Learning from Millions of 3D Scans for Large-scale 3D Face Recognition

Syed Zulqarnain Gilani Ajmal Mian
Computer Science and Software Engineering,
The University of Western Australia
{zulqarnain.gilani,ajmal.mian}@uwa.edu.au

Deep networks trained on millions of facial images are believed to be closely approaching human-level performance in face recognition. However, open world face recognition still remains a challenge. Although, 3D face recognition has an inherent edge over its 2D counterpart, it has not benefited from the recent developments in deep learning due to the unavailability of large training as well as large test datasets. Recognition accuracies have already saturated on existing 3D face datasets due to their small gallery sizes. Unlike 2D photographs, 3D facial scans cannot be sourced from the web causing a bottleneck in the development of deep 3D face recognition networks and datasets. In this backdrop, we propose a method for generating a large corpus of labeled 3D face identities and their multiple instances for training and a protocol for merging the most challenging existing 3D datasets for testing. We also propose the first deep CNN model designed specifically for 3D face recognition and trained on 3.1 Million 3D facial scans of 100K identities. Our test dataset comprises 1,853 identities with a single 3D scan in the gallery and another 31K scans as probes, which is several orders of magnitude larger than existing ones. Without fine tuning on this dataset, our network already outperforms state of the art face recognition by over 10%. We fine tune our network on the gallery set to perform end-to-end large scale 3D face recognition which further improves accuracy. Finally, we show the efficacy of our method for the open world face recognition problem.

1. Introduction

Face recognition, being a highly non-intrusive biometric [10], is fast becoming the tool of choice [38] in the domains of surveillance (*e.g.*, border control, suspect tracking, identification), security (*e.g.*, system login, banking, file encryption) and entertainment (*e.g.*, human computer interaction, 3D animation, virtual reality). Advancements in Deep Learning have brought about revolutionary improvements in various computer vision tasks where CNN based face recognition is claimed to have surpassed human performance [54]. However, the recent MegaFace challenges [27, 40] have shattered this myth, revealing that face recognition is still an unsolved problem.

Table 1. State-of-the-art 2D face recognition networks are trained on millions of images and tested on thousands of identities. However, 3D face recognition algorithms are tested on just a few hundred identities. The proposed FR3DNet is trained on 3.1M 3D scans and tested on 1.85K identities.

Modality	Model \ Technique	Input Size	Training		Testing			NW Param
			IDs	Scans	IDs	Scans	Dataset	
2D	VGG-Face [41]	224 × 224	2.6K	2.6M	5K	13K	LFW	134M
	DeepFace [54]	152 × 152	4K	4.4M	5K	13K	LFW	120M
	FaceNet [49]	220 × 220	8M	200M	5K	13K	LFW	140M
	MF2 [40]	-	672K	4.7M	690K	1M	MegaFace	-
3D	MMH [34]	-	-	-	0.46K	4K	FRGCv2	-
	K3DM [18]	-	-	-	0.46K	4K	FRGCv2	-
	Kim *et al.* [28]	224 × 224	0.7K	123K	0.1K	4.6K	Bosphorus	134M
3D	FR3DNet	160 × 160	100K	3.1M	1.85K	31K	LS3DFace	29M

Two-dimensional face recognition using CNNs on conventional photographs has shown remarkable performance on benchmarks like LFW [24] and Janus [29]. One of the main factors for this accomplishment is the ability of CNNs to learn from massive training data which is readily available. For instance, FaceNet [49] was trained on 200M textured images of 8M identities while VGG-Face [41] used 2.6M photos of 2,622 distinct subjects for training. Despite this phenomenal performance and availability of data, 2D face recognition is challenged by changes in illumination, pose and scale [1]. Furthermore, facial texture is not always stable for identities as it can change with make up. On the other hand, 3D face recognition has the potential to address these shortcomings. Although this modality in face recognition is gaining popularity [2, 4, 8, 17, 18, 32, 35], literature survey shows that there is no deep CNN designed specifically for 3D face recognition. This is primarily because of the lack of huge amounts of 3D training and test data. 3D face data cannot be obtained by crawling the web [27,40,41] and it requires great efforts to collect a respectable sized dataset. For instance, the largest publicly available 3D face dataset, ND-2006 [15] (a superset of FRGCv2 [45]) has

only 13,540 scans of 888 unique identities and took over two years to collect.

The problem of addressing the dearth of labeled 3D face data for training CNNs has been addressed through data augmentation. This is either done by creating synthetic faces from an existing 3D face model [13, 46] or by manipulating the facial appearance of existing data by introducing expressions [28, 33]. The former method is restricted to the linear space of the specific model resulting in faces with confined shape variations. The latter method only generates more scans per subject without increasing the number of unique identities in the data. In this paper, we present a technique for data augmentation that introduces non-linear heterogeneous variations in 3D shape, facial expressions, pose and occlusions to generate a training dataset of 3.1M 3D scans of 100K unique identities. The closest numbers in literature [28] for fine tuning VGG-Face on depth images are 127K scans of 700 identities, several orders of magnitude lower than ours (See Table 1 for details).

Another notable challenge to face recognition systems is the need for large-scale of test data. Recognition accuracies on small datasets like LFW (99.6% [49]) and FRGCv2 (98.7% [18]) have already saturated indicating the need for larger gallery sizes as it is well known that increasing the gallery size degrades the face recognition performance [16]. The MegaFace Challenges [27, 40] show that the performance of even the best 2D face recognition networks drop significantly when the gallery size increases. The identification accuracy of VGG network with triplet loss reduced by more than 20% on FaceScrub when only 10^2 distractors were added to the gallery set [40]. FaceNet [49] behaved similarly when one million distractors were added to the gallery [27]. Literature has no such statistics for 3D face recognition as large-scale 3D face recognition has never been attempted. Absence of large 3D face datasets with huge galleries is the prime reason for this massive gap in research. While millions of 2D face datasets have been generated by crawling the Internet [21, 27, 40], 3D domain still depends on physical collection of data from real subjects.

We present a unique solution by merging the most challenging publicly available 3D face datasets for large-scale face recognition testing. Our gallery consists of 1,853 identities while the probe set contains 31,860 3D scans of these individuals. Through extensive experiments, we show how existing methods and CNN models perform on this large scale dataset. We use the challenging protocol of a sin- gle sample per identity in the gallery as, most often than not, this would be the case in practical real world scenarios. Note that in the domain of 3D face recognition, the largest dataset (FRGCv2 [45]) on which results have mostly been reported has only 466 identities in the gallery.

Apart from data, the recognition algorithm itself is a very important component. The literature contains a variety of state-of-the-art deep CNN architectures for 2D face recognition [23, 41, 49, 52]. Using networks trained on 2D images to perform 3D face recognition is simplistic and suboptimal as 3D data has its own peculiarities defined by the underlying shape and geometry. To the best of our knowledge, there is no deep network designed specifically for 3D face recognition. We cover this research gap and propose a Deep 3D Face recognition Network coined *FR3DNet* (pronounced frednet) suited for 3D face data and trained from scratch on 3.1M 3D faces. We also analyze the affects of input image sizes and suitability of kernel sizes for 3D faces.

In a nutshell, our contributions are as follows: (1) *Training Data*: We present a method for generating a large corpus of labeled 3D face data for training CNNs. Our dataset contains 3.1M 3D scans of 100K identities highly rich in shape variations. Our training data does not include the public datasets. (2) *Large-scale Test Data*: Owing to the limitations of physically collecting huge 3D datasets, we merge the most challenging existing public 3D face datasets and propose a protocol for large-scale face recognition using a single sample per identity in the gallery. The test data contains 31,860 3D scans of 1,853 identities. To the best of our knowledge, this is the largest gallery size of 3D faces on which face recognition results have ever been reported. (3) *Deep 3D Face Recognition Network (FR3DNet)*: We propose the first ever deep CNN designed specifically for 3D face recognition and trained on 3.1M 3D faces. We fine tune *FR3DNet* on the 1,853 gallery identities in our large-scale dataset and achieve an end-to-end Rank-1 recognition rate of 98.74% on 27K probes, significantly outperforming the state-of-the-art on constituent datasets. The trained and end-to-end fine tuned *FR3DNet* will be made public.

2. Related Work

Face recognition is one of the most researched topics in Computer Vision and many detailed surveys exist [10, 43, 51, 60]. Here, we present the most relevant works to this paper and divide them into conventional methods which use hand crafted local and global features, deep learning based methods which are mainly based on various CNN architectures and data augmentation methods which focus on the problem of limited training data for learning.

Conventional Methods for 3D Face Recognition: These methods can be grouped into local or global descriptor based techniques [1, 10] where the latter also include 3D morphable model based methods. Local descriptor based techniques match local 3D point signatures derived from the curvatures, shape index and/or normals. For instance, Mian *et al*. [35] proposed a highly repeatable keypoint detection algorithm for 3D facial scans. They fused the 3D keypoints with 2D Scale Invariant Feature Transform (SIFT) to develop multimodal face recognition. However, the keypoint detection method and features were both sensitive to facial

expressions. For robustness to facial expressions, Mian *et al.* [34] proposed a parts based multimodal hybrid method (MMH) which exploited local and global features in the 2D and 3D modalities. A key component of their method was a variant of the ICP [5] algorithm which is computationally expensive due to its iterative nature. Gupta *et al.* [22] matched the 3D Euclidean and geodesic distances between pairs of fiducial landmarks to perform 3D face recognition. Berretti *et al.* [4] represented a 3D face with multiple mesh-DOG keypoints and local geometric histogram descriptors while Drira *et al.* [14] represented the facial surface by radial curves emanating from the nosetip.

Model based methods construct a 3D morphable face model and fit it to each probe face. Face recognition is performed by matching the model parameters to those in the gallery. Gilani *et al.* [18] proposed a keypoint based dense correspondence model and performed 3D face recognition by matching the parameters of a statistical morphable model called K3DM. Blanz *et al.* [6, 8] used the parameters of their 3DMM [7] for face recognition. Passalis *et al.* [42] proposed an Annotated Face Model (AFM) based on an average facial 3D mesh. Later, Kakadiaris *et al.* [25] proposed elastic registration using this AFM and performed 3D face recognition by comparing the wavelet coefficients of the deformed images obtained from morphing. Model fitting algorithms can be computationally expensive and do not perform well on large galleries as shown in our results. Both local and global techniques were tested on individual 3D datasets, the largest one being FRGCv2 with a gallery size of 466 identities. To the best of our knowledge, none of the conventional methods have performed large-scale 3D face recognition.

Deep Learning: Akin to progress in other applications of computer vision, deep learning has given a quantum jump in 2D face recognition. Three years ago, Facebook AI group proposed a nine-layer DeepFace model [54] mainly consisting of two convolutional, three locally-connected and two fully-connected (FC) layers. The network was trained on 4.4M 2D facial images of 4,030 identities and achieved an accuracy of 97.35% on the benchmark LFW [24] dataset which is 27% higher than the previous state of the art. This was followed by Google Inc., a year later, with FaceNet [49] based on eleven convolutional and three FC layers. The distinction of this network was its training dataset of 200M face images of 8M identities and a triplet loss function. The authors reported face recognition accuracy of 98.87% on LFW. DeepFace and FaceNet were both trained on pri- vate datasets which are not available to the broader research community. Consequently, Parkhi *et al.* [41] proposed a method for crawling the web to collect a face database of 2.6M 2D images from 2,622 identities and presented the VGG-Face model comprising of 16 convolutional and three FC layers. Despite training on a smaller dataset, the authors reported face recognition accuracy of 98.95% on the LFW dataset. However, recently the MegaFace Challenges [27, 40] claimed that the existing 2D benchmark datasets have reached saturation and proposed adding millions of faces to the galleries of these datasets to match the real world scenarios. They showed that the face recognition accuracy of state-of-the-art 2D networks dropped by more that 20% when just a few thousand distractors were added to the gallery of public face recognition benchmark datasets. The take away for the 3D domain is that CNNs on 2D data perform best when they learn from massive training sets and are particularly designed for the 2D modality, and yet, their real performance can be validated only when they are tested with large gallery sizes.

To the best of our knowledge, only Kim *et al.* [28] have presented deep 3D face recognition results. They reported results on three public datasets after fine tuning the VGG-Face network [41] on 3D depth images. They used an augmented dataset of 123,325 depth images to fine-tune the VGG-Face network and then tested it on the Bospho- rus [47], BU3DFE [59] and 3D-TEC (twins) [56] datasets individually. Except for the Bosphorus dataset, their results do not outperform the state-of-the-art conventional methods. Moreover, they have not reported results on the challenging FRGCv2 dataset and their fine-tuned model is not publicly available.

Data Augmentation: Dou *et al.* [13] and Richardson *et al.* [46] generated thousands of synthetic 3D images for face reconstruction using BFM [44], AFM [25] and 3DMM [7]. This method generates 3D faces within the linear space of a specific statistical face model. The faces generally have a variation of ± 3 standard deviations from the model mean with highly smooth surfaces. Gilani *et al.* [17] generated synthetic images using a similar approach. However, these images were used to train a 3D landmark identification network. Kim *et al.* [28] fitted the BFM [44] to 577 identities of FRGCv2 [45] database and induced 25 expressions in each identity. They also introduced minor pose variations between $\pm 10^{\circ}$ in yaw, pitch and roll for each original scan. To simulate occlusions, the authors introduced eight random occlusion patches to each 2D depth map to increase the dataset to 123,325 scans. This method only increases the intra-person variations without augmenting the number of identities, which in this case remained 577.

3. Proposed Data Generation for Training

We use 3D facial scans of 1,785 individuals (a propriety dataset) to train our deep network. The number of identities in this dataset is larger than any 3D dataset but still not sufficient for deep learning. Inspired by the recent works of Gilani *et al.* [18], we establish dense correspondence over 15K 3D vertices on the faces from this dataset, using the keypoints based algorithm. The goal now is to grow the

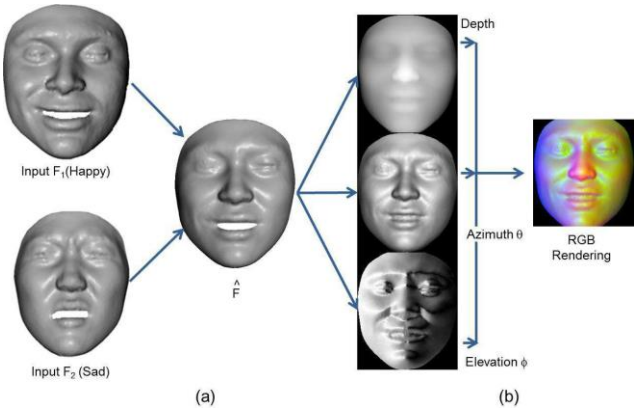

Figure 1. (a) Our data generation process. Notice the non-linearity introduced in the new face while at the same time preserving the high frequency shape variations. (b) Data preparation for input to our *FR3DNet*.

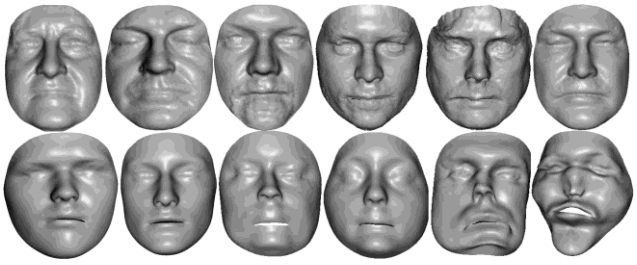

Figure 2. Example 3D faces generated by our method (row 1) and a statistical model [44] (row 2). The same identities were used for generating faces for both techniques. The 3D faces from our method look more realistic and have richer shape variations, especially around high curvature regions.

dataset by generating faces from the space spanned by pairs of densely corresponding real 3D faces of distinct identities. To ensure that the identities in the pair are as "distinct" as possible, we select the face pair with the maximum non-rigid shape difference. Let the faces be represented by $\mathrm{F}i = [x_p, y_p, z_p]^T$, where $i = 1, \ldots, N$, $p = 1, \ldots, P$; $N = 1,000$ and P =15,000. The shape difference between faces $\mathrm{F}i$ and $\mathrm{F}j$ is defined as

$$\mathrm{D}(i, j) = \frac{\gamma_{ij} + \gamma_{ji}}{2}, \quad (1)$$

where, γ_{ij} is the amount of bending energy required to deform 3D face $\mathrm{F}i$ to face $\mathrm{F}j$. Extending the 2D thin-plate spline model [9] to our case, we calculate the bending energy as, $\gamma(i, j) = \mathrm{x}^T\mathrm{B}\mathrm{x} + \mathrm{y}^T\mathrm{B}\mathrm{y} + \mathrm{z}^T\mathrm{B}\mathrm{z}$ where x, y and z are the vectors containing the x, y and z coordinates of P points in face $\mathrm{F}j$ and B is the bending matrix, which is defined as the $P \times P$ upper left matrix of $\begin{matrix} \mathrm{K} & \mathrm{S} \\ \mathrm{S}^T & 0 \end{matrix}^{-1}$. Here,

$$\mathrm{K}(a, b) = ||\mathrm{F}^a_i - \mathrm{F}^b_i||^2 \, log||\mathrm{F}^a_i - \mathrm{F}^b_i||,$$

with $a, b = 1, \ldots, P$, $\mathrm{S} = [1, x^j, y^j, z^j]$, and 0 is a $P \times 4$ matrix of zeros.

We select 90,100 pairs of 3D faces with maximum shape difference $\mathrm{D}(i, j)$ from the possible $\binom{N}{2} = 499,500$ pairs. Since the 3D faces in each pair are in dense correspondence to each other, a new face $\hat{F}$ is generated from the linear space of each pair (i, j) as $\hat{F} = \frac{[x^p_i, y^p_i, z^p_i]^T + [x^p_j, y^p_j, z^p_j]^T}{2}$. The process is depicted in Figure 1.

It is important to note here that our proposed method is significantly different from generating synthetic faces from a statistical face model. Varying the parameters of a statistical model generates faces that are over smooth and devoid of details and high frequency shape variations because of the low dimensional space that is used to generate them. On the contrary, our synthetic faces are generated from high dimensional raw 3D faces. Furthermore, not all faces generated by statistical models are *faces* unless strict constraints are imposed on the variation of the model parameters [37]. Such constraints will further limit the variations in identities that can be generated from the model. Finally, faces generated from statistical models span the linear space of the model whereas our method introduces non-linearity in the generated identities by varying the expressions of the face pair used to generate $\hat{F}$. By interpolating between identities and expressions, we generate new identities that do not necessarily lie in the linear space of the original identities. This is illustrated in Figure 1. Thus, we can choose the most dissimilar faces generating new identities that have maximum inter-person variations. The differences in the two methods of face generation can be seen clearly in Figure 2. Note that it is guaranteed that our method will never create deformed unrealistic faces like the ones generated by the statistical model (*e.g.* last two faces of bottom row).

The second source of 3D faces for our training data is a commercial software [1] that generates densely corresponded faces of varying facial shapes, ethnicities and expressions. We generate 300 identities, each in four different expressions with three intensity levels and follow the protocol above to create 9,950 new identities from the 44,850 possible pairs. However, in this case we select the pairs of faces that are "similar" and have smaller inter-person distance as per definition in Equation 1. The motivation for placing this condition comes from real world scenarios where face recognition systems are required to recognize people who look quite identical, for example in extreme cases, identical twins or triplets. A face recognition system trained on identities that look similar would have the power to distinguish between probes that are very similar in shape. Note that there is still ample inter-person variation in the original pairs for our *FR3DNet* to learn high level face identity features.

Finally, we simulate pose variations and large occlusions in each 3D scan by deploying 15 synthetic cameras on a

[1]Singular Inversions, Facegen Modeller,