Eurographics Workshop on 3D Object Retrieval (2019)
S. Biasotti, G. Lavoué, B. Falcidieno, and I. Pratikakis (Editors)

Depth-Based Face Recognition by Learning from 3D-LBP Images

João Baptista Cardia Neto[1], Aparecido Nilceu Marana[2], Claudio Ferrari[3], Stefano Berretti[3] and Alberto Del Bimbo[3]

[1]São Carlos Federal University (UFSCAR), São Carlos SP 13565-905, Brazil
[2]São Paulo State University (UNESP), Bauru SP 17033-360, Brazil
[3]Media Integration and Communication Center, University of Florence, Italy

Abstract

In this paper, we propose a hybrid framework for face recognition from depth images, which is both effective and efficient. It consists of two main stages: First, the 3DLBP operator is applied to the raw depth data of the face, and used to build the corresponding descriptor images (DIs). However, such operator quantizes relative depth differences over/under ±7 to the same bin, so as to generate a fixed dimensional descriptor. To account for this behavior, we also propose a modification of the traditional operator that encodes depth differences using a sigmoid function. Then, a not-so-deep (shallow) convolutional neural network (SCNN) has been designed that learns from the DIs. This architecture showed two main advantages over the direct application of deep-CNN (DCNN) to depth images of the face: On the one hand, the DIs are capable of enriching the raw depth data, emphasizing relevant traits of the face, while reducing their acquisition noise. This resulted decisive in improving the learning capability of the network; On the other, the DIs capture low-level features of the face, thus playing the role for the SCNN as the first layers do in a DCNN architecture. In this way, the SCNN we have designed has much less layers and can be trained more easily and faster. Extensive experiments on low- and high-resolution depth face datasets confirmed us the above advantages, showing results that are comparable or superior to the state-of-the-art, using by far less training data, time, and memory occupancy of the network.

CCS Concepts

• ***Computing methodologies*** → ***Biometrics; Neural networks;*** *Matching;*

1. Introduction

In recent years, many works have demonstrated the potential of applying Deep Convolutional Neural Networks (DCNN) in the facial recognition task from static images and videos [PVZ15, SKP15, TYRW14,SWT14,YL18]. Despite the success of facial recognition based on DCNN, many aspects of the behavior of such networks, their strengths and limitations are not yet fully understood. In addition, various problems related to revealing a person's identity from his/her facial images remain open, with new perspective challenges that still need a solution to advance the field into completely realistic scenarios. One tendency is to experiment with more stringent protocols, such as *open set* and *open world* face recognition, in contrast to the classic and easier ones where a closed set of identities is given. Another trend is to consider more difficult data in terms of quality and resolution, quantity and source, extending the imaging modes to less conventional ones, including infrared or depth. In most of the cases, this accompanies with deeper and deeper network designs that, as such, require ever-increasing amounts of data to robustly train their internal parameters. The result is the need for machines with high parallelism and extreme computational power, but such infrastructures are expensive and may not be accessible in many cases.

The above considerations suggested us that in facial recognition based on DCNNs there are some aspects that have been under-investigated. This is particularly true when not so conventional imaging modalities, such as infrared or depth, are taken into consideration. In fact, there are several low-cost/low-resolution devices that can capture RGB-D data (*e.g.*, Kinect camera). The depth resolution of such devices is not comparable to that achieved by high-resolution 3D scanners; however, those may be still more effective than RGB images in some cases, such as when lighting conditions are difficult or large facial expressions should be considered.

This work focuses on the problem of face recognition from low-resolution depth data as acquired by a Kinect like sensor. The main goal is to demonstrate that when utilizing low-resolution depth data, comparable or even better face recognition results can be achieved with shallow network architectures and much fewer training data. In doing so, one founding idea of the present solution is that of using a descriptor image (DI) built by encoding hand-crafted low-level features in place of the depth data, in order to capture and emphasize details of the face. We propose a modified version of the 3DLBP [HWT06] descriptor for such task: 3DLBP is computed from the depth image producing a DI. Then, we designed a shallow convolutional neural network (SCNN) that learns

DOI: 10.2312/3dor.20191062

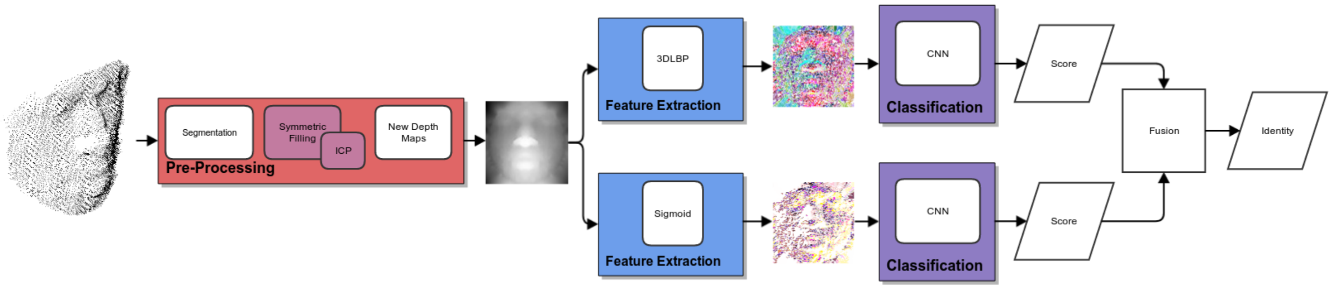


Figure 1: *Overview of the proposed method. First, some pre-processing operations (*i.e.*, noise removal, holes closing, and pose normalization) are applied to the 3D face scans given in input; then, a depth image of the face is derived; the next step is to compute the traditional 3DLBP and Sigmoid descriptors images; CNN training and classification is performed on the descriptor images, and the scores for each type of image descriptor are fused to derive the subject's identity.*

on the top of these DIs. Since DIs encode low-level image features similar to what the first layers do in a deep architecture, the resulting network has a smaller number of layers and can thus be trained with a limited amount of data with reduced computational cost. Figure 1 summarizes the two-stage architecture of the proposed approach. Results on the EURECOM low-resolution depth face dataset, and on the Bosphorus high-resolution 3D face dataset show a good compromise in the performance on both low and high-resolution data. This cannot be achieved by existing state-of-the-art DCNN solutions that are tailored to either low or high-resolution data. In summary, there are two main contributions in the proposed face recognition method from depth data:

- To the best of our knowledge, this work is the first one that combines hand-crafted features and a specifically designed shallow network in a framework that can be trained quickly and with a limited number of examples;
- We demonstrate through a comprehensive experimental validation that the proposed approach can effectively be applied either to high or low-resolution data as captured by RGB-D cameras.

The rest of the paper is organized as follows: In Section 2, we summarize the works in the literature that are more close to our proposed solution; In Section 3, the main aspects of the 3D LBP descriptor computation and our proposed variant are reported; The combination of such depth descriptor with a newly designed shallow network architecture is proposed in Section 4; An extensive experimental evaluation on low- and high-resolution depth data is presented in Section 5; Finally, in Section 6, we report some discussion on the relevant aspects of our method and the conclusions.

2. Related work

In this section, we summarize the works in the literature that focus on 3D face recognition and, more specifically, on low-resolution depth data. Those methods are classified into two groups: methods that rely on hand-crafted features, and methods that are based on CNN architectures.

Hand-crafted features based 3D face recognition: In recent times, the literature on 3D face recognition has been focusing on methods that describe surfaces, specifically for capturing geometric properties of the facial geometry [DBS*13, FBF08, MBO07, BDP10, KPT*07, Spr11]. Even with the several advances, most of the work that has been done utilizes high-resolution data in controlled environments with costly devices. Some of the works that utilize Kinect-like devices try to increase the resolution of the data utilizing the temporal redundancy of frames in a sequence [BPBD16, DMT13, HCM12]; The main problem of those approaches is the increase in the necessary computational power and, as a consequence, they are unable to operate in real time.

Only a handful of methods performs face recognition directly from low-resolution data. In the work proposed by Min *et al.* [MCMD12] a real-time 3D face identification system that receives a depth sequence as input is built. Initially, the region of the face is detected and segmented by utilizing a threshold on the depth values. The next step is to reduce the faces to common resolutions and the matching is obtained by registering a probe with several intermediate references in the gallery with the EM-ICP algorithm. Mantecon *et al.* [MdJG14] proposed the Depth Local Quantized Pattern as a modification of the original LBP operator. This modification introduces a quantification step that allows the descriptor to distinguish between different patterns. The descriptor has been used to train and test an SVM classifier. In another work, Mantecon *et al.* [MdJG16] proposed an algorithm for face recognition based on an image descriptor called bag of dense derivative depth patterns. Dense spatial derivatives are first computed and quantized in a face-adaptive fashion to encode the 3D local structure. Then, a multi-bag of words creates a compact vector description from the quantized derivatives.

CNN-based 3D face recognition: The development of deep architectures that deal with 3D data has had a slower expansion than the image-based counterpart, mainly because of the data representation problem; while CNNs were designed to work with 2D images, the wide variety of 3D data (*e.g.*, point-clouds, triangular meshes, etc.) makes it difficult to work in the same standardized way without making significant modifications in the whole framework. An example of a possible way to make use of existing deep architectures for 3D face recognition is the work proposed by Kim *et al.* [KHCM17], where the authors utilized a pre-trained version of the VGG-Face and fine-tuned it for depth data. To deal with the shortage of depth data for training, the authors expanded the dataset by generating expressions and occlusions. With the necessity of

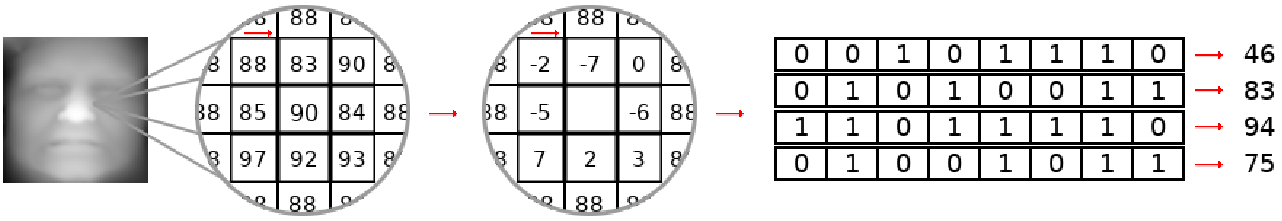


Figure 2: *3DLBP computation on a depth image of the face. A* 3×3 *neighborhood region of a central pixel with depth value of* 90 *is shown; First, the difference between the neighborhood and the central pixel is computed (clockwise, starting from the top-left pixel); Then, each column of four bits encodes the difference value using the first bit for the sign (0 for negative, 1 for zero or positive values), and the three subsequent bits for the absolute value. Using a positional encoding for the bits in each row, the 3DLBP values associated to the central pixel are computed and reported on the right.*

large amounts of data, Gilani *et al.* [GM18] proposed a synthetic data generation technique that they used to build a dataset of $\approx 3M$ scans. The authors utilized such data to train a deep architecture that follows a VGG-like structure and consists of 13 convolutional layers, 3 fully connected layers and the final softmax layer. One of the conclusions reached in this work is that, because of the smooth nature of the face surface, there is the need for larger kernels for the convolutional filters with respect to the ones commonly used. Lee *et al.* [LCTL16a], proposed a face recognition system based on deep learning that utilizes face images captured with a consumer-level RGB-D camera. For this task three steps are performed: depth image recovery, deep learning for feature extraction, and joint classification. To alleviate the problem of the limited size of available RGB-D data for deep learning, the deep network is firstly trained with a standard RGB face dataset and later fine-tuned on depth face images for transfer learning. The main difference between this work and previously cited one is that it focuses on low-resolution data instead of high-resolution.

3. 3DLBP descriptor images

Utilizing depth images from consumer like cameras (*i.e.*, Kinect) can be an interesting solution to perform 3D face recognition in unconstrained environments. In this case, the main problem is that normally this type of devices obtain data with fewer details of the face compared to those acquired by high-resolution scanners. In this sense, training from scratch a DCNN on such data is difficult for two main reasons: the nature of the data is less smooth than for RGB images; large volumes of depth images of the face with subject labels are difficult to collect (*i.e.*, the web does not represent a viable source of data in this case). The most practiced solution in the literature is that of taking a DCNN architecture pre-trained on RGB data and fine tune it with a small set of depth images.

In this work, we propose a different approach, where the learning tools are applied to the top of intermediate images generated from the original raw data by applying a low-level feature extractor. With this approach, the 3DLBP [HWT06] feature is a potential candidate due to its computational efficiency and the fact it has been proven to be effective in describing 3D depth images of the face [CNM18]. The 3DLBP is a variation of the traditional LBP as originally proposed by Ojala *et al.* [OPH96]. Its computation starts in a 3×3 region (containing 8 neighbors) defined around a center pixel, or more generally a region with radius R in which P points are sampled (in this latter case, if a point does not fall into a position with a defined value, a bi-linear interpolation is utilized). The depth value of the central pixel is subtracted from its neighbors and each of those values is truncated in the range [-7,+7]. This truncation is motivated by the fact the face is a smooth surface and the majority of those differences fall between that range [HWT06]. In the subsequent step, the depth differences are encoded as a feature. With the [-7,+7] range, 15 different values are to be encoded, which results in a four-bit representation. Each bit is regarded as a separate layer: the first layer encodes the sign of the difference, *i.e.*, 0 if the difference is negative, 1 otherwise; the other layers encode the absolute value of the difference transformed in a binary code of 3 bits. Figure 2 shows the generation of a 3DLBP descriptor for a pixel with a 3×3 neighborhood region. The 3DLBP descriptors of the whole depth are then transformed into an image; each one of the four bits of the 3DLBP is regarded as a separate channel of the final descriptor image (DI). To encode four bits, we used a four-channel RBGA image, being the last one the alpha channel.

Actually, a possible limitation of the standard 3DLBP approach is that, within the [-7,+7] interval, negative or positive difference values share the same binary code, except for the sign bit. This implies that some regions of the resulting DI might have the same values on three out of four channels. One way to account for this is to incorporate a sigmoid function in the computation of the 3DLBP operator. In this case, instead of using four bits for representing values in the [-7,+7] range, while truncating the exceeding values, a sigmoid function is used to map larger intervals to four bits:

$$f(x) = \frac{1}{1+\exp(-x)}\ , \tag{1}$$

where x is the depth difference between a point in the neighborhood and its center. To encode the sigmoid values, 8 bins are defined in the interval between 0 and 1. Then, each $f(x)$ is mapped to its closest bin, in a histogram-like fashion. Note that, even though the sign channel is maintained to build the four-channel image, $f(x)$ is computed considering the depth difference along with the sign, so that same values with opposite sign are put into different bins. This ensures that different maps with respect to the classic 3DLBP are generated. This has the advantage of encoding a larger range of variations, though with a coarser resolution. Figure 3 compares the

DI obtained using the standard 3DLBP approach in (a), with the DI derived by applying the sigmoid correction in (b).

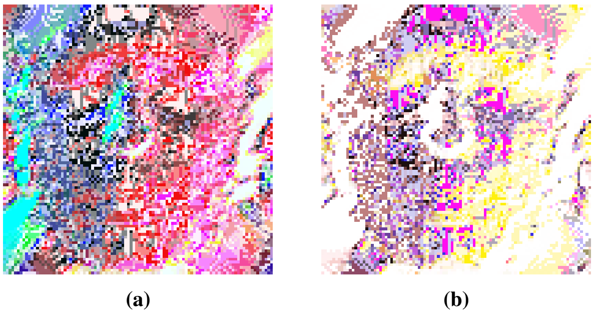


Figure 3: *Examples of the descriptor image obtained by computing: a) the standard (original) 3DLBP operator; (b) the Sigmoid 3DLBP (sample subject from the EURECOM Kinect face dataset). It is possible to see the color representation after the feature extraction process radically changes depending on the strategy.*

4. Shallow network architecture

The main idea behind this work is that the DIs computed in Section 3 provide data that are richer compared to the original depth images, where several details of the face have been enhanced. With this in mind, we expect it is possible to utilize much shallower CNNs with way less parameters to learn. The network architecture we have designed is constructed with three convolutional layers as illustrated in Figure 4: the first one operates with 64 filters, an 8×8 kernel and a stride of 4; The next convolutional layer has 192 filters, a kernel of 5 and a stride of 1; the last convolutional layer has 384 filters and a kernel of 3. After the second and last convolutional layer, there is a max pooling layer with a kernel of 3 and stride of 2. After the last max pooling layer, there are two fully connected layers and a softmax at the output for subject classification. The network has a total of 31,648 trainable parameters, excluding the softmax layer. The code for the network is publicly distributed[†].

To better assess the potential of the above architecture, and the impact of each layer on the results, we have evaluated our approach with three different network configurations. Initially, we considered a very shallow network with only one convolutional layer ("Experiment 1" in Figure 4); The second experiment is performed using a network with one more convolutional layer and a max pooling layer ("Experiment 2" in Figure 4); In the last experiment, the full network architecture is used with three convolutional layers and two max-pooling layers ("Experiment 3" in Figure 4). The experiments and results are discussed in the following section.

5. Experimental results

In this section, the results of the proposed approach are presented. Initially, we describe the databases used; next, the network training details and an ablation study on the network architecture and the DIs are presented; finally, we compare our best configurations with respect to the state-of-the-art on both high and low-resolution depth data.

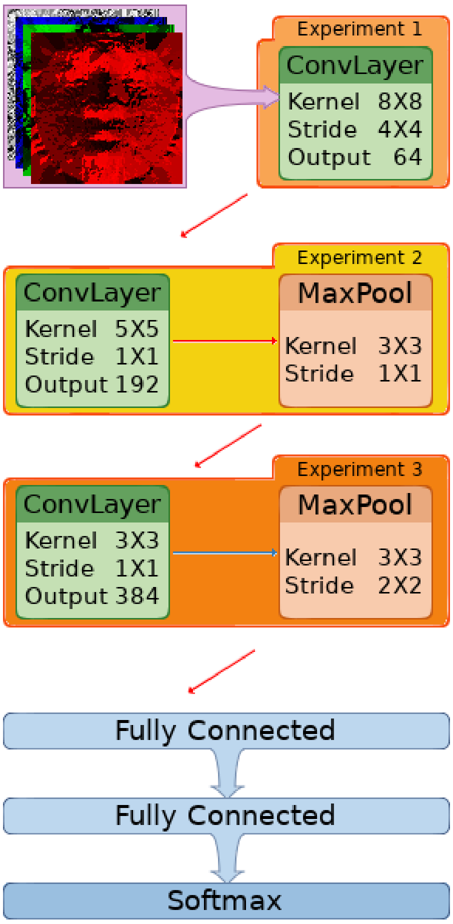


Figure 4: *Architecture of the proposed network. The network takes the DIs as input (four channel images in RGBA format). Three variants of the network, indicated in the diagram as "Experiment 1," 2, and 3 have been evaluated in the experiments.*

5.1. Datasets

The datasets used in this work are: *(i)* the Face Recognition Grand Challenge v2.0 (FRGC) [PFS*05], *(ii)* the Bosphorus [SAD*08], and the *(iii)* EURECOM [MKD14]. The FRGC is used to train our shallow network from scratch, while tests are conducted on the Bosphorus and EURECOM, where the gallery is used for the fine-tuning process of the network and train a new softmax layer for the final classification.

FRGC – The FRGC dataset includes 4007 high-resolution scans of 466 different individuals acquired in two separated sessions. About 60% of the scans have neutral expression, while the rest show slight spontaneous expressions.

Bosphorus – The Bosphorus dataset comprises 4,666 high-resolution scans of 150 individuals; there are about 54 scans per

[†] https://github.com/jbcnrlz/biometricprocessing

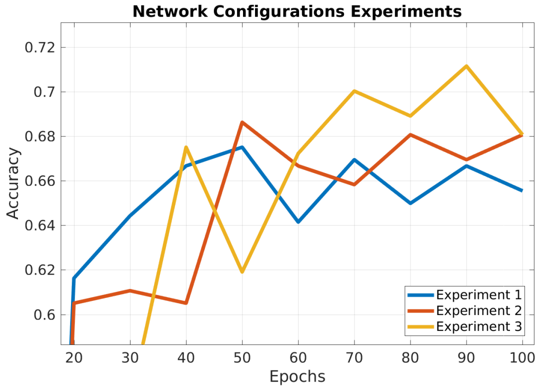


Figure 5: *Rank-1 recognition on the EURECOM dataset as a function of the number of training epochs for the three network configurations with, respectively, 1, 2 or 3 convolutional layers.*

subject, which include expression variations, facial action units activation, rotations and occlusions. This dataset was used for evaluating the framework effectiveness on high-resolution data.

EURECOM – The EURECOM database collects RGB-D images acquired with a Kinect sensor of 52 subjects, taken in two separate sessions with 7 variations each: neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion and open mouth. This dataset is employed for evaluating our approach with low-resolution data.

5.2. Training details

We used the FRGC dataset to train the network for 200 epochs. We augmented the training set by applying a 3D rotation in the Y (pitch angle) and X (yaw angle) axis from -30 to +30 degrees to each training face. The final number of images for training is 48,867. Before validation, the network undergoes through a fine-tuning process for 100 epochs and the data utilized for the fine-tuning process is augmented in the same way as the FRGC data. When fine-tuning, the last fully connected layer, *i.e.*, the classification layer, is substituted and re-trained so as to account for the new identities.

To perform recognition, in the following we also employed a score fusion between the 3DLBP and the Sigmoid approaches, in order to investigate whether a coarser encoding of larger depth intervals could bring complementary information to the original formulation. The fusion is a weighted sum of the classification scores:

$$FSCORE = 3DLBP_W * 3DLBP_S + SIG_W * SIG_S \qquad (2)$$

being $FSCORE$ the final score for a subject, $3DLBP_W$ the weight for the 3DLBP approach, $3DLBP_S$ the original score for the 3DLBP, SIG_W the weight for the sigmoid approach, and SIG_S the score for the original sigmoid.

5.3. Network depth analysis

The focus of this set of experiments is to evaluate the impact of the number of convolutional layers on the recognition accuracy. These refer to "Experiment 1," 2 and 3 depicted in Figure 4. To this aim, the EURECOM Kinect face dataset was used. In this experiment, the gallery is composed of images from Session 1 with all the variations mentioned except the "paper occlusion" one, while probe images come from the same classes in Session 2. The network was trained from scratch for 100 epochs and, as it is possible to observe in Figure 5, the best result is obtained when three convolutional layers are used. For all the subsequent experiments, we will employ this latter configuration; nevertheless, the gap between the results is rather slight, making our framework versatile and suitable also for lower-powered devices.

5.4. Results on high-resolution data

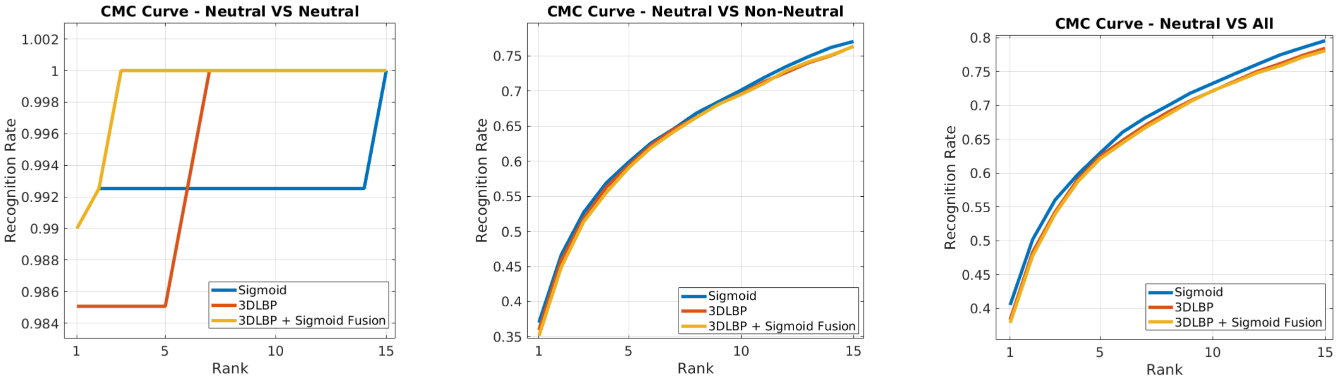
In this section, the results obtained on the high-resolution scans of the Bosphorus database for our best performing configuration (*i.e.*, with 3 convolutional layers) are reported and compared with the state-of-the-art. For this experiment, three protocols are examined (gallery vs. probe set): *(i)* Neutral vs. Neutral (N vs. N), *(ii)* Neutral vs. Non-Neutral (N vs. NN), and *(iii)* Neutral vs. All (N vs. A), which includes neutral scans as well. In our approach, the gallery for the Bosphorus is augmented by rotating each face on the X and Y axis and generating synthetic expressive faces of the gallery subjects. Even with the data augmentation step the performance radically drop when several types of rotation and expression are presented on the probe, as emerges from the CMC curves in Figure 6. Another thing to be noted is that the sigmoid approach has a slight advantage over the traditional 3DLBP.

Analyzing the results in a more deepened way, we reached the conclusion that most of the errors are introduced by the rotated scans; this because *(i)* rotated scans undergo self-occlusions and thus information is missing, and *(ii)*, the pre-processing operations involving the cropping of the central face region is highly prone to failures. One way to confirm those suspicions is to look at the results illustrated in Figure 7, when the rotations are removed from the probe and become part of the gallery. This also suggests us that the representation is quite robust to expressions, which are still present in the probe set. However, this can be seen as a limitation of the current proposal and might be ascribable to the characteristics of the 3DLBP operator; still, results show that a hybrid solution is feasible and such limitation can eventually to be accounted for. Indeed, in the Neutral vs. Neutral scenario, the system performed very accurately.

Method	*N vs. N*	*N vs. NN*	*N vs. A*
$3DLBP_W$	0.4	0.9	0.9
SIG_W	0.6	0.1	0.1

Table 1: *Values of the weights in the weighted sum score fusion for the Bosphorus dataset in the Neutral vs. Neutral (N vs. N), Neutral vs. Non-Neutral (N vs. NN), Neutral vs. All (N vs. A) experiments. These values were defined with empirical tests.*

One important thing to be noted is that the score fusion only increases the system performance when both approaches have acceptable results. Table 1 reports the values of the score weights of Eq. (2) that were used in each experiment (best performing values).



(a) *Neutral vs. Neutral* **(b)** *Neutral vs. Non-Neutral* **(c)** *Neutral vs. All*

Figure 6: *Bosphorus dataset: CMC curves for the three experiments with the proposed approach. The scales on the vertical axis are different.*

It is possible to see that, when both approaches perform well, the weights are more balanced, as in the Neutral vs. Neutral setting. In the other experiments, the weights are totally unbalanced with the 3DLBP original approach contributing more than the sigmoid one. In the Neutral vs. Neutral experiment, the system improves its performance after the rank-2, getting to 100% rank-1 sooner. In the experiment illustrated in Figure 7, there is instead a huge improvement caused by the score fusion. In this case, as in the previous one, the 3DLBP operator contributes more to the fusion than the sigmoid; this can indicate that, in the presence of expressions, the sigmoid approach brings complementary information that is critical for effective recognition.

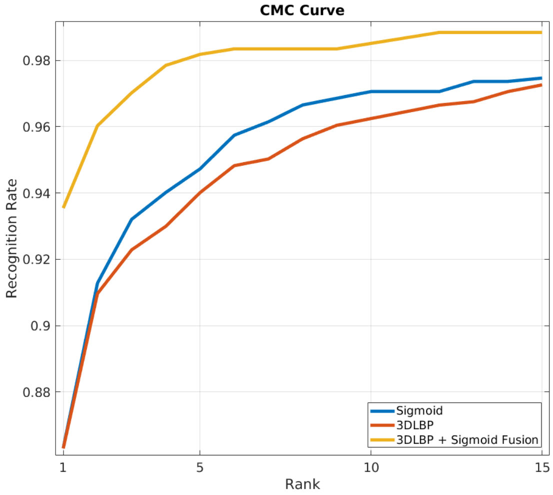


Figure 7: *Bosphorus dataset: CMC curves for the experiment using neutral and rotated face images for the gallery, and occluded, neutral, and expression images for the probe. In this experiment the value of* $3DLBP_W$ *is 0.9 and* SIG_W *0.1.*

Table 2 compares the results of the first experiment (*i.e.*, Neutral vs. Neutral) with respect to state-of-the-art methods. In particular, we considered the DCNN based approach by Kim *et al.* [KHCM17], and the hand-crafted features based approach by Li *et al.* [LHM*15]. Regarding the latter, we reported the results as presented in the original paper, while we utilized the feature extraction method publicly available in [KHCM17], but implemented the matching algorithm for the extracted feature.

Method	*Neutral vs. Neutral*
Proposed - Sigmoid	99.0%
Proposed - 3DLBP	98.0%
Proposed - Fusion	99.0%
VGG [KHCM17]	99.2%
Li *et al.* [LHM*15]	**100%**

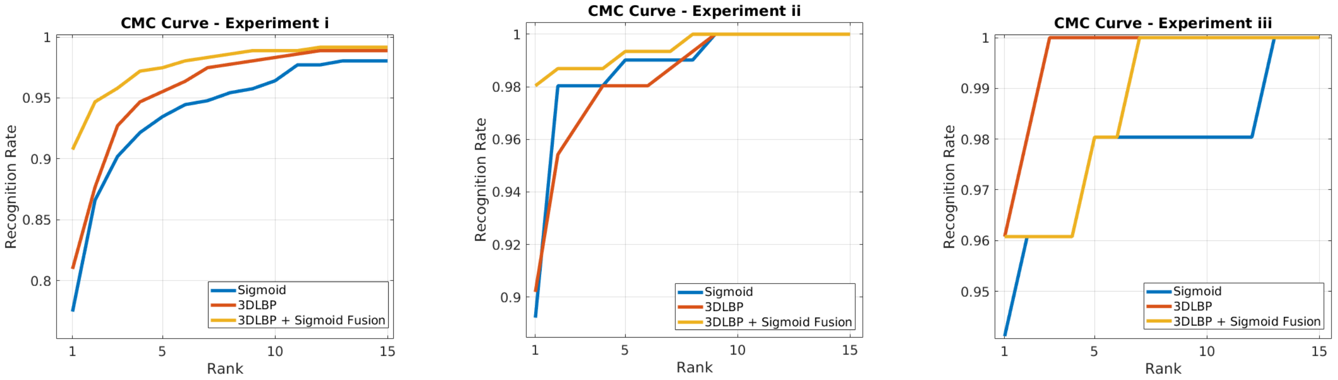
Table 2: *Bosphorus: Rank-1 results for the N vs. N protocol.*

5.5. Results on low-resolution data

In this section, the results obtained on the low-resolution Kinect scans of the EURECOM database are reported. The protocol for this experiment is structured in the following manner:

- *(i)* Gallery composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth) from Session 1, Probe composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion and open mouth) from Session 2;
- *(ii)* Gallery composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth) from Session 1, Probe composed of three variations (neutral, smile, illumination) from Session 2;
- *(iii)* Gallery composed of seven variations (neutral, smile, illumination, paper occlusion, mouth occlusion, eyes occlusion, and open mouth) from Session 1, Probe composed of one variation (neutral) from Session 2.

Also in this case, during the fine-tuning process, a new softmax layer is stacked in place of the old one and re-trained. Figure 8 shows the CMC curves for the three experiments. It is possible to



(a) **(b)** **(c)**

Figure 8: *EURECOM dataset: CMC curves for the three experiments with the proposed approach: (a) Gallery from Session 1 vs. probe from Session 2 (7 variants each), (b) Gallery from Session 1 vs. probe with the three variants without occlusions from Session 2, and (c) Gallery from Session 1 vs. neutral probes from Session 2. The scales on the vertical axis are different.*

see that there is a small advantage of the traditional 3DLBP approach over the Sigmoid, this representing the opposite of what we have obtained on high-resolution data. However, the score fusion led to a noticeable accuracy improvement. In this experiment the values for $3DLBP_W$ and SIG_W are 0.6 and 0.4, respectively, these values being the same for all the experiments. It is worth to note that the benefit of the fusion strategy is way more evident in harder protocols, and only in the experiment *(iii)* it tied with the 3DLBP approach in rank-1 recognition. Lastly, even though there is a huge difference in the data quality, the proposed method demonstrated to be effective on both high and low-resolution data.

Table 3 reports the results for the different experiments also in comparison to state-of-the-art solutions. For this experiment, the VGG [KHCM17] features are extracted in the same way as the high-resolution and, even in this case, it does not undergo any preprocessing step. The matching is performed as a Nearest-Neighbor search employing the cosine similarity between descriptors. In Lee *et al.* [LCTL16b] a pipeline to include the 3D face shape in a deep network is proposed. It performs depth face image recovery and enhancement, extraction of deep representation, and joint classification for depth and RGB data. Since our work deals with depth data only, the results reported for [LCTL16b] are those for depth data. To make the comparison possible, we designed our experiments with the same protocol as in [LCTL16b].

In the case of low-resolution data, a more balanced score fusion led to more accurate recognition. This can be a piece of evidence that, due to the coarser nature of the depth maps, it is beneficial to utilize more information from both sides.

6. Conclusions

This work proposes a new 3D face recognition approach based on depth data, which revealed to be particularly effective in the case images are captured by low-resolution cameras. Our proposed method is built on the assumption that a "hybrid" approach can be built, composed of a shallow convolutional neural network that learns from descriptor images (DIs) generated from hand-crafted feature descriptors. This solution takes advantage of both the components: 3DLBP revealed effective in emphasizing the traits of the face in the depth images, while being sufficiently robust to the acquisition noise; This permitted us to generate descriptor images of the face that can be used as training data for a shallow convolutional neural network; this learns, in an effective and efficient way, to discriminate identities from their descriptor images. We also showed that a fusion of the scores provided by the 3DLBP and a novel variant can help the performance of the approach, as illustrated in Figure 7 and in all the experiments on the EURECOM dataset.

Method	*(i)*	*(ii)*	*(iii)*
Proposed - Sigmoid	77.8%	89.2%	94.1%
Proposed - 3DLBP	80.9%	90.2%	**96.1%**
Proposed - Fusion	**90.75%**	**98.0%**	**96.1%**
VGG [KHCM17]	13.9%	14.8%	13.5%
Lee *et al.* [LCTL16b]	-	80.8%	78.8%

Table 3: *EURECOM dataset: Rank-1 results. The gallery is from Session 1, while the probe set is composed of: (i) Session 2, (ii) the three variants without occlusions from Session 2, (iii) neutral scans from Session 2.*

When dealing with low-resolution images, the proposed approach outperforms the results obtained by state-of-the-art methods, either using DCNN or hand-crafted features. With high-resolution data, the hybrid approach can compete with state-of-the-art methods, but only on the Neutral vs. Neutral setting, while rotated scans still constitute a quite severe issue for our solution. Despite this latter limitation, our approach is the first one capable of obtaining competitive performance on depth data that span a large variety of resolution. In addition to this, we proved our framework can be trained efficiently even with a limited amount of data, thus making it a viable solution for applications where hardware resources are limited both in terms of computational power and memory.

7. Acknowledgments

The Titan Xp used for this research was donated by the NVIDIA Corporation.

CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) supported the first author in the development of this work, with its PDSE grant.

References

[BDP10] Berretti S., Del Bimbo A., Pala P.: 3D face recognition using isogeodesic stripes. *IEEE Trans. on Pattern Analysis and Machine Intelligence 32*, 12 (Dec 2010), 2162–2177. 2

[BPBD16] Bondi E., Pala P., Berretti S., Del Bimbo A.: Reconstructing high-resolution face models from kinect depth sequences. *IEEE Trans. on Information Forensics and Security 11*, 12 (Dec 2016), 2843–2853. 2

[CNM18] Cardia Neto J. B., Marana A. N.: Utilizing deep learning and 3DLBP for 3D face recognition. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)* (2018), pp. 135–142. 3

[DBS*13] Drira H., Ben Amor B., Srivastava A., Daoudi M., Slama R.: 3D face recognition under expressions, occlusions, and pose variations. *IEEE Trans. on Pattern Analysis and Machine Intelligence 35*, 9 (Sept 2013), 2270–2283. 2

[DMT13] Drosou A., Moschonas P., Tzovaras D.: Robust 3D face recognition from low resolution images. In *Int. Conf. of the BIOSIG Special Interest Group* (Sept 2013), pp. 1–8. 2

[FBF08] Faltemier T. C., Bowyer K. W., Flynn P. J.: A region ensemble for 3-D face recognition. *IEEE Trans. on Information Forensics and Security 3*, 1 (March 2008), 62–73. 2

[GM18] Gilani S. Z., Mian A.: Learning from millions of 3D scans for large-scale 3D face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 1896–1905. 3

[HCM12] Hernandez M., Choi J., Medioni G.: Laser scan quality 3-D face modeling using a low-cost depth camera. In *European Signal Processing Conf. (EUSIPCO)* (Aug 2012), pp. 1995–1999. 2

[HWT06] Huang Y., Wang Y., Tan T.: Combining statistics of geometrical and correlative features for 3D face recognition. In *British Machine Vision Conf. (BMVC)* (2006), pp. 90.1–90.10. doi:10.5244/C.20.90. 1, 3

[KHCM17] Kim D., Hernandez M., Choi J., Medioni G.: Deep 3D face identification. In *IEEE Int. Joint Conf. on Biometrics (IJCB)* (Oct 2017), pp. 133–142. doi:10.1109/BTAS.2017.8272691. 2, 6, 7

[KPT*07] Kakadiaris I. A., Passalis G., Toderici G., Murtuza M. N., Lu Y., Karampatziakis N., Theoharis T.: Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence 29*, 4 (April 2007), 640–649. 2

[LCTL16a] Lee Y., Chen J., Tseng C. W., Lai S.-H.: Accurate and robust face recognition from RGB-D images with a deep learning approach. In *British Machine Vision Conf. (BMVC)* (Sept. 2016), pp. 123.1–123.14. doi:10.5244/C.30.123. 3

[LCTL16b] Lee Y., Chen J., Tseng C. W., Lai S.-H.: Accurate and robust face recognition from rgb-d images with a deep learning approach. In *British Machine Vision Conf. (BMVC)* (Sept. 2016), pp. 123.1–123.14. 7

[LHM*15] Li H., Huang D., Morvan J.-M., Wang Y., Chen L.: Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *Int. Journal of Computer Vision 113*, 2 (2015), 128–142. 6

[MBO07] Mian A., Bennamoun M., Owens R.: An efficient multimodal 2D-3D hybrid approach to automatic face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence 29*, 11 (Nov 2007), 1927–1943. 2

[MCMD12] Min R., Choi J., Medioni G., Dugelay J.: Real-time 3D face identification from a depth camera. In *Int. Conf. on Pattern Recognition (ICPR)* (Nov 2012), pp. 1739–1742. 2

[MdJG14] Mantecón T., del-Blanco C. R., Jaureguizar F., García N.: Depth-based face recognition using local quantized patterns adapted for range data. In *IEEE Int. Conf. on Image Processing (ICIP)* (Oct 2014), pp. 293–297. 2

[MdJG16] Mantecón T., del-Blanco C. R., Jaureguizar F., García N.: Visual face recognition using bag of dense derivative depth patterns. *IEEE Signal Processing Letters 23*, 6 (June 2016), 771–775. 2

[MKD14] Min R., Kose N., Dugelay J.-L.: Kinectfacedb: A kinect database for face recognition. *IEEE Trans. on Systems, Man, and Cybernetics: Systems 44*, 11 (Nov 2014), 1534–1548. doi:10.1109/TSMC.2014.2331215. 4

[OPH96] Ojala T., Pietikäinen M., Harwood D.: A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition 29*, 1 (jan 1996), 51–59. 3

[PFS*05] Phillips P. J., Flynn P. J., Scruggs T., Bowyer K. W., Chang J., Hoffman K., Marques J., Min J., Worek W.: Overview of the face recognition grand challenge. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2005), vol. 1, pp. 947–954. 4

[PVZ15] Parkhi O. M., Vedaldi A., Zisserman A.: Deep face recognition. In *British Machine Vision Conf. (BMVC)* (2015), vol. 1, p. 6. 1

[SAD*08] Savran A., Alyüz N., Dibeklioğlu H., Çeliktutan O., Gökberk B., Sankur B., Akarun L.: Bosphorus database for 3D face analysis. In *European Workshop on Biometrics and Identity Management* (2008), Springer, pp. 47–56. 4

[SKP15] Schroff F., Kalenichenko D., Philbin J.: FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2015), pp. 815–823. 1

[Spr11] Spreeuwers L.: Fast and accurate 3D face recognition. *Int. Journal of Computer Vision 93*, 3 (July 2011), 389–414. 2

[SWT14] Sun Y., Wang X., Tang X.: Deep learning face representation from predicting 10,000 classes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2014), pp. 1891–1898. 1

[TYRW14] Taigman Y., Yang M., Ranzato M., Wolf L.: DeepFace: Closing the gap to human-level performance in face verification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2014), pp. 1701–1708. 1

[YL18] Yin X., Liu X.: Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Trans. on Image Processing 27*, 2 (Feb 2018), 964–975. 1