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

Abstract

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 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.

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 inserting 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 single 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 sub-optimal 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. Berrettie 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 nosecirt.

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 private 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 than 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 Bosphorus [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 ±10° 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
We select 90,100 pairs of 3D faces with maximum shape difference \( \mathbf{D}(i,j) \) 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:

\[
\hat{F} = \frac{[x_i^p, y_i^p, z_i^p]^T + [x_j^p, y_j^p, z_j^p]^T}{2},
\]

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 un-realistic 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, www.facegen.com

---

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 \( \mathbf{F}_i = [x_i^p, y_i^p, z_i^p]^T \), where \( i = 1, \ldots, N, \ p = 1, \ldots, P; \ N = 1,000 \) and \( P = 15,000 \). The shape difference between faces \( \mathbf{F}_i \) and \( \mathbf{F}_j \) is defined as

\[
\mathbf{D}(i,j) = \frac{\gamma_{ij} + \gamma_{ji}}{2},
\]

where, \( \gamma_{ij} \) is the amount of bending energy required to deform 3D face \( \mathbf{F}_i \) to face \( \mathbf{F}_j \). Extending the 2D thin-plate spline model \([9]\) to our case, we calculate the bending energy as,

\[
\gamma(i,j) = \mathbf{x}^T \mathbf{B} \mathbf{x} + \mathbf{y}^T \mathbf{B} \mathbf{y} + \mathbf{z}^T \mathbf{B} \mathbf{z},
\]

where \( \mathbf{x}, \mathbf{y} \) and \( \mathbf{z} \) are the vectors containing the \( x \), \( y \) and \( z \) coordinates of \( P \) points in face \( \mathbf{F}_j \) and \( \mathbf{B} \) is the bending matrix, which is defined as the \( P \times P \) upper left matrix of \( \begin{bmatrix} \mathbf{K} & \mathbf{S} \\ \mathbf{S}^T & 0 \end{bmatrix}^{-1} \). Here,

\[
\mathbf{K}(a,b) = ||\mathbf{F}_a^i - \mathbf{F}_b^i||^2 \log||\mathbf{F}_a^i - \mathbf{F}_b^i||,
\]

with \( a, b = 1, \ldots, P; \ S = [1, \mathbf{x}^i, \mathbf{y}^i, \mathbf{z}^i] \), and \( \mathbf{0} \) is a \( P \times 4 \) matrix of zeros.
Table 2. Details of the dataset generated for training FR3DNet.

<table>
<thead>
<tr>
<th>Type</th>
<th>IDs</th>
<th>Expressions</th>
<th>Poses</th>
<th>Total Scans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Correspondence Model</td>
<td>90,100</td>
<td>2</td>
<td>15</td>
<td>*1,680,900</td>
</tr>
<tr>
<td>Real 3D Faces</td>
<td>1,785</td>
<td>1</td>
<td>15</td>
<td>26,775</td>
</tr>
<tr>
<td>Synthetic</td>
<td>8,120</td>
<td>12</td>
<td>15</td>
<td>1,461,600</td>
</tr>
<tr>
<td>Total</td>
<td>100,005</td>
<td>12</td>
<td>15</td>
<td>3,169,275</td>
</tr>
</tbody>
</table>

*Randomly selected from 2,703,000 scans.

4. **FR3DNet**: Deep Network for 3D Face Recognition

4.1. Training Data

The 3D pointcloud of each scan in the training data is used to generate a three channel image. The first channel is the depth image which is generated by fitting a surface of the form \( z(x, y) \) to the 3D pointcloud using the \texttt{gridfit} algorithm [12]. The surface normals of the original pointcloud are calculated in spherical coordinates \( (\theta, \phi) \) where \( \theta \) and \( \phi \) are the azimuth and elevation angles to make the second and third channels of the 3D image representation we used to train our network. The three channels are normalized on the 0-255 range and can be rendered as an RGB image. This image is passed through a landmark identification network [17] to detect the nosetip. With the face centered at the nosetip, we crop a square of \( 224 \times 224 \) pixels. This process is depicted in Figure 1. The \( 224 \times 224 \) size is chosen for comparison with existing networks. These images are down-sampled to \( 160 \times 160 \) for use in our network.

4.2. Network Architecture and Feature Extraction

Inspired by the success of recent deep networks [41, 50] in 2D face recognition, we propose a deep convolutional neural network that is suited to 3D data. The VGG network was designed for 2D images which exhibit significant texture variations over small regions. In contrast, 3D facial surfaces are generally smooth and hence filters with larger kernel sizes would better suite this type of data. For example, Figure 4 shows that surface patches of \( 7 \times 7 \) contain more variation than patches of \( 3 \times 3 \) and this is true even for the high curvature areas. This claim is empirically verified by calculating the average variance and average number of keypoints [36] over kernel sizes of 3, 7 and 9 in 10,000 3D images randomly selected from our training data. Average keypoints are calculated as the number of points on the 3D facial image that qualify as keypoints for a given kernel size, using the criterion in [36], divided by the number of possible kernels of that size on the image. The average kernel variance and average number of keypoints per kernel size of \( 7 \times 7 \) is significantly higher than size \( 3 \times 3 \). Results depicted in Figure 4 are compelling in favor of a kernel size of 7 for our initial convolutional layers.

The skeleton architecture of our FR3DNet follows [41] but with a change in the \texttt{conv} layers (see Figure 5 for details). We aim to minimize the average prediction log-loss after the softmax layer by learning the parameters of a network designed to classify \( N = 100,005 \) identities. After the network is trained, we remove the drop out layers. The embedded feature vectors of length 1,024 from FC7 can be used for face recognition by minimizing the cosine distance between a probe and the gallery faces in the feature space. We also fine-tune the FR3DNet on the gallery scans of the large-scale test data and denote it as FR3DNet\(_{FT}\).
The network was trained for 50 epochs. The input to our network is the 160 × 160 × 3 image where the three channels correspond to depth, azimuth and elevation angles of the normal vector. We train the proposed FR3DNet in MatConvNet [55] with randomly selected 90% scans of each identity in training and use the remaining scans for validation. We optimize the learning by Stochastic Gradient Descent (SGD) with standard L2 norm over the learned weights using mini batches of 20 images. The model is regularized using dropout layers after FC6 and FC7 with 0.5 rate. The learning rate was initially set to 0.01 and reduced by a factor of 10 after every 10 epochs. The filter weights of each layer were initialized with parameters drawn from a Gaussian distribution with zero mean and a standard deviation adjusted using the Xavier’s method [19]. The network was trained for 50 epochs. For fine-tuning FR3DNetFT, the network weights were frozen except for the last layer which was learnt with a rate of 0.01. Since the gallery contains only one 3D scan per identity, we render it from multiple viewpoints to generate more training data.

5. Large-scale 3D Face Test Dataset

FRGCv2 [45] by far still remains the largest 3D face recognition benchmark dataset with 446 identities. Table 1 shows that there is a huge disparity between 2D and 3D face dataset sizes. In the absence of any alternate means to collect real 3D faces for testing face recognition systems, we propose a protocol for merging the most challenging public datasets and call this dataset “LS3DFace”. The proposed technique is akin to crawling the web for augmenting 2D datasets and enables us to create a 3D face dataset of 1,853 identities with 31,860 scans. Our dataset enshrines every possible challenging scenario in face recognition and contains extreme variations in expressions, pose, occlusion, missing data, sensor type and similarities of faces in the form of identical twins. Table 4 lists the details of LS3DFace. For ease of comparison with our method, we provide the gallery and probe lists for various experiments. Since all these datasets are publicly available, other researchers can reproduce our results using our FR3DNet model. Moreover, newly released 3D face datasets can be added to make the protocol more challenging.

Since, ND-2006 [15] is a superset of FRGCv2 [45] dataset, we include the scans common in both datasets only once in the LS3DFace to avoid repetitions. Furthermore, BU-4DFE [58] dataset contains 3D video sequences of six expressions per identity. We only retain five frames equally spaced apart for each expression type.

6. Evaluation Protocol

We first evaluate the affects of input image size and the convolutional kernel size on face recognition accuracy. We train our network for 50 epochs on 500K 3D faces and report validation accuracy on 100K faces using three image sizes (96, 160 and 224) and kernel sizes of 3, 5, 7 and 9. The results are shown in Table 3. The improvement in validation accuracy from image size 96 × 96 to 160 × 160 and from kernel size 5 to 7 is significant and hence we select these parameter settings for the remaining experiments. Table 3 also validates our claim that bigger kernel sizes are more suitable for 3D data.

We feed forward the 3D images (containing depth, azimuth and elevation normal angles) of LS3DFace through FR3DNet and use the image representations from FC7 as features. The first available neutral scan of each identity is placed in the gallery while the remaining scans are used as probes. Where a neutral scan is not available, we use the first available scan in the gallery. Face identification is performed by matching the features of a probe with all identities in the gallery and based on minimum cosine distance, an identity is assigned to the probe. We report the results in the form of Cumulative Matching Curves (CMC). In case of face verification, the probe is matched with each claimed identity in the gallery. The result is a binary accept or reject decision based on some threshold applied to the match score. We report the results as ROC curves for varying thresholds of False Acceptance Rate (FAR).

Figure 5. Architecture of our proposed FR3DNet. Every convolutional layer is followed by a rectifier layer.

Table 3. Affect of image size (at K = 3) and kernel size K (for 160 × 160 image size) on validations accuracy. The kernel sizes of only the first two Conv layers are changed.

<table>
<thead>
<tr>
<th>Image size</th>
<th>96 × 96</th>
<th>160 × 160</th>
<th>224 × 224</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy%</td>
<td>82.27</td>
<td>86.33*</td>
<td>86.85</td>
</tr>
<tr>
<td>Kernel size</td>
<td>K = 3</td>
<td>K = 5</td>
<td>K = 7</td>
</tr>
<tr>
<td>Accuracy%</td>
<td>86.33</td>
<td>86.60</td>
<td>88.73*</td>
</tr>
</tbody>
</table>

* Significant improvement over smaller kernels (p << 0.001)

Table 4. Details of the constituent datasets of LS3DFace.

<table>
<thead>
<tr>
<th>Name</th>
<th>IDs</th>
<th>Scans</th>
<th>Expressions</th>
<th>Pose</th>
<th>Occlusion</th>
<th>Scanner</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRGCv2 [45]</td>
<td>466</td>
<td>4,007</td>
<td>Multiple</td>
<td>±15°</td>
<td>None</td>
<td>Laser</td>
</tr>
<tr>
<td>BU3DFE [59]</td>
<td>100</td>
<td>2,500</td>
<td>6 × 4</td>
<td>Frontal</td>
<td>None</td>
<td>Stereo</td>
</tr>
<tr>
<td>Bosphorus [57]</td>
<td>105</td>
<td>4,666</td>
<td>7</td>
<td>±90°</td>
<td>None</td>
<td>Stereo</td>
</tr>
<tr>
<td>GavabDB [39]</td>
<td>61</td>
<td>488</td>
<td>Multiple</td>
<td>±30°</td>
<td>None</td>
<td>Laser</td>
</tr>
<tr>
<td>Texas FRD [23]</td>
<td>118</td>
<td>1,151</td>
<td>Multiple</td>
<td>Frontal</td>
<td>None</td>
<td>Stereo</td>
</tr>
<tr>
<td>BU4DFE [58]</td>
<td>101</td>
<td>3,030</td>
<td>6 × 5</td>
<td>Frontal</td>
<td>None</td>
<td>Stereo</td>
</tr>
<tr>
<td>CASIA [57]</td>
<td>123</td>
<td>4,674</td>
<td>6</td>
<td>±90°</td>
<td>None</td>
<td>Laser</td>
</tr>
<tr>
<td>3D-TEC [56]</td>
<td>214</td>
<td>428</td>
<td>2</td>
<td>Frontal</td>
<td>None</td>
<td>Laser</td>
</tr>
<tr>
<td>ND-2006 [15]</td>
<td>422</td>
<td>9,443</td>
<td>Multiple</td>
<td>±15°</td>
<td>None</td>
<td>Laser</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1853</td>
<td>31,860</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

All datasets excepts GavabDB come with texture maps (RGB face images).

The file names of the scans used in gallery will be released.
Figure 6. Comparison of closed world (a) CMC and (b) ROC curves with state-of-the-art algorithms on LS3DFace dataset. (c) Comparison of open world Rank-1 identification rates at varying thresholds of unknown face acceptance rate (UFAR). The gallery of LS3DFace has been reduced by 50%. Note that FR3DNet performs better than VGG-Face by a margin of 10%. Curves for only the best performing networks from Table 5 are shown.

Our FR3DNetFT is fine-tuned on the gallery set mentioned above which contains a single sample per subject. This is a highly challenging scenario but the most practical one in real world. We learn an $N$-way ($N = 1853$) classifier and output the classification decision from the final soft-max layer. We compare our results with the state-of-the-art algorithms on each constituent dataset.

Closed World Face Recognition: This is a scenario where all the probes are enrolled in the gallery. Such probes are referred to as previously known. We report face identification and verification results on LS3DFace and its constituent datasets. We also compare our closed world results with four state-of-the-art 2D face recognition CNNs (RGB and 3D) as well as four state-of-the-art conventional methods using the same protocol. Note that wherever we report results on the constituent datasets of LS3DFace, the gallery always contains all 1,853 identities and not just the identities of that particular dataset.

Open World Face Identification: A real world scenario in face recognition occurs when the probe set contains unknown identities which are not enrolled in the gallery. Open world or open set 3D face recognition has not been studied in the context of a single sample per gallery. Scheirer et al. [48] and more recently Günther et al. [20] discuss this problem in the context of a gallery which contains multiple 2D images of a person and where training a classifier on the gallery faces is involved. Following [48] we define openness ($\Psi$) with a slight change to account for the single sample per person case used in our experiments:

$$\Psi = 1 - \sqrt{\frac{2 \times N_{\text{TargetID}}}{N_{\text{TestID}} + N_{\text{TargetID}}}}$$  \hspace{1cm} (2)$$

where $N_{\text{TargetID}}$ and $N_{\text{TestID}}$ denote the number of identities in the gallery and probe sets respectively. $\Psi = 0$ ($N_{\text{TargetID}} = N_{\text{TestID}}$) denotes the conventional closed world face recognition where as $N_{\text{TargetID}} = 1, N_{\text{TestID}} = 0$ gives varying levels of openness for open world face recognition. The robustness of a face recognition system in open world is tested by varying the Unknown Person Acceptance Rate (UPAR) denoted by $\tau$. When $\tau = 0$ all probes are classified as an unknown identity whereas when $\tau = 1$ every probe is assigned a known identity. At each $\tau$ the system outputs the Rank-1 identification rate of all probes (both known and unknown) identified correctly. We
Table 6. Comparison of Rank-1 recognition accuracy (%) with the state-of-the-art methods on constituent datasets of LS3DFace. Note that FR3DNetFT uses the full gallery of LS3DFace dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Xie et al. [57]</td>
<td>95.0</td>
<td>92.3</td>
<td>96.3</td>
<td>98.0</td>
<td>98.5</td>
<td>99.5</td>
</tr>
<tr>
<td>Kim et al. [28]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Falctinier et al. [15]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gupta et al. [22]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Al-Osaimi et al. [3]</td>
<td>97.8</td>
<td>97.5</td>
<td>97.0</td>
<td>97.1</td>
<td>97.2</td>
<td>97.0</td>
</tr>
<tr>
<td>Li et al. [31,32]</td>
<td>96.3</td>
<td>95.3</td>
<td>95.5</td>
<td>95.3</td>
<td>95.8</td>
<td>95.8</td>
</tr>
<tr>
<td>Lei et al. [30]</td>
<td>96.4</td>
<td>97.3</td>
<td>96.9</td>
<td>96.2</td>
<td>97.0</td>
<td>97.0</td>
</tr>
<tr>
<td>Mian et al. [34]</td>
<td>96.2</td>
<td>95.9</td>
<td>94.2</td>
<td>96.4</td>
<td>95.6</td>
<td>95.6</td>
</tr>
<tr>
<td>Gilani et al. [18]</td>
<td>98.5</td>
<td>97.2</td>
<td>96.0</td>
<td>98.6</td>
<td>98.1</td>
<td>97.2</td>
</tr>
<tr>
<td>FR3DNetFT</td>
<td>99.9</td>
<td>99.9</td>
<td>98.0</td>
<td>100.0</td>
<td>99.7</td>
<td>99.7</td>
</tr>
</tbody>
</table>

FR3DNetFT report open world face recognition on LS3DFace dataset at different openness values and compare the results with the state-of-the-art algorithms.

7. Results and Analysis

7.1. Closed World Face Recognition

Table 5 details the closed world Rank-1 identification results on LS3DFace and compares them with the state-of-the-art deep and conventional methods. We perform 2D face recognition on the RGB images that accompany the datasets. The main conclusions that can be drawn from these results is that 3D face has more to offer in terms of correctly identifying a person and 3D face recognition results are superior to its 2D counterpart. Note that the conventional methods that report near saturated results on small 3D face datasets (see Table 6), fail to achieve high accuracies on the large LS3DFace dataset. This shows that increasing the gallery size has a strong inverse affect on the performance of these algorithms. FR3DNet outperforms state-of-the-art conventional 3D face recognition algorithms by more than 14% and the best 2D face recognition method by 4.7%. In a single sample per face gallery scenario, only 3D faces have the advantage to generate more training data to finetune any network. FR3DNetFT outperforms VGG-Face by 8% and the margins are more significant (over 15%) when the probes are more challenging as in the cases of 3D-TEC (twins) and UMBDB datasets. We compare our CMCs and ROCs with VGG-Face and GoogleNet in Figure 6.

A straightforward comparison of FR3DNetFT on the individual datasets in Table 6 shows that our deep network fine-tuned on a single scan per person outperforms the state-of-the-art conventional algorithms. The results for all other methods are reported from their original papers. Note that these results are biased against FR3DNetFT which tests each probe against the full gallery of LS3DFace dataset whereas other methods test the probes against only the gallery identities of that particular dataset. Results are remarkably especially on the UMBDB [11] (containing occlusions) and CASIA datasets where our network outperforms the nearest competitors by 18.6% and 14.3% respectively.

7.2. Open World Face Identification

Unlike other methods [20, 48], we do not use classifiers to train on a subset of gallery faces since we perform face recognition with a single sample per person in the gallery. Figure 6(c) shows Rank-1 identification results when half of the gallery (925) identities are removed to simulate an open world scenario. Hence, the probes belonging to these identities are unknown. The curves demonstrate the detection power of FR3DNet which outperforms VGG-Face (RGB) by a significant margin. In Table 7 we report the average Rank-1 recognition rate over τ = [0, 1] for varying levels of openness (See Equation 2). For both experiments we performed ten random fold selection of unknowns and the figures presented in are the mean results of ten random folds. The standard deviations in all cases was less than 1%.

8. Conclusion

This paper bridges the vast gap between research advancements in 2D and 3D face recognition algorithms especially in the context of deep learning. It proposes a technique to generate millions of 3D facial images of unique identities by simultaneously interpolating between the facial identity and facial expression spaces. Additional factors such as subtle variations in facial shape, major variations in facial shape, camera viewpoint and self occlusions are introduced to generate a training dataset of 3.1M scans of 100K identities. A purpose designed 3D face recognition CNN is proposed and trained from scratch on this dataset. To test the network, existing 3D face datasets are merged and comparative results are reported on the largest 3D face dataset to date. The proposed training and test datasets are several orders of magnitude larger than the existing 3D datasets reported in the literature. The proposed FR3DNet outperforms the state-of-the-art 3D as well as 2D face recognition algorithms in closed and open world recognition scenarios.

Acknowledgments

This research was partly supported by ARC Discovery grant DP160101458. We are grateful to NVIDIA Corporation for donating the Titan Xp GPU used for this research.
References


