3D Face Recognition using Distinctiveness Enhanced Facial Representations and Local Feature Hybrid Matching

Di Huang, Guangpeng Zhang, Mohsen Ardabilian, Yunhong Wang, Member, IEEE and Liming Chen, Member, IEEE

Abstract—This paper presents a simple yet effective approach for 3D face recognition. A novel 3D facial surface representation, namely Multi-Scale Local Binary Pattern (MS-LBP) Depth Map, is proposed, which is used along with the Shape Index (SI) Map to increase the distinctiveness of smooth range faces. Scale Invariant Feature Transform (SIFT) is introduced to extract local features to enhance the robustness to pose variations. Moreover, a hybrid matching is designed for a further improved accuracy. The matching scheme combines local and holistic analysis. The former is achieved by comparing the SIFT-based features extracted from both 3D facial surface representations; while the latter performs a global constraint using facial component and configuration. Compared with the state-of-the-art, the proposed method does not require time-consuming accurate registration or any additional data in a bootstrap for training special thresholds. The rank-one recognition rate achieved on the complete FRGC v2.0 database is 96.1%. As a result of using local facial features, the approach proves to be competent for dealing with partially occluded face probes as highlighted by supplementary experiments using face masks.

I. INTRODUCTION

The face is potentially one of the best biometrics for people identification-related applications, since it is non-instructive, contactless and socially well accepted. Unfortunately, human faces are similar in their configurations and hence offer low distinctiveness, unlike the other biometrics, i.e., iris and fingerprint [1]. Furthermore, intra-class variations, due to factors as diverse as pose, facial expression etc. are often greater than inter-class ones. The past several decades have witnessed the tremendous efforts firstly focused on 2D face images [2] and more recently on 3D face scans [3]. Despite the great progress achieved so far in the field [2], 2D images are still not reliable enough [4], especially in the presence of pose and illumination changes [5]. Along with the development in 3D imaging systems, 2.5D or 3D scans have emerged as a major solution to deal with unsolved issues in 2D face recognition, such as variations of illumination and pose [3][6]. Meanwhile, though 3D face scans capture the facial surface structure, and they are therefore theoretically reputed to be robust to lighting variations, they are likely to be more sensitive to facial expression changes. In addition, they always require an accurate registration step before shape-based 3D matching.

Zhao et al. [2] categorized 2D image-based face recognition techniques into three main approaches: the holistic ones such as PCA [7] and LDA [8]; the feature-based ones such as Elastic Bunch Graph Matching (EBGM) [9]; the hybrid ones such as Component Eigenfaces [10]. This taxonomy can also be extended to 3D model-based face recognition techniques. For instance, the holistic class includes ICP (Iterative Closest Point) based matching [11], annotated deformable model [12], isometry-invariant description [13] etc. The matching scheme based on holistic facial features generally requires an accurate normalization step with respect to pose and scale, and it has proved sensitive to expression changes and partial occlusions. On the other hand, feature-based matching compares the local descriptive points or regions of 3D face scans and has been explored in several tasks in the literature, containing the point signature approach [14] and more recently, multi-modal local feature-based matching [15]. Feature-based matching has the potential advantages of being robust to facial expression, pose lighting changes and even to partial occlusions. The downside of this scheme is the difficulty extracting sufficient repeatable informative feature points from similar or smooth facial surfaces. There also exist some papers in the literature presenting hybrid matching schemes which combine global features with local ones: Region-ICP [16], multiple region-based matching [17] and the component and morphable model-based method [18]. The hybrid matching scheme is theoretically the most powerful [2]. However, it also risks inheriting both types of shortcomings: sensitivity to pose variations, difficulty generating enough stable descriptive feature points, etc.

In this study, we propose a simple yet effective approach for robust 3D face recognition. Its major contributions can be summarized as:

1) Since faces are similar and 3D facial surfaces are very smooth, a novel 3D facial representation, namely Multi-Scale Local Binary Pattern (MS-LBP) Depth Map, is proposed and used along with the Shape Index Map to efficiently highlight the distinctiveness of range data.

2) Most of the existing works in the literature depend on online ICP-based accurate registration which computes iteratively and expensively. In contrast to them, an off-the-shelf operator, SIFT, is investigated for 3D face recognition. Due to its insensitiveness to pose variations, the proposed method does not require the time-consuming registration step when dealing with nearly frontal face data, such as the FRGC v2.0 database [20].

3) A new matching step is designed to measure similarities between gallery and probe face samples. It works in a hybrid way by performing local feature-based matching based on the SIFT features and holistic matching using the facial compo-
rent and configuration constraint.

Tested on the complete FRGC v2.0 database, the proposed method displays a rank-one face recognition rate up to 96.1%, which is comparable to the best accuracy so far known in the state-of-the-art [12]. Meanwhile, the proposed approach also proves effective to deal with expression variations and partial occlusions without the need for any prior threshold by using additional data in an offline bootstrap as [12] and [21] did.

The remainder of this paper is organized as follows. The framework of the proposed approach is described in section II. Both distinctiveness enhanced facial surface representations, the MS-LBP Depth and SI Map, are introduced in section III, and section IV presents the SIFT-based local feature extraction. The hybrid matching process is illustrated in section V. Experimental results of face recognition are shown and analyzed in section VI. Section VII concludes the paper.

II. THE PROPOSED APPROACH OVERVIEW

Figure 1 illustrates the framework of the proposed method. In both gallery and probe set, only range face images are used. We calculate two distinctiveness enhanced 3D facial surface representations: the MS-LBP Depth Map and the Shape Index (SI) Map, for SIFT-based local feature extraction instead of original range images which capture smooth facial surfaces and similar facial configurations. The SI and MS-LBP Depth Map describe two different aspects of local shape of surfaces: the former shows the geometric character, i.e., spherical cup, spherical cup, saddle etc.; while the latter shows the bending direction and is achieved by introducing LBP operators using neighborhoods with different radius values to range images. Thus, combining both 3D facial representations, local shape of 3D facial surfaces can be described more comprehensively. We hence proceed to extract the SIFT-based local repeatable features. The hybrid matching process then carries out local feature-based matching based on the SIFT features as well as holistic matching using the facial component and configuration constraint. The similarity scores of the MS-LBP and SI Map are further fused for the final decision.

III. DISTINCTIVENESS ENHANCED FACIAL REPRESENTATION

As all the range images capture smooth facial surfaces and are similar in appearance, we propose to apply two intermediate 3D facial surface representations, the Multi-Scale LBP (MS-LBP) Depth and Shape Index (SI) Map, to highlight the distinctiveness of face data.

A. Multi-Scale LBP Depth Map

Local Binary Patterns (LBP), a non-parametric algorithm, was originally proposed to describe local texture attributes of 2D images [22]. The most important properties of LBP are its tolerance to monotonic illumination variations and computational simplicity. Hence, it has been extensively adopted for 2D face recognition during the last several years [23].

Specifically, the original LBP operator labels each pixel of a given image by thresholding in a $3 \times 3$ neighborhood. If the values of the neighboring pixels are not lower than that of the central pixel, their corresponding binary bits are assigned to 1; or they are assigned to 0. A binary number is thus formed by concatenating all the eight binary bits, and the resulting decimal value is utilized for labeling. Figure 2 gives a process example.

Formally, given a pixel at $(x_c, y_c)$, the derived LBP decimal value is:

$$LBP(x_c, y_c) = \sum_{n=0}^{8} s(i_n - i_c)2^n; \quad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (1)$$

where $n$ covers the eight neighbors of the central pixel, $i_c$ and $i_n$ are gray level values of the central pixel and its surrounding pixels respectively.

According to (1), LBP is invariant to monotonic gray-scale transformations preserving pixel order in local neighborhoods; therefore, it has been considered one of the most effective and popular 2D texture descriptors. Based on the definition of the original LBP operator, when it works on the depth information of range images, the generated local binary patterns can also describe local geometry structures.

The original LBP operator was extended later with various local neighborhood sizes to handle different scales. The local neighborhood of LBP is defined as a set of sampling points evenly spaced on a circle centered at the pixel to be labelled. The sampling points that do not fall exactly on the pixels are expressed using bilinear interpolation, and thus allowing any value of radius and any number of points in the neighborhood.
LBP facial representation can be achieved in two ways: one is LBP histogram, the other is LBP face. The general idea of the former is that a face can be seen as a composition of micro-patterns described by LBP. The images are divided into a certain number of local regions, from which LBP histograms are extracted. The LBP histograms are concatenated and thus contain both local and global information about the faces. The latter approach regards the corresponding decimal number of the LBP binary code as the intensity value of each pixel, and generates the LBP face.

A few LBP histogram-based tasks changed the neighborhood of the LBP operator to improve performance. By varying the value of radius \( R \), LBP of different resolutions is thus obtained, namely Multi-Scale LBP (MS-LBP). It was firstly applied to texture classification [22], and this technique was also introduced to 2D face recognition [24] [25]. In [26], Shan and Gritti investigated MS-LBP for facial expression recognition by first using MS-LBP to extract LBP histogram features, and then applying AdaBoost to learn the most discriminative bins. The boosted classifier of MS-LBP is observed to consistently outperform that of the single-scale LBP, and the selected LBP bins distribute at all scales. Hence, MS-LBP can be considered as an efficient method for facial representation.

In 3D domain, LBP histogram based approaches have been used for face recognition [19] [34]. However, LBP histogram loses 2D spatial information for representing 3D surfaces. In this study, the LBP image is investigated for 3D facial surface description. We propose to extract LBP images from original range data using the same multi-scale definition of LBP histogram. These LBP images, named MS-LBP Depth Maps, are finally utilized to represent 3D facial surfaces, which is quite a different and novel use of LBP. The local binary patterns at different radii capture local shape variations. As a result, the associated LBP Depth Maps emphasize local shape variations and thus highlight details, compared with the range face images. They are, therefore, more suitable for the extraction of local prominent features, such as SIFT-based features.

MS-LBP Depth Maps of range face image can be achieved by varying the neighborhood size of LBP operator, or by first down-sampling the range image and then adopting the LBP operator with a fixed radius. The samples are shown in Fig.4. The number of sampling points is 8, and the value of radius varies from 1 to 8. As we can see, the original range image is very smooth, while the resulting MS-LBP Depth Map contains many more local shape details, thus enhancing the distinctiveness of 3D face models.

### B. Shape Index Map

The Shape Index (SI) was firstly proposed by Koenderink and Doorn [27] to describe surface attributes of shapes. The Shape Index value at point \( p \) is defined as:

\[
S(p) = \frac{1}{2} - \frac{1}{\pi} \arctan \left( \frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \right)
\]

(2)

where \( k_1 \) and \( k_2 \) represent maximum and minimum principal curvatures respectively. As principal curvature is invariant to variations of pose, so is the Shape Index. The value of the Shape Index lies in the scale of \([0, 1]\). The local shape at point \( p \) is a spherical cup when \( S(p) = 0 \), and a spherical cap when \( S(p) = 1 \). When the value of the Shape Index varies from 0 to 1, the local shape changes from spherical cup to spherical cap. Figure 5 (c) shows an example of the SI Map, in which dark pixels represent lower values; bright pixels represent higher values. As seen in Fig.5, local shape variations are also highlighted in such a 3D facial representation. More details are provided by [11] for calculating the Shape Index.

### IV. LOCAL FEATURE EXTRACTION

The use of local feature to describe facial appearance provides the robustness to pose changes, facial expression variations and even to partial occlusions. Currently local feature extraction, i.e., anthropometric landmarks, is mostly operated on 2D or 3D smooth face images directly, leading to few local features with low distinctiveness. In this study, we propose to apply the widely-used SIFT (Scale Invariant Feature Transform) [28] features extracted from both the previous distinctiveness enhanced 3D facial surface representations: SI Map and MS-LBP Depth Map. Alternatively, other local feature descriptors can also be explored.

SIFT makes use of the scale-space Difference-of-Gaussian (DOG) to detect keypoints in images. The original image is repeatedly convolved with Gaussians of different scales separated by a constant factor \( k \) to produce an octave in scale space. As for an input image, \( I(x, y) \), the scale space is defined as a function, \( L(x, y, \sigma) \), produced from the convolution of a variable scale Gaussian \( G(x, y, \sigma) \) with the input image \( I \), and the DOG function \( D(x, y, \sigma) \) can be computed from the difference of two nearby scales:

\[
D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)
\]

\[
= L(x, y, k\sigma) - L(x, y, \sigma)
\]

Then extrema of \( D(x, y, \sigma) \) are detected by comparing each pixel with its 26 neighbors in \( 3 \times 3 \) regions at the current and adjacent scales (8 at the current scale and 9 at both adjacent scales, thus 26 in total). At each scale, the gradient magnitude,

---

Figure 3 shows different LBP neighborhoods. The notation \((P, R)\) denotes the neighborhood of \( P \) sampling points on a circle of radius \( R \).

Fig. 3. Operator examples: circular (8, 1), (16, 2), and (8, 2).

Fig. 4. LBP Depth Maps of a range image with different radii, the upper from 1 to 4, the lower from 5 to 8 (from left to right).
with the local features of all the gallery faces. Here, \( N_G \) and \( N_Si \) denote the number of matched keypoints in, respectively, the \( i_{th} \) LBP Depth Map pair, generated by the LBP operator from the range image with a parameter setting of \( L_{si} \) and the SI Map pair.

### B. Holistic Facial Matching

Unlike samples utilized in object detection, all human faces possess the same physical components and share the similar global configuration. Holistic matching is hence carried out to constrain the matched local features with respect to the facial components and configuration.

#### 1) Facial Component Constraint

We propose to divide the entire range face image into some sub-regions, each of which describes roughly a component of the nearly frontal faces, to restrict the matched keypoints of gallery and probe face scans only to those with similar physical meaning. That means the matched keypoints from the same facial component should be more important. Instead of the costly clustering process carried out in [29] to automatically construct sub-regions based on the keypoint locations from training samples, we simply make use of our common sense of facial component position, and divide the entire face region into 3 \( \times \) 3 rectangle blocks of the same size. The similarity measure of the facial component constraint is defined from this facial composition scheme. A range face image \( I \) is represented as \((m_1, m_2, \ldots, m_k)\); \( k \) is 9 in our case and \( m_i \) is the number of detected SIFT keypoints fall within the \( i_{th} \) component. The local SIFT-based descriptors in all the \( k \) components can be denoted by:

\[
I = (f_1^1, f_1^{m_1}, f_2^1, f_2^{m_2}, \ldots, f_i^1, f_i^{m_i})
\]

where \( f_i^j \) means the \( f_{ij} \) descriptor in the \( i_{th} \) facial component. Then the similarity between a gallery face \( I_g \) and a probe face \( I_p \) is computed by:

\[
C(I_p, I_g) = \frac{1}{k} \sum_{i=1}^{k} (\max_{j=1}^{m_i} (\langle f_i^j, f_i^{m_i} \rangle / ||f_i^j|| / ||f_i^{m_i}||))
\]

where \( x \in [1, \ldots, m_x]; y \in [1, \ldots, m_y]; \langle \cdot, \cdot \rangle \) denotes the inner product of two vectors, and \( || \cdot || \) denotes the norm of one vector. The bigger \( C \) indicates the more similar attribute of the two faces represented by the LBP Depth Map or SI Map. We obtain the similarity values, \( C_{Li} \) for each LBP Depth Map and \( C_{Si} \) for the SI Map.

#### 2) Facial Configuration Constraint

The former paragraph on facial component constraint emphasizes the importance of matching between the local facial features of the same component-based face region in the gallery and probe set, and we further improve the holistic constraint by facial configuration. All the range face images are normalized to a certain size to build a public coordinate system. For each 3D face model, the LBP Depth Maps and SI Map are extracted from the range image, and there is a pixel-to-pixel correspondence between these images. Therefore, all the keypoints of two facial surface representations share the same XY-plane with the range face image, and the pixel values of the corresponding range image can be regarded as the Z-axis values of these keypoints. Thus every keypoint has its position in 3D space. After local
feature-based matching, a 3D graph is formed for each LBP Depth Map or the SI Map of one probe \( F_p \), by simply linking every two of its keypoints which have a matching relationship with the corresponding keypoints on a gallery face image \( F_g \). The matched keypoints of \( F_g \) also construct a corresponding graph of \( F_p \). Intuitively, if faces \( F_g \) and \( F_p \) are from the same subject, the corresponding graphs should have similar shapes and locations in 3D space.

One similarity measure between the two graphs is

\[
d_s = \frac{1}{n_e} \sum_{i=1}^{n_e} |d_{p_i} - d_{g_i}|
\]

(8)

where \( d_{p_i} \) and \( d_{g_i} \) are the lengths of corresponding edges in the probe and gallery graphs respectively. The value \( n_e \) is the total number of edges. If the number of matched keypoints is \( n_e \), \( n_e \) will be \( n_e(n_e - 1)/2 \). Equation 8 is an efficient way to measure the spatial error between the matched keypoint pairs of probe and gallery features.

Another similarity between two graphs is calculated as the mean Euclidean distance \( d_g \) between corresponding nodes:

\[
d_g = \frac{1}{n_e} \sum_{i=1}^{n_e} |n_{p_i} - n_{g_i}|
\]

(9)

where \( n_{p_i} \) and \( n_{g_i} \) are the coordinates of corresponding keypoints in a gallery and its probe graph respectively. \( n_e \) is the number of matched keypoints.

The final similarity measure value of the facial configuration constraint is:

\[
D = w_c \cdot d_c + w_d \cdot d_g
\]

(10)

where \( w_c \) and \( w_d \) are the corresponding weights of \( d_c \) and \( d_g \), designed according to the scheme adopted in similarity fusion, and the smaller \( D \) indicates the more similar attribute of two range faces represented by the LBP Depth Map or the SI Map. As in the facial component constraint, \( D_{cf} \), denotes the similarity of each LBP Depth Map and \( D_{fS} \) denotes that of the SI Map.

C. Similarity Fusion:

In summary, the matching step of gallery and probe range face images contains three types of similarities: the number of matched keypoint pairs \( N \), similarity \( C \) of the facial component constraint as well as similarity \( D \) of the facial configuration constraint. Except for \( D \), all the other similarity measures are with a positive polarity (a bigger value means a better matching relationship). A face of probe set is matched with every face in the gallery, resulting in three vectors \( S_N \), \( S_C \) and \( S_D \). The \( n_h \) element of each score vector corresponds to the similarity measure between the probe and the \( n_h \) gallery face. Each vector is normalized to the interval of \([0, 1]\) using the min-max rule. Elements of \( S_D \) are subtracted from 1 in order to reverse its polarity. The final similarity of the probe face with each face in the gallery set is then calculated using a basic weighted sum rule:

\[
S = w_N \cdot S_N + w_C \cdot S_C + w_D \cdot (1 - S_D)
\]

(11)

The corresponding weights: \( w_N \), \( w_C \) and \( w_D \) are calculated dynamically during the online step using the scheme in [15]:

\[
w_N = \frac{\text{mean}(S_N) - \min_1(S_N)}{\text{mean}(S_N) - \min_2(S_N)}
\]

(12)

where \( i \) corresponds to the three similarities: \( N \), \( C \), and \( D \), and operators \( \min_1(S) \) and \( \min_2(S) \) produce the first and second minimum value of the vector \( S \), The gallery face image which has the maximum value in vector \( S \) is declared as the identity of the probe face image when the decision is to be made on each LBP Depth Map or the SI Map independently. Matching scores are also normalized and fused according to the same weighted sum rule to combine the scores of the LBP Depth Maps with different scales to generate the MS-LBP score, and so is the final score produced by using a fusion of the result of the MS-LBP Depth Map and the SI Map. Figure 6 shows an example of matching results using the original range image, the LBP Depth Map as well as the SI Map.

Fig. 6. An matching example: the upper shows matching between the same subject; the lower shows matching between different subjects.

VI. EXPERIMENTAL RESULTS

The experiments were carried out on the FRGC v2.0 [20], one of the most comprehensive and popular databases, containing 4007 3D face models of 466 subjects. One range face image is extracted from 3D each face model. A preprocessing step was applied to remove spikes with the median filter and fill holes using cubic interpolation. As a result of using local features, we did not perform any registration on these 3D face scans, in contrast to the work done in [12] and [30]. The range images were cropped using a basic bounding box based on the mask provided by a 3D scanner indicating if the point is valid or not. Cropped images have pose, expression variations, and partial occlusions caused by hair. All the faces are normalized to \(100 \times 100\) pixels for computation simplicity.

Fig. 7. Probes with eight types of missing quadrant.

The proposed method was tested by face recognition tasks. One 3D face scan with a neutral expression was selected from each subject to make a gallery of 466. The remaining 3D face scans (4007-466=3541) were treated as probes. The probe set was then divided into two subsets according to their expression labels to evaluate its insensitiveness to facial expression variations. The first subset contains face scans with the neutral expression; while the other one with face scans possessing non-neutral expressions. Besides the experiment of Neutral vs. All, two additional experiments of Neutral vs. Neutral and Neutral vs. Non-Neutral were also carried out. In Neutral vs. Neutral and Neutral vs. Non-Neutral, only the neutral and non-neutral probe subsets were used, respectively. In addition, we also analyze the proposed method with partially occluded probes, using eight types of face models with up to 50% of the quadrants missing (see Fig.7).
Table I shows the results based on LBP Depth Maps with different parameters. Recall that $P$ is the number of sampling points and $R$ is the radius value. As we can see, for the same $R$ value, using 8 sampling points usually achieves better results; while for the same $P$ value, the best result is often obtained by the radius value in the middle of the 1 to 8 range of $R$.

The proposed method is compared with ICP as well as the state-of-the-art: Chang et al. [31], Kakadiaris et al. [12], Mian et al. [15] [30], and Wang et al. [32]. The rank-one recognition rates are listed in Table II. Except for ICP, all the results are quoted from the original papers.

The results in brackets of (2), (3) and (4) of Table II are quoted from the original papers.

Table IV gives the performance figures when dealing with the eight proposed missing quadrants, highlighting the effectiveness of the proposed method with partially occluded face probes.

Currently, an unoptimized implementation of our approach with MATLAB can perform about 65 matches between the gallery and probe faces per second using a 2.66 GHz Pentium IV machine with 4 GB RAM.

VII. CONCLUSIONS AND DISCUSSIONS

This paper presents a simple yet effective approach for 3D face recognition. We first propose a novel 3D facial surface representation, named the Multi-Scale LBP (MS-LBP) Depth Map, which is used along with the Shape Index (SI) Map to enhance the distinctiveness of 3D facial surface. SIFT-based local features are then extracted from both 3D facial surface representations instead of directly operating on smooth and similar range faces. The proposed matching process adopts a hybrid strategy combining local feature-based matching with holistic matching. The former is achieved through comparing SIFT-based local facial features; while the latter performs the holistic constraint of facial component and configuration. By the test on the complete FRGC v2.0, our method displayed a
rank-one recognition rate up to 96.1% which is comparable to the best performance on the same dataset so far known in the literature. Meanwhile, thanks to the utilization of local facial features extracted from the 3D facial surface representations, the proposed approach does not require the time-consuming ICP-based registration step or any additional data in a bootstrap to train special thresholds. Moreover, it proves effective to deal with partial face occlusions.

The range images of 3D faces are generally pose-sensitive, though the extraction and matching step of local features like those, e.g. by SIFT, softens such sensitiveness, making our method registration-free only for nearly frontal 3D face scans. To deal with faces with severe pose variations, only a coarse pose correction based on a few automatic landmarking points, i.e. nose tip and the two inner corners of the eyes, is sufficient to adjust them to nearly frontal status as [33] did.

REFERENCES