L, \tag{8}$$

and then the corresponding location-sensitive reconstruction residuals is

$$r_i(\boldsymbol{y}) = \sum_{k=1}^{K} w_k \|\boldsymbol{y}_k - \boldsymbol{D}_k \delta_i(\hat{\boldsymbol{x}})\|_2^2, i = 1, 2, \ldots, N. \tag{9}$$

To solve eqn. (8), we notice that it equals to solve

$$\min \sum_{k=1}^{K} \|w_k \boldsymbol{y}_k - w_k \boldsymbol{D}_k x\|_2^2 \text{ s.t. } \|\boldsymbol{x}\|_0 \leq L. \tag{10}$$

We denote $W(\boldsymbol{D}) = [w_1\boldsymbol{D}_1; w_2\boldsymbol{D}_2; \ldots; w_K\boldsymbol{D}_K]$, and $W(\boldsymbol{y}) = [w_1\boldsymbol{y}_1; w_2\boldsymbol{y}_2; \ldots; w_K\boldsymbol{y}_K]$. Then eqn. (10) equals to

$$\min \|W(\boldsymbol{y}) - W(\boldsymbol{D})\boldsymbol{x}\|_2^2 \text{ s.t. } \|\boldsymbol{x}\|_0 \leq L. \quad (11)$$

Problem (11) means that the location-sensitive sparse representation model as expressed in eqn.(8) amounts to solving a single SRC with global feature vectors in simply stacking weighted features at a certain location. Once determined the sparse representation coefficient $\hat{\boldsymbol{x}}$ of eqn.(11), location-sensitive reconstruction residuals in eqn.(9) can be computed. Then the minimal $r_i(\boldsymbol{y})$ can be used to determine the identity of $\boldsymbol{y}$. We call this sparse representation-based classifier enhanced by spatial location weights as Location-Sensitive Sparse Representation-based Classifier (LS-SRC) in the subsequent. To solve eqn.(6) and eqn.(11), the Orthogonal Matching Pursuit (OMP) [38] algorithm with the sparse number $L$ = 30 is used for all the following experiments. In practice, the weights can be learned from a given training datasets.

# 5. Experimental Results

## 5.1. 3D databases and evaluation protocols

To evaluate the expression-robustness of the proposed recognition approach, three databases: FRGC v2.0, BU-3DFE and Bosphorus 3D face dataset, which depicting a rich set of facial expressions (e.g., subtle, prototypical and exaggerated) are used in our experiments. The basic information of these datasets and their evaluation protocols are introduced as follows.

**FRGC version 2.0 database** [39]: FRGC v2.0 is the largest public 3D face dataset which has been widely used for 3D face recognition in the past decade. It consists of 4007 textured 3D face scans of 466 subjects with different facial expressions (1642 samples) captured under controlled lighting conditions. The face scans of this dataset were captured using the Minolta laser sensors over two different sets of sessions: Fall 2003 and Spring 2004. The resolution of each face scan is $640 \times 480$. To evaluate our approach on this dataset, two standard experimental protocols are used. For the first protocol, 466 samples of the first scans of each subject are used to make a gallery and the remaining 3541 samples are treated as probe. The purpose of the second protocol is used to evaluate the performance degradations caused by replacing neutral probes by non-neutral probes. That is, we compare the recognition rates between neutral gallery vs. neutral probe and neutral gallery vs. non-neutral probe. In our experiments, the location weights were learned from BU-3DFE and Bosphorus datasets, and namely LS-SRC$_{bu}$ and LS-SRC$_{bos}$, respectively.

**BU-3DFE database** [56]: This database contains 100 subjects (56 female and 44 male) with age ranging 18 to 70 years old, and with a variety of ethnic ancestries. For each subject, 24 samples with six prototypic expressions (happiness, disgust, fear, angry, surprise, and sadness) of four intensity levels in addition to a neutral one are included. All the 3D face scans are captured using the 3dMD imaging system, and each one is saved as a polygonal mesh with a resolution range 20,000 to 35,000 polygons. Due to the diversity of different expression types and the variability of different levels of expression intensities, this dataset is regarded as one the most challenging benchmarks for expression-robustness 3D face recognition algorithms. In our experiments, the single neutral sample of each subject is used as gallery, and the scans with different types of expressions are used as probe samples, respectively. The location weights were learned from Bosphorus dataset, namely LS-SRC$_{bos}$.

**Bosphorus database** [45]: This database contains 4666 textured 3D face scans of 105 subjects with six types of basic facial expressions in addition to neutral, 28 types of facial action units caused by movements of lip(s), mouth, cheek, nose, chin, eye(s), and brow(s), systematic head poses (13 yaw and pitch rotations), and varieties of face occlusions (beard, moustache, hair, hand, eyeglasses). The 3D face scans are acquired using structured-light based 3D system with the sensor resolution in $x$, $y$, and $z$ (depth) dimensions are 0.3mm, 0.3mm and 0.4mm respectively. Since this paper focuses on expression-robust face recognition, the first neutral scan per subject is used in the gallery (105 samples) and the remaining scans with different expressions, different action units, and near frontal head poses (without occlusions) are used as probes (2797 samples). In practice, the location weights were learned from BU-3DFE dataset, namely LS-SRC$_{bu}$.

### 5.1.1 Experimental results on the FRGC v2.0 database

**(1) Deep Normal Patterns vs. Deep Depth Patterns** To show the effectiveness of DNP, we compare it with deep depth patterns (DDP), *i.e.* deep representations of facial depth (range) image extracted from the same deep net and net layer. For fair comparison, Neighbor neighbor (NN) classier is used for similarity measurement. Table 2 reports the rank-one recognition rates on the whole FRGC v2.0 database. We can see that DNP performs much better (more than 20% higher) than DDP. This indicates the strong discriminative power of DNP.

**(2) NN Classifier vs. SRC Classifier.** Table 3 compares the rank-one recognition rates achieved by NN and SRC on the whole FRGC v2.0 database. We can see that SRC outperforms NN about 10%, 7%, 4%, and 5% for DNP$_x$, DNP$_y$, DNP$_z$, DNP$_{xyz}$, respectively. These results highlight the effectiveness of SRC when using deep normal patterns (DPN) based 3D facial representation.

Table 2. Comparison of the rank-one scores on the whole FRGC v2.0 database: deep depth patterns (DDP) vs. deep normal patterns (DNP) using the nearest neighbor (NN) classifier.

| Approaches | Rank-one Scores |
|---|---|
| (1) $DDP_z$ + NN | 71.83% |
| (2) $DNP_x$ + NN | 80.26% |
| (3) $DNP_y$ + NN | 87.38% |
| (4) $DNP_z$ + NN | 91.35% |
| (4) $DNP_{xyz}$ + NN | **92.03%** |

Table 3. Comparison of the rank-one scores on the whole FRGC v2.0 database using deep normal patterns (DNP): nearest neighbor (NN) classifier vs. sparse representation-based classifier (SRC).

| Approaches | Rank-one Scores |
|---|---|
| (1) $DNP_x$ + NN | 80.26% |
| (2) $DNP_x$ + SRC | **90.70%** |
| (3) $DNP_y$ + NN | 87.38% |
| (4) $DNP_y$ + SRC | **94.52%** |
| (5) $DNP_z$ + NN | 91.35% |
| (6) $DNP_z$ + SRC | **95.35%** |
| (7) $DNP_{xyz}$ + NN | 92.03% |
| (8) $DNP_{xyz}$ + SRC | **97.30%** |

**(3) SRC Classifier vs. LS-SRC Classifier.** In this experiment, we show the effectiveness of LS-SRC, in which the location sensitive weights are learned from BU-3DFE and Bosphorus databases, respectively. In particular, these weights are archived by normalizing the rank-one recognition scores on the training dataset based on the DNP at a certain location and nearest neighbor classifier. The comparison results on the whole FRGC v2.0 database are shown in Table 4. We can see that LS-SRC is significantly efficient to boost the performance of SRC. LS-$SRC_{bu}$ achieves slightly better results than LS-$SRC_{bos}$.

Table 4. Comparison of the rank-one score improvements on the whole FRGC v2.0 database: location sensitive weights are learned using NN classifier on BU-3DFE (LS-$SRC_{bu}$) and Bosphorus (LS-$SRC_{bos}$), respectively.

| | SRC | LS-$SRC_{bu}$ | LS-$SRC_{bos}$ |
|---|---|---|---|
| $DNP_x$ | 90.70% | **92.03%** | **91.77%** |
| $DNP_y$ | 94.52% | **94.98%** | **94.87%** |
| $DNP_z$ | 95.35% | **95.71%** | **95.86%** |
| $DNP_{xyz}$ | 97.30% | **98.01%** | **97.93%** |

**(4) Robustness to facial expression variations.** To show the robustness to facial expression variations, we run the second protocol on the FRGC v2.0 dataset, where probe set is divided into neutral subset and non-neutral subset. The results in Table 5 show that the proposed method is quite robust to facial expression variations, and achieving the best rank-one score (99.39%) on the neutral subset and competitive score (96.29%) on the non-neutral subset.

Table 5. Comparing the degradations of rank-one scores influenced by facial expression changes on the FRGC v 2.0 database (Subset I: neutral probe samples; Subset II: non-neutral probe samples).

| | Subset I | Subset II |
|---|---|---|
| Mian *et al.* (2008) [33] | 99.00% | 86.70% |
| Alyuz *et al.* (2010) [2] | 98.39% | 96.40% |
| Queirolo *et al.* (2010) [41] | 99.50% | 94.80% |
| Huang *et al.* (2012) [22] | 99.20% | 95.10% |
| Berretti *et al.* (2013) [6] | 97.30% | 92.80% |
| Drira *et al.* (2013) [16] | 99.20% | 96.80% |
| Li *et al.* (2014) [28] | 98.00% | 94.20% |
| Emambakhsh *et al.* (2016) [17] | 98.45% | **98.50%** |
| Our method (DNP + LS-$SRC_{bu}$) | **99.39%** | 96.29% |

**(5) Comparison with the state-of-the-art.** Table 6 gives a comprehensive performance comparisons between the proposed method and the state of the arts on the whole FRGV v2.0 dataset. Our method achieves a rank-one recognition rate of 98.01%, which is better than most other ones and only slightly worse than [54], [41] and [50]. Notice that sophisticated machine learning or optimization tools such as boosting, linear discriminant analysis, simulated annealing were used in [54], [50] and [41], respectively.

#### 5.1.2 Experimental results on the BU-3DFE database

Table 7 reports the performance comparisons of the proposed method and state-of-the-art ones on different expression subsets and the whole dataset of BU-3DFE. The location-sensitive weights are learned from the Bosphorus dataset. Our method achieves rank-one scores more than 98% except the disgust subset, which outperform about 5%, 7%, and 7% on the subsets of happy, surprise and fear respectively comparing the best results. On the whole dataset, our method archives a rank-one score of 96.1%, which is much higher than the state-of-the-art results. Expression robustness of our method is also demonstrated by these results.

#### 5.1.3 Experimental results on the Bosphorus database

Table 8 reports the performance comparisons of the proposed method and state-of-the-art ones on the expression subset of Bosphorus database. Our method achieves rank-one scores of 97.89%, 97.82%, and 97.60% when using NN, SRC, and LS-$SRC_{bos}$ as classifier, respectively. These results are quite closed to the best rank-one scores. It should

Table 7. Rank-one recognition rates for different expression types (400 samples for each type of expression), and all 2400 expression samples from the BU-3DFE database used as probes, with 100 neutral samples used as gallery.

| Approaches | Happy | Surprise | Fear | Sadness | Anger | Disgust | All |
|---|---|---|---|---|---|---|---|
| Hajati *et al.* (2012) [20] | 86.0% | 84.0% | 82.0% | 85.0% | 93.0% | 79.0% | 84.83% |
| Li *et al.* (2014) [28] | 93.8% | 83.8% | 91.8% | 98.5% | 97.0% | 88.5% | 92.2% |
| Emambakhsh *et al.* (2016) [17] | 88.5% | 91.0% | 89.8% | 92.3% | 90.1% | 81.8% | 88.9% |
| Our method (DNP + LS-SRC$_{bos}$) | **98.3%** | **98.5%** | **98.3%** | **99.8%** | **98.0%** | 84.0% | **96.1%** |

Table 6. Rank-one recognition rates on the FRGC v2.0 database.

| Approaches | Rank-one scores |
|---|---|
| Chang *et al.* (2006) [10] | 91.90% |
| Cook *et al.* (2006) [12] | 92.90% |
| Kakadiaris *et al.* (2007) [25] | 97.00% |
| Mian *et al.* (2007) [32] | 96.20% |
| Mian *et al.* (2008) [33] | 93.50% |
| Faltemier *et al.* (2008) [18] | 97.20% |
| Osaimi *et al.* (2009) [1] | 96.50% |
| Wang *et al.* (2010) [54] | **98.39%** |
| Queirolo *et al.* (2010) [41] | **98.40%** |
| Berretti *et al.* (2010) [4] | 94.15% |
| Alyuz *et al.* (2010) [2] | 97.50% |
| Spreeuwers *et al.* (2011) [50] | **99.00%** |
| Huang *et al.* (2012) [22] | 97.60% |
| Smeets *et al.* (2013) [49] | 89.60% |
| Berretti *et al.* (2013) [6] | 95.60% |
| Drira *et al.* (2013) [16] | 97.00% |
| Li *et al.* (2014) [28] | 96.30% |
| Li *et al.* (2015) [29] | 96.30% |
| Emambakhsh *et al.* (2016) [17] | 97.90% |
| Guo *et al.* (2016) [19] | 97.00% |
| Our method (DNP+LS-SRC$_{bu}$) | 98.01% |

Table 8. Rank-one recognition rates on the Bosphrous database.

| Approaches | Rank-one scores |
|---|---|
| Alyuz *et al.* (2010) [2] | 98.20% (2814/105) |
| Kakadiaris *et al.* (2011) [36] | 98.20% (2797/105) |
| Smeets *et al.* (2013) [49] | 97.70% (3186/105) |
| Li *et al.* (2014) [28] | 95.40% (2797/105) |
| Li *et al.* (2015) [29] | **98.82%** (2797/105) |
| Emambakhsh *et al.* (2016) [17] | 95.35% (2797/105) |
| Our method (DNP + NN) | **97.89%** (2797/105) |
| Our method (DNP + SRC) | 97.82% (2797/105) |
| Our method (DNP + LS-SRC$_bos$) | 97.60% (2797/105) |

be notice that LS-SRC$_{bos}$ performs slightly worse than SRC. We guess the reason is due to the inconsistency of sample distribution between Bosphorus (34 types of action units and six types of basic emotions) and BU-3DFE (six types of basic emotions with different levels of intensity) datasets.

## 6. Conclusion and Discussion

In this paper, we propose a simple yet efficient 3D face recognition approach by employing deep surface normal representations and location-sensitive spare representation based classifier. The effectiveness of the proposed approach, including DNP, SRC, and LS-SRC and its robustness to the variations of facial expression changes are comprehensively demonstrated on the the FRGC v2.0, BU-3DFE, and Bosphorus datasets. Current method has also some limitations. For example, in this paper, we only use the "vgg-deep-face-net" to generate deep normal patterns, the effectiveness of other pre-trained deep face nets has not been evaluated. Furthermore, it will be more interesting if we can learn a end-to-end CNN for 3D face normalization, representation and recognition. Finally, we can also consider to learn the location-sensitive weights for local facial pathes around a certain landmarks.

Including improving the above limitations, in the future, we will also study some related issues such as the robustness to pose and occlusion variations, the way of location-sensitive weight learning, the effect of using 3D face scans for network fine-tuning, and the effectiveness of the proposed method for face verification, etc.

## Acknowledgment

Huibin Li was supported in part by the NSFC under grant 11401464, Chinese Postdoctoral Science Foundation under grant 2014M560785, and International Exchange Foundation of China NSFC and United Kingdom RS under grant 61711530242. Jian Sun was supported in part by the NSFC under grants 61472313 and 11622106. Liming Chen was supported in part by the French Research Agency, l'Agence Nationale de Recherche (ANR), through the Jemime project (N° contract ANR-13-CORD-0004-02) and the Biofence project (N° contract ANR-13-INSE-0004-02) and the PUF 4D Vision project funded by the Partner University Foundation.

## References

[1] F. Al-Osaimi, M. Bennamoun, and A. Mian. An expression deformation approach to non-rigid 3d face recognition. *Int.*

*J. Comput. Vision*, 81(3):302–316, Mar. 2009.

[2] N. Alyuz, B. Gokberk, and L. Akarun. Regional registration for expression resistant 3-d face recognition. *IEEE Transactions on Information Forensics and Security*, 5(3):425–440, 2010.

[3] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3d face recognition with a morphable model. In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, Sep. 2008.

[4] S. Berretti, A. D. Bimbo, and P. Pala. 3d face recognition using isogeodesic stripes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2162–2177, 2010.

[5] S. Berretti, A. D. Bimbo, and P. Pala. Superfaces: A super-resolution model for 3d faces, 2012.

[6] S. Berretti, A. del Bimbo, and P. Pala. Sparse matching of salient facial curves for recognition of 3-d faces with missing parts. *IEEE Transactions on Information Forensics and Security*, 8(2):374–389, Feb 2013.

[7] T. Bolkart and S. Wuhrer. A robust multilinear model learning framework for 3d faces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 4911–4919, 2016.

[8] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition. *Computer Vision and Image Understanding*, 101:1–15, 2006.

[9] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant representations of faces. *IEEE Transactions on Image Processing*, 16(1):188–197, Jan. 2007.

[10] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1695–1700, 2006.

[11] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation, 2012.

[12] J. Cook, V. Chandran, and C. Fookes. 3d face recognition using log-gabor templates. In *British Machine Vision Conference (BMVC)*, pages 769–778, 2006.

[13] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):518–531, March 2016.

[14] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2014.

[15] S. L. Dong Yi, Zhen Lei and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[16] H. Drira, B. B. Amor, A. Srivastava, M. Daoudi, and R. Slama. 3d face recognition under expressions, occlusions and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2013.

[17] M. Emambakhsh and A. Evans. Nasal patches and curves for expression-robust 3d face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):995–1007, May 2017.

[18] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. A region ensemble for 3d face recognition. *IEEE Transactions on Information Forensics and Security*, 3(1):62–73, 2008.

[19] Y. Guo, Y. Lei, L. Liu, Y. Wang, M. Bennamoun, and F. Sohel. Ei3d: Expression-invariant 3d face recognition based on feature and shape matching. *Pattern Recognition Letters*, 83, Part 3:403 – 412, 2016.

[20] F. Hajati, A. A. Raiea, and Y. Gao. 2.5d face recognition using patch geodesic moments. *Pattern Recognition*, 45(3):969 – 982, 2012.

[21] R. Hoffman and A. K. Jain. Segmentation and classification of range images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):608–620, 1987.

[22] D. Huang, M. Ardabilian, Y. Wang, and L. Chen. 3-d face recognition using elbp-based facial description and local feature hybrid matching. *IEEE Transactions on Information Forensics and Security*, 7(5):1551–1565, 2012.

[23] T. Huynh, R. Min, and J.-L. Dugelay. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Proceedings of the 11th International Conference on Computer Vision*, pages 133–145, 2013.

[24] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4188–4196, 2016.

[25] I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):640–649, 2007.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[27] H. Li, D. Huang, L. Chen, Y. Wang, and J.-M. Morvan. A group of facial normal descriptors for recognizing 3d identical twins. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 271–277, 2012.

[28] H. Li, D. Huang, J. Morvan, L. Chen, and Y. Wang. Expression-robust 3d face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns. *Neurocomputing*, 133:179–193, 2014.

[29] H. Li, D. Huang, J. Morvan, Y. Wang, and L. Chen. Towards 3d face recognition in the real: A registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision*, 113(2):128–142, 2015.

[30] X. Li, T. Jia, and H. Zhang. Expression-insensitive 3d face recognition using sparse representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2575–2582, Jun. 2009.

[31] X. Lu and A. K. Jain. Deformation modeling for robust 3d face matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1346–1357, 2008.

[32] A. S. Mian, M. Bennamoun, and R. A. Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1927–1943, 2007.

[33] A. S. Mian, M. Bennamoun, and R. A. Owens. Keypoint detection and local feature matching for textured 3d face recognition. *International Journal of Computer Vision*, 79(1):1–12, 2008.

[34] H. Mohammadzade and D. Hatzinakos. Iterative closest normal point for 3d face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):381–397, Feb 2013.

[35] I. Mpiperis, S. Malassiotis, and M. G. Strintzis. Bilinear models for 3-d face and facial expression recognition. *IEEE Transaction on IEEE Transactions on Information Forensics and Security*, 3:498–511, 2008.

[36] O. Ocegueda, G. Passalis, T. Theoharis, S. Shah, and I. Kakadiaris. Ur3d-c: Linear dimensionality reduction for efficient 3d face recognition. In *Proc. IEEE Int. Joint Conf. on Biometrics*, 2011.

[37] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[38] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proc. 27th Asilomar Conf. on Signals, Systems and Computers*, 1993.

[39] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2005.

[40] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3418–3427, 2016.

[41] C. Queirolo, L. Silva, O. Bellon, and M. Segundo. 3d face recognition using simulated annealing and the surface interpenetration measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):206–219, 2010.

[42] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

[43] C. Samir, A. Srivastava, and M. Daoudi. Three-dimensional face recognition using shapes of facial curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1858–1863, 2006.

[44] C. Samir, A. Srivastava, M. Daoudi, and E. Klassen. An intrinsic framework for analysis of facial surfaces. *International Journal of Computer Vision*, 82(1):80–95, 2009.

[45] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. 3d face recognition benchmarks on the bosphorus database with focus on facial expressions. In *Proc. Workshop on Biometrics and Identity Management*, 2008.

[46] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[48] D. Smeets, P. Claes, J. Hermans, D. Vandermeulen, and P. Suetens. A comparative study of 3-d face recognition under expression variations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(5):710–727, 2012.

[49] D. Smeets, J. Keustermans, D. Vandermeulen, and P. Suetens. meshsift: Local surface features for 3d face recognition under expression variations and partial data. *Computer Vision and Image Understanding*, 117(2):158–169, 2013.

[50] L. Spreeuwers. Fast and accurate 3d face recognition using registration to an intrinsic coordinate system and fusion of multiple region classifiers. *International Journal of Computer Vision*, 93(3):389–414, 2011.

[51] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, June 2014.

[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[53] V. Vijayan, K. W. Bowyer, P. J. Flynn, D. Huang, L. Chen, M. Hansen, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris. Twins 3d face recognition challenge. In *Proc. Int. Joint Conf. on Biometrics*, 2011.

[54] Y. Wang, J. Liu, and X. Tang. Robust 3d face recognition by local shape difference boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1858–1870, 2010.

[55] S. Xie, S. Shan, X. Chen, and J. Chen. Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Transactions on Image Processing*, 19(5):1349–1361, May 2010.

[56] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d facial expression database for facial behavior research. In *Proc. 7th Int. Conf. on Automatic Face and Gesture Recognition*, 2006.

[57] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3dface: A large-scale database of low-cost kinect 3d faces. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, June 2016.

[58] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, June 2016.