ABSTRACT

In this paper, we present a novel approach for fusing shape and texture local binary patterns (LBP) for 3D face recognition. Using a recently proposed framework [1], we compute LBP directly on the face mesh surface, then we construct a grid of the regions on the facial surface that can accommodate global and partial descriptions. Compared to its depth-image counterpart, our approach is distinguished by the following features: a) inherits the intrinsic advantages of mesh surface (e.g., preservation of the full geometry); b) does not require normalization; c) can accommodate partial matching. In addition, it allows early-level fusion of texture and shape modalities. Through experiments conducted on the BU-3DFE and Bosphorus databases, we assess different variants of our approach with regard to facial expressions and missing data.

Index Terms— mesh-LBP, fusion, 3D face recognition

1. INTRODUCTION

3D face image has emerged as a promising modality for face recognition addressing the limitations of its 2D counterpart, such as pose and luminance variation, while opening-up new horizons for enhancing the reliability of face-based identification. This trend has been further fueled by the advances in 3D scanning technology, which provides now 3D textured scans encompassing aligned shape and photometric data.

Recently, Local Binary Patterns (LBP) [2] descriptors have been extensively used in 2D face description and representation and rapidly have been extended to the 3D modality. 3D-LBP approaches advanced the state of the art and proved to be competitive with other classes of methods. However, their applications is hindered by the intrinsic limitations of the 2D image support. Indeed, most if not all 3D-LBP approaches operate on depth images, in which depth is mapped to a gray level via 2D projection. As such, depth images require normalization to accommodate with pose variation. Yet, they still remain vulnerable to self-occlusion (caused for instance by lateral areas of the nose).

To address these problems, we propose a novel LBP-based face representation that can be constructed over triangular mesh manifolds. This representation, which is based on the recently proposed mesh-LBP concept [1], relieves the recognition process from the need for normalization, while it preserves the full 3D geometry of the shape. In another hand, given the consensus on the advantageous aspect of multi-modal face recognition [3], LBP construction on the mesh allows boosting recognition by offering an elegant framework for fusing, over a mesh support, texture and shape information at data and feature level, in addition to score and decision level. To the best of our knowledge, this work is the first one to propose texture and shape fusion for face recognition using LBP patterns constructed on the mesh. Our method encompasses the following stages: 1) Computation of LBP descriptors from the face mesh surface; 2) Construction of grid of points on the face surface, to obtain an ordered set of regions (equivalent to blocks in the 2D case); 3) Computing a histogram at each region, then concatenating the regional histograms into a structure encoding either a global or partial description of the face; 4) Performing the face matching.

The rest of the paper is organized as follows: in Sect. 2, we review previous 3D-LBP methods. In Sect. 3, we give an overview on the mesh-LBP concept. In Sect. 4 and Sect. 5, we describe, respectively, how mesh-LBP is used for constructing face representation, and for fusing shape and texture information. Experimental results are exposed in Sect. 6. Concluding remarks are discussed in Sect. 7.

2. RELATED WORK

In the literature, most if not all the LBP-based face recognition methods operate on depth images. This format allowed a straightforward application of the 2D-LBP operator as it was demonstrated in the pioneering work of Li et al. [4]. Later, Huang et al. [5, 6] proposed the multi-scale extended LBP (eLBP), which consists of several LBP codes in multiple layers accounting for the exact gray value differences between the central pixel and its neighbors. Sandbach et al. [7], introduced the local normal binary pattern (LNBP), which used the angle between normals at two points, rather than the depth value to obtain the local binary code. This novel LNBP concept has been adopted in subsequent works in different variants. Li et al. [8] extracted surface normals in 3D, then the values of the normal components along the direction of the three coordinate axes are interpreted as depth values, and LBP is computed on these depth maps reporting the values of the normal components. In a further extension, Sandbach
et al. [9] constructed images of azimuthal projection distance. The azimuthal equidistant projection is able to project normals onto points in an Euclidean space according to the direction. Though the projected information is not the depth, depending on the normals of the 3D surface, 2D LBP are still computed on the projection images. The 3D-LBP method proposed in [10], computed the difference of the depth value or the angle between the normal of a central vertex and the eight neighboring vertices on a mesh. Using this descriptor, a region based representation of the face similar to the one developed in [11] for 2D face recognition is derived. This work includes the idea of using normals computed on the mesh, but the mesh require an elaborated preprocessing in order to extract LBP constrained to the eight vertices near to a central one. Also, the circular ordering procedure of these vertices, necessary to perform LBP computation is not revealed. In addition, multi-resolution LBP is not supported, and the partitioning of the face into regions is defined based on a set of 48 landmarks manually annotated. More recently Bayramoglu et al. [12], combined a central symmetric variant (CS-3DLBP) patterns and a set of geometrical features in a decision-level fusion using a robust random forest classifier. This method operates on depth images and adopted also surface normal orientation as a shape function. All the aforementioned methods, except [10], operate on depth images, and therefore when dealing with mesh model as input have to convert it into a depth image via assiduous normalization procedures. This makes handling incomplete face scan resulting, for instance, from pose variation and occlusion, quite problematic for these methods. The method of Tang et al. [10], while it constructs LBP patterns on the mesh, lacks the elegance, the simplicity and the multi-resolution aspects of the original LBP.

3. LBP DESCRIPTORS ON THE MESH

In its simplest form, an LBP is an 8-bit binary code obtained by comparing an image pixel’s value (e.g., gray level, depth) with each pixel’s value in its 3 × 3 neighbour. The outcome of this comparison is 1 if the difference between the central pixel’s value and its neighbour pixel’s counterpart is less or equal than a certain threshold, and 0 otherwise.

Wergli et al. [1] elegantly extended the LBP concept to the mesh model by proposing a simple yet efficient technique for constructing sequences of facets ordered in a circular fashion around a central facet. The approach consists in categorizing the facets on the contour defined by a central facet’s edges in two categories, namely, the \( F_{out} \) facet and the \( F_{gap} \) facets. An \( F_{out} \) facet (respectively, an \( F_{gap} \) facet) shares an edge (respectively, a single vertex) with a central facet (referred by \( f_c \), in Fig. 1). Starting with three—clockwise or anticlockwise—ordered \( F_{out} \) facets \((f_{out1}, f_{out2}, f_{out3})\) in Fig. 1), the construction algorithm iteratively extracts the \( F_{gap} \) facets located between each pair of consecutive \( F_{out} \) facets following the same order in which the \( F_{out} \) facets have been initially arranged, and closing the loop at the pair composed by the last \( F_{out} \) facet (the third one) and the first one. The outcome of this procedure is a ring of ordered facets arranged clockwise or anticlockwise around the central facet. From this ring, a new sequence of ordered \( F_{out} \) facets located on the ring’s outer-contour can be extracted, thus allowing the ring construction procedure to be iterated, and to generate a sequence of concentric rings around the central facet (see Fig. 1).

The so obtained structure of ordered and concentric rings around a central facet forms an adequate support for computing LBP operators (referred as mesh-LBP in [1]) at different radial and azimuthal resolutions, while preserving the simplicity of the original LBP. Let \( h(f) \) be a scalar function defined on the mesh, which can incarnate either a geometric (e.g., curvature) or photometric (e.g., color) information. The mesh-LBP operator, as proposed in [1], is defined as follows:

\[
\text{meshLBP}_m(f_c) = \sum_{k=0}^{m-1} s(h(f_k^c) - h(f_c)) \cdot \alpha(k),
\]

where \( r \) is the ring number, and \( m \) is the number of facets uniformly spaced on the ring. The parameters \( r \) and \( m \) control, respectively, the radial resolution and the azimuthal quantization. \( \alpha(k) \) is a discrete function introduced for deriving different LBP variants. For example, \( \alpha(k) = 2^k \) results into the mesh counterpart of the basic LBP operator firstly suggested by Ojala et al. [2]; with \( \alpha(k) = 1 \) we obtain the sum of the
digits composing the binary pattern. In the experiments, we will refer to these two functions by $\alpha_1$ and $\alpha_2$, respectively.

4. FACE REPRESENTATION

In the standard LBP-based face representation [11], the 2D face image is divided into a grid of rectangular blocks, then histograms of LBP descriptors are extracted from each block and concatenated afterwards to form a global description of the face. To extend this scheme to the face manifold, we need first to partition the facial surface into a grid of regions (the counterpart of the blocks in the 2D-LBP), compute their corresponding histograms, and then group them into a single structure. To this end, first the plane formed by the nose tip and the two eyes inner-corner landmark points is initially computed (see Fig. 2(a)). These three landmarks are used since they are the most accurate detectable landmarks on the face, in addition to be robust to facial expressions. From these landmarks we derive, via simple geometric calculation, an ordered and regularly spaced set of points on that plane. Afterwards, the plane is tilted slightly, by a constant amount, to make it more aligned with the face orientation, and then we project this set of points on the face surface, along the plane’s normal direction. The outcome of this procedure is an ordered grid of points (see Fig. 2(b)), which defines an atlas for the facial regions that divide the facial surface. Around each grid point, we extract a neighborhood of facets. These can be defined by the set of facets confined within a geodesic disc or a sphere, centered at that grid point.

Fig. 2. (a) Construction of the face grid; (b) Partition of the grid points into a top, middle and bottom band.

Then, we compute a multi-resolution mesh-LBP descriptor using (1) for each facet in a region, considering both shape-valued and texture-valued functions (gray level, $GL$). Finally, histograms of these descriptors are computed and integrated into a single histogram describing either the whole face or part of it (see Fig. 3(a)).

5. FUSION SCHEMES

Four levels of fusion are considered in biometry applications, namely, data, feature, score, and decision [13]. As mentioned in [14], it is believed that low-level fusion performs better than its higher level counterparts (score and decision) [15]. Looking at the spectrum of region methods fusing texture and 3D shape modalities, we found much concentration in the score-level category [4, 16, 17, 18, 19], as compared to the feature-level [4, 20, 21]. The work of [4], in particular, fused LBP features derived from depth and texture image.

In our approach, we have investigated a score-level fusion and two variants of feature-level fusion. We have chosen the sum rule for the score-level, as it has been proven to be the optimal one [22]. In the first variant of the feature-level, we concatenate the two mesh-LBP regional histograms, corresponding to the shape and the texture functions. For example, considering an azimuthal quantization $m = 12$ and $\alpha_1$, we obtain a 13-bins histogram for each function, thus leading to a one-dimensional 26-bins histogram for each radial resolution $r$, that is a $r \times 26$ histogram. In the second variant, we used a 2-D accumulator that counts for the co-occurrences of the mesh-LBP corresponding to the shape and the texture functions. For the same aforementioned parameters’ values, we obtain an $r \times 13 \times 13$ histogram (Fig. 3(b) depicts some examples). In the rest of the paper, we will refer to these first and second variants by $FF1$ and $FF2$ respectively, whereas the score-level fusion will be referred by $SF$.

6. EXPERIMENTS

We evaluated our approach on the BU-3DFE database from Binghamton University [23], and the Bosphorus database [24].

The BU-3DFE contains scans of 56 males and 44 females, acquired in a neutral plus six different expressions (anger, disgust, fear, happiness, sadness, and surprise). Apart of the neutral expression, all the other facial expressions have been acquired at four levels of intensity (levels 1 to 4). This combination results in a total of 2500 scans, with both texture and shape data. We considered as gallery and probe the sets of neutral scans and the expression scans, respectively.

We set the radial resolution $r$ and the azimuthal quantization $m$ to 7 and 12, respectively. The choice of 12 for $m$ is justified by the fact that the number of facets in the first ring is always equal to 12 for regular meshes, regardless of the res-
olution [1]. Choosing this value allows then to account for all facets in the first ring. This number is used for the subsequent rings, so as to have patterns taking values in the same range. The number of rings \( r \) is related to the resolution of the mesh. The rationale behind the choice of \( r \) is to cover an area around a point of the sampling grid wide enough to capture local surface information. With the BU-3DFE mesh, we found that \( r = 7 \) covers about 7mm around the point, representing a good compromise. To account for the effects of facial expressions, we segmented the grid points into three bands, dubbed top (T), middle (M) and bottom (B). Then we tested our recognition approach considering the full grid (TMB) and the top and middle bands (TM) only (see Fig. 2(b)). The TMB and the TM grids contain 35 and 26 points, respectively. For the choice of the local descriptors we tested a variety of descriptors that include the mean (\( H \)) and the Gaussian (\( K \)) curvatures, the curvedness (\( C \)), and the shape index (SI), in combination with \( \alpha_1 \) and \( \alpha_2 \). We found that the \( H \) and \( C \) descriptors perform better than the rest, so we will report results related to these descriptors, mainly.

Results are reported in Table 1 for different fusion and descriptors of our method and for three variants of the interest points method proposed in [25]. Methods in [21, 26] also used the BU-3DFE database for 3D face recognition, but they are not directly comparable with our due to the different settings. From Table 1, we first notice that our method outperforms [25] even with variants using single modality (see scored related to \( H \), \( C \) and \( GL \)). In particular, we obtained full recognition rate for the surprise category. The disgust category, which is the most radical expression, exhibited the lowest rate (93.5% for lower levels 1& 2). The distribution of the best scores, highlighted in bold, clearly indicates the recognition enhancement brought by the fusion schemes.

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Level 1 &amp; 2</th>
<th>Level 3 &amp; 4</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>[25] HOG</td>
<td>88.30</td>
<td>97.90</td>
<td>83.80</td>
</tr>
<tr>
<td>SHOT</td>
<td>91.60</td>
<td>83.40</td>
<td>87.50</td>
</tr>
<tr>
<td>GH</td>
<td>86.50</td>
<td>81.50</td>
<td>84.00</td>
</tr>
<tr>
<td>TM ( \alpha_2 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>96.4</td>
<td>90.9</td>
<td>93.7</td>
</tr>
<tr>
<td>C</td>
<td>96.3</td>
<td>90.0</td>
<td>91.2</td>
</tr>
<tr>
<td>GL</td>
<td>87.6</td>
<td>75.8</td>
<td>81.7</td>
</tr>
<tr>
<td>FF1 H</td>
<td>97.3</td>
<td>93.4</td>
<td>95.4</td>
</tr>
<tr>
<td>FF1 C</td>
<td>97.1</td>
<td>91.7</td>
<td>94.4</td>
</tr>
<tr>
<td>SF H</td>
<td>97.3</td>
<td>93.4</td>
<td>95.4</td>
</tr>
<tr>
<td>SF C</td>
<td>97.1</td>
<td>91.7</td>
<td>94.4</td>
</tr>
</tbody>
</table>

Table 1. BU-3DFE: Recognition rate obtained for the different expression subsets compared with [25].

The Bosphorus database [24], contains 4666 scans of 105 subjects. The subjects were scanned in different poses, action units, and occlusion conditions. We assessed our method on the database’s subsets corresponding to Neutral and expressive scans, Lower Face Action Unit (LFAU), Upper Face Action Unit (UFAU) and Combined Action Unit (CAU), comparing it with Beretti et al. [25] and Li et al. [27], which used the same experimental protocol. Sandbach [9] and Bayramoglu [12] used also the same database, but their purpose and setting are different from ours. Table 2 depicts the comparison results. We can notice that our method neatly outperforms [25] and [27] in most of the cases, performing reasonably across the different subsets. For the intra-comparison side, referring to the computational cost and pattern repeatability, the \( \alpha_1 \) variant is more appealing than \( \alpha_2 \). This last takes advantage, theoretically, in its discriminative power, given the wider range of its related pattern value. While the results confirm the superiority of \( \alpha_2 \) variant overall, we notice that at some instances, \( \alpha_1 \) equates or performs better than \( \alpha_2 \). While we do not have a definitive postulate explaining such a consistency, we believe that the most plausible one is the intrinsic repeatability of the \( \alpha_1 \) variant.

<table>
<thead>
<tr>
<th></th>
<th>[27]</th>
<th>[25]</th>
<th>TMB ( \alpha_1 ) H</th>
<th>TMB ( \alpha_2 ) H</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF2</td>
<td>90.9</td>
<td>90.9</td>
<td>90.9</td>
<td>90.9</td>
</tr>
<tr>
<td>SF1</td>
<td>91.7</td>
<td>91.7</td>
<td>91.7</td>
<td>91.7</td>
</tr>
<tr>
<td>SF2</td>
<td>95.4</td>
<td>95.4</td>
<td>95.4</td>
<td>95.4</td>
</tr>
</tbody>
</table>

Table 2. Bosphorus: Rank-1 recognition accuracy obtained with a selection of our method variants compared with to [25] and [27]. The maximum recognition rate obtained for each subset is highlighted in bold.

7. CONCLUSIONS

In this paper, we presented an original approach for constructing a multi-modal LBP-based face representation on a triangular mesh-model. It is the first approach of its kind that integrates texture and shape information in LBP-patterns derived from a mesh support. This marriage between mesh-model and LBP-based face recognition will open-up new horizons that go quite beyond the limits imposed by the depth image constraints. We proposed a face representation that encompasses a face-centric grid, at each point of which, LBP histograms constructed using geometric and photometric data are attached. Contrary to its depth-image counterpart, this representation supports partial facial matching, and does not require normalization. In addition, it preserves the full geometry of the facial shape, which might be partially lost in depth images because of self-occlusion. The experiments conducted with BU-3DFE and Bosphorus database showcased the boosting of the recognition performance brought by our fusion framework, and its superiority with regard to the most closest approaches.
8. REFERENCES


