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

Syed Zulqarnain Gilani   Ajmal Mian

School of 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 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 [14], is fast becoming the tool of choice [39] in the domains of surveillance (for example, border control, suspect tracking, identification), security (for example, system login, banking, file encryption) and entertainment (for example, human computer interaction, 3D animation, virtual reality). Advancements in Deep Learning have brought

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.

Modal-ity	Model \ Technique	Input Size	Training		Testing			NW Param
			IDs	Scans	IDs	Scans	Dataset	
2D	VGG-Face [45]	224 × 224	2.6K	2.6M	5K	13K	LFW	134M
	DeepFace [58]	152 × 152	4K	4.4M	5K	13K	LFW	120M
	FaceNet [53]	220 × 220	8M	200M	5K	13K	LFW	140M
	MF2 [42]	-	672K	4.7M	690K	1M	MegaFace	-
3D	MMH [35]	-	-	-	0.46K	4K	FRGCv2	-
	K3DM [13]	-	-	-	0.46K	4K	FRGCv2	-
	Kim *et al.* [29]	224 × 224	0.7K	123K	0.1K	4.6K	Bosphorus	140M
3D	**FR3DNet**	160 × 160	100K	3.1M	1.85K	31K	LS3DFace	**29M**

about revolutionary improvements in various computer vision tasks where CNN based face recognition is claimed to have surpassed human performance [58]. However, the recent MegaFace challenges [28, 42] 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 [25] and Janus [30]. 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 [53] was trained on 200M textured images of 8M identities while VGG-Face [45] 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 [2]. 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 [3, 5, 9, 11, 13, 33, 36], 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 [28, 42, 45] and it requires great efforts to collect a respectable sized

dataset. For instance, the largest publicly available 3D face dataset, ND-2006 [19] (a superset of FRGCv2 [49]) 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 [17, 50] or by manipulating the facial appearance of existing data by introducing expressions [29, 34]. 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 [29] 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% [53]) and FRGCv2 (98.7% [13]) 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. The MegaFace Challenges [28, 42] 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 [42]. FaceNet [53] behaved similarly when one million distractors were added to the gallery [28]. 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 [22, 28, 42], 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 [49]) 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 [24, 45, 53, 56]. 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 [14, 47, 55, 66]. 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 [2, 14] 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*. [36] 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.* [35] 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 [7] algorithm which is computationally expensive due to its iterative nature. Gupta *et al.* [23] matched the 3D Euclidean and geodesic distances between pairs of fiducial landmarks to perform 3D face recognition. Berretti *et al.* [5] represented a 3D face with multiple mesh-DOG keypoints and local geometric histogram descriptors while Drira *et al.* [18] 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.* [13] 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.* [8, 11] used the parameters of their 3DMM [10] for face recognition. Passalis *et al.* [46] proposed an Annotated Face Model (AFM) based on an average facial 3D mesh. Later, Kakadiaris *et al.* [26] 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 [58] 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 [25] dataset which is 27% higher than the previous state of the art. This was followed by Google Inc., a year later, with FaceNet [53] 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.* [45] 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 [28, 42] 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.* [29] have presented deep 3D face recognition results. They reported results on three public datasets after fine tuning the VGG-Face network [45] 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 [51], BU3DFE [65] and 3D-TEC (twins) [61] 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.* [17] and Richardson *et al.* [50] generated thousands of synthetic 3D images for face reconstruction using BFM [48], AFM [26] and 3DMM [10]. 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.* [9] generated synthetic images using a similar approach. However, these images were used to train a 3D landmark identification network. Kim *et al.* [29] fitted the BFM [48] to 577 identities of FRGCv2 [49] database and induced 25 expressions in each identity. They also introduced minor pose variations between $\pm 10°$ 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) who were participants of various studies in our institution to train our deep network. The number of identities in this dataset is larger than any 3D dataset but still not
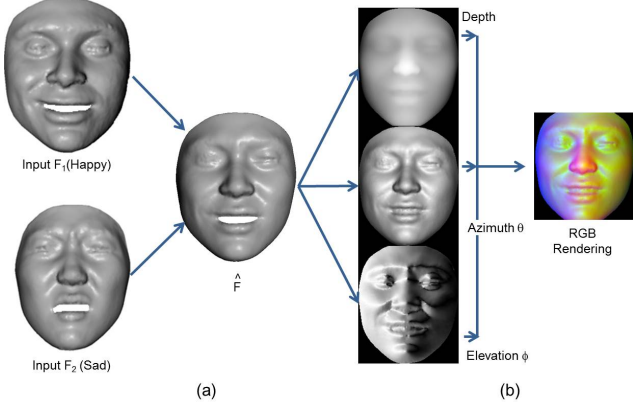
Figure 1. (a) Our data generation process. Notice the non-linearity introduced in the new face while at the same time preserving the high frequency shape variations. (b) Data preparation for input to our *FR3DNet*.

Figure 2. Example 3D faces generated by our method (row 1) and a statistical model [48] (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.

sufficient for deep learning. Inspired by the recent works of Gilani *et al.* [13], 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 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 $\mathbf{F}_i = [x_p, y_p, z_p]^T$, where $i = 1, \dots, N$, $p = 1, \dots, P$; $N = 1,000$ and $P = 15,000$. The shape difference between faces $\mathbf{F}_i$ and $\mathbf{F}_j$ is defined as

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

where, γ_{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 [12] to our case, we calculate the bending energy as, $\gamma(i,j) = \mathbf{x}^T \mathcal{B} \mathbf{x} + \mathbf{y}^T \mathcal{B} \mathbf{y} + \mathbf{z}^T \mathcal{B} \mathbf{z}$ 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 $\mathcal{B}$ is the bending matrix, which is defined as the $P \times P$ upper left matrix of $\begin{bmatrix} \mathbf{K} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{0} \end{bmatrix}^{-1}$. Here

$\mathbf{K}(a,b) = ||\mathbf{F}_i^a - \mathbf{F}_i^b||^2 \, log||\mathbf{F}_i^a - \mathbf{F}_i^b||$ with $a, b = 1, \dots, P$, $\mathbf{S} = [\mathbf{1}, \mathbf{x}^j, \mathbf{y}^j, \mathbf{z}^j]$, and $\mathbf{0}$ is a $P \times 4$ matrix of zeros.

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 [38]. 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 (for example 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.

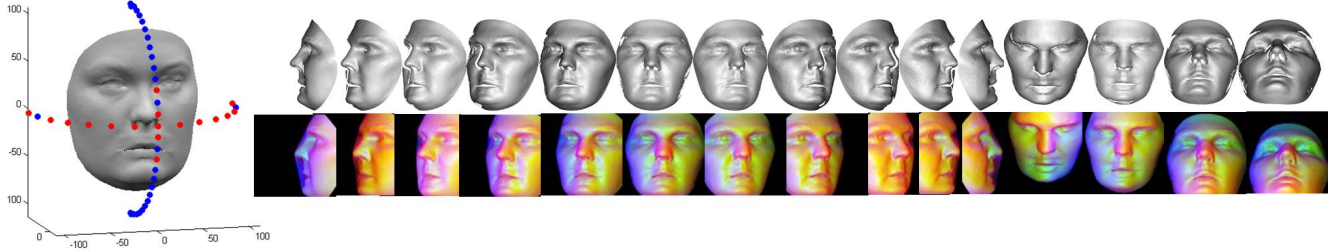[1] Singular Inversions, Facegen Modeller, www.facegen.com

Figure 3. Position of cameras on a hemisphere surrounding the 3D face and the 15 poses generated as a result.

Finally, we simulate pose variations and large occlusions in each 3D scan by deploying 15 synthetic cameras on a hemisphere in front of the 3D face. The cameras are deployed in the range of $[-90°, 90°]$ on the longitude and $[-30°, 30°]$ on the latitude of the hemisphere; all at $15°$ intervals. We do not deploy cameras at $-75°$ and $75°$ on the longitude. The self-occluded 3D points from the camera view point are removed by applying the hidden point removal algorithm [27]. Note that this step creates missing data in varying amounts on each scan, thereby simulating realistic self occlusion. Figure 3 depicts the placement of cameras and displays the output images.

Our final training dataset consists of 3,169,275 facial scans from 100,005 identities (approx 31 scans per identity). Table 2 gives details of the augmented 3D face dataset.

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 *gridfit* algorithm [16]. The surface normals of the original pointcloud are calculated in spherical coordinates (θ, ϕ) where θ, ϕ are the azimuth and elevation angles of the normal vector. Using a similar x, y grid to the depth image, surfaces of the form $\theta(x, y)$ and $\phi(x, y)$ are fitted to 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 [9] to detect the nosetip. With the face centered at the nosetip, we crop a square of 224×224 pixels. This process is depicted in Figure 1. The 224×224 size is chosen for comparison with existing networks. These images are down-sampled to 160×160 for use in our network.
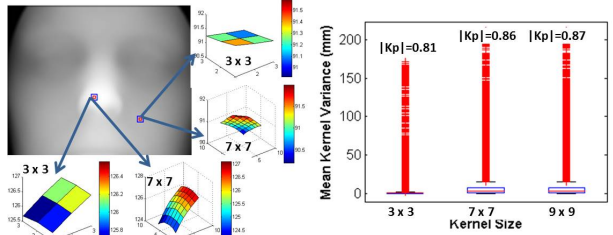

Figure 4. Variation in 3D depth frequency with different kernel sizes. Notice that patches of 3×3 are almost quasi-planar. $|K_p|$ denotes average number of keypoints per kernel size.

Table 2. Details of the dataset generated for training *FR3DNet*.

Type	IDs	Expressions	Poses	Total Scans
Dense Correspondence Model	90,100	2	15	*1,680,900
Real 3D Faces	1,785	1	15	26,775
Synthetic	8,120	12	15	1,461,600
Total	**100,005**	**12**	**15**	**3,169,275**

*Randomly selected from 2,703,000 scans.

4.2. Network Architecture and Feature Extraction

Inspired by the success of recent deep networks [45, 54] 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 7×7 contain more variation than patches of 3×3 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 [37] 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 [37], 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 7×7 is significantly higher than size 3×3. Results depicted in Figure 4 are compelling in favor of a kernel size of 7 for our initial convolutional layers.
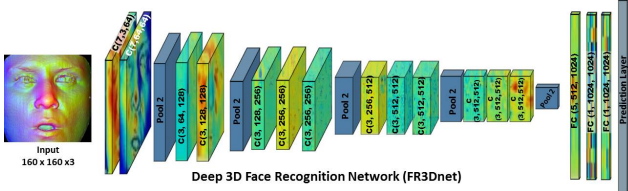
Figure 5. Architecture of our proposed *FR3DNet*. Every convolutional layer is followed by a rectifier layer.

The skeleton architecture of our *FR3DNet* follows [45] but with a change in the *conv* layers, details of which are given in Figure 5. 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 $FR3DNet_{FT}$.

4.3. Implementation Details

The input to our network is the $160 \times 160 \times 3$ image where the three channels correspond to depth, azimuth and elevation angles of the normal vector. We train the proposed *FR3DNet* in MatConvNet [60] 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 [20]. The network was trained for 50 epochs. For fine-tuning $FR3DNet_{FT}$, 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 [49] 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 aug-

menting 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 3 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 [19] is a superset of FRGCv2 [49] dataset, we include the scans common in both datasets only once in the *LS3DFace* to avoid repetitions. Furthermore, BU-4DFE [64] 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 4. 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 4 also validates our claim that bigger kernel sizes are more suitable for 3D data.

Table 3. Details of the constituent datasets of *LS3DFace*.

Name	IDs	Scans	Expressions	Pose	Occlusion	Scanner
FRGCv2 [49]	466	4,007	Multiple	$\pm 15°$	None	Laser
BU3DFE [65]	100	2,500	6×4	Frontal	None	Sterio
Bosphorus [51]	105	4,666	7	$\pm 90°$	4 types	Sterio
GavabDB [41]	61	488	Miltiple	$\pm 30°$	None	Laser
Texes FRD [23]	118	1,151	Miltiple	Frontal	None	Sterio
BU4DFE [64]	101	3,030	6×5	Frontal	None	Sterio
CASIA [63]	123	4674	6	$\pm 90°$	None	Laser
UMB DB [15]	143	1,473	4	Frontal	7 types	Laser
3D-TEC [61]	214	428	2	Frontal	None	Laser
ND-2006 [19]	422	9,443	Multiple	$\pm 15°$	None	Laser
TOTAL	**1853**	**31,860**	-	-	-	-

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

Table 4. 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.

Image size	96×96	160×160	224×224	
Accuracy(%)	82.27	**86.33***	86.85	
Kernel size	$K = 3$	$K = 5$	$K = 7$	$K = 9$
Accuracy(%)	86.33	86.60	**88.73***	88.92

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

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[2]. 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).

Our *FR3DNet$_{FT}$* 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 person gallery. Scheirer *et al*. [52] and more recently Günther *et al*. [21] 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 [52] we define openness (Ψ) with a slight change to account for the single sample per person case used in our experiments:

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

where N_{TargetID} and N_{TestID} denote the number of identities in the gallery and probe sets respectively. $\Psi = 0$

---

[2]The file names of the scans used in gallery will be released.

($N_{\text{TargetID}} = N_{\text{TestID}}$) denotes the conventional closed world face recognition where as $N_{\text{TargetID}} = [1, N_{\text{TestID}} - 1]$ 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 τ. When $\tau = 0$ all probes are classified as an unknown identity whereas when $\tau = 1$ every probe is assigned a *known* identity. At each τ the system outputs the Rank-1 identification rate of all probes (both *known* and *unknown*) identified correctly. We 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 than its 2D counterpart. Note that the conventional methods that report near saturated results on small 3D face datasets, 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 fine-tune any network. *FR3DNet$_{FT}$* 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 the our CMCs and ROCs with VGG-Face and GoogleNet in Figure 6.

A straightforward comparison of *FR3DNet$_{FT}$* 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 *FR3DNet$_{FT}$* 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 remarkable especially on the UMBDB [15] (containing occlusions) and CASIA datasets where our network outperforms the nearest competitors by 18.6% and 14.3% respectively.

Table 5. Comparison of face recognition accuracy (%) on *LS3DFace* with state-of-the-art deep and conventional methods. We have used the complete gallery of *LS3DFace* in all experiments reported here.

| Method	Model \ Technique	Modality	Gallery of LS3DFace										
			LS3DFace This paper	FRGC [49]	BU3DFE [65]	BU4DFE [64]	Bosphorus [51]	CASIA [63]	GavabDB [41]	TexasFRD [23]	3D-TEC [61]	UMBDB [15]	ND-2006 [19]
CNN	GoogleNet [57]	RGB	53.97	21.51	50.76	65.41	63.44	85.91	-	53.08	79.95	65.78	24.14
	Resnet152 [24]	RGB	15.05	13.53	8.04	9.64	7.05	52.85	-	20.94	72.66	34.08	10.92
	VGG-Face [45]	RGB	90.85	87.92	97.68	96.51	96.39	94.18	-	99.73	83.30	81.54	82.86
	GoogleNet [57]	3D	38.66	35.54	46.56	41.88	26.81	50.81	66.56	67.59	67.29	47.66	30.81
	Resnet152 [24]	3D	12.49	14.40	5.80	10.13	3.84	25.34	44.26	16.25	60.98	22.20	12.08
	VGG-Face [45]	3D	61.20	62.42	71.16	53.17	48.14	71.95	77.38	85.58	78.04	67.48	60.81
Conventional	MMH [35]	3D + 2D	83.08	89.37	88.50	84.93	85.10	85.24	86.64	85.67	80.85	77.32	86.71
	3D Keypoint [36]	3D	81.76	86.59	85.14	82.50	82.64	81.38	84.41	84.99	75.63	71.68	82.30
	R3DM [9]	3D	82.89	87.50	87.13	83.21	86.06	84.51	85.60	85.47	78.27	77.11	84.84
	K3DM [13]	3D	84.67	89.50	89.24	86.05	88.60	85.35	87.90	86.13	79.55	78.64	87.77
CNN	**FR3DNet**	3D	95.51	97.06	98.64	95.53	96.18	98.37	96.39	100.00	97.90	91.17	95.62
	FR3DNet$_{FT}$	**3D**	**98.75**	**99.88**	**99.96**	**98.04**	**100.00**	**99.74**	**99.70**	**100.00**	**99.12**	**97.20**	**99.13**

(a)

(b)

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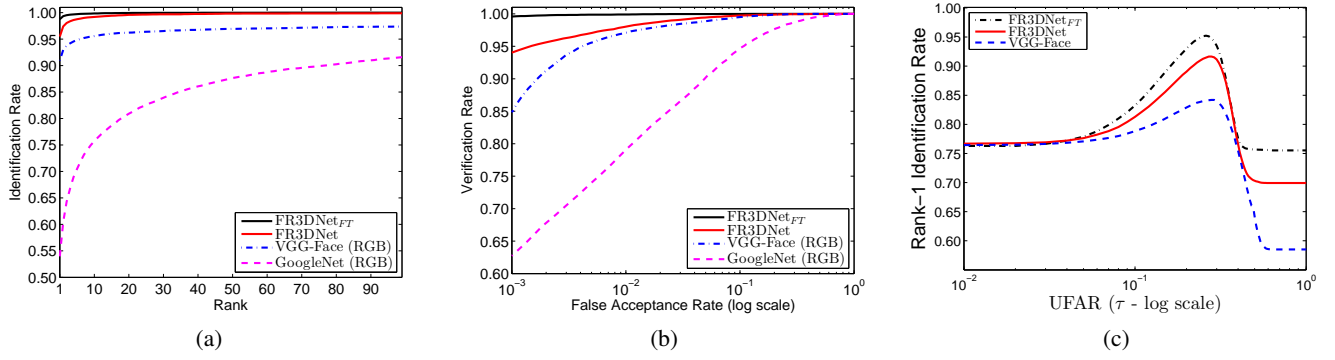
Figure 6. Comparison of closed world (a) CMC and (b) ROC curves with state-of-the-art algorithms on *LS3DFace* dataset. (c) Comparison of open world Rank-1 identification rates at varying thresholds of unknown face acceptance rate (UFAR). The gallery of *LS3DFace* has been reduced by 50%. Note that *FR3DNet*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

Table 6. Comparison of Rank-1 recognition accuracy (%) with the state-of-the-art methods on constituent datasets of *LS3DFace*. Note that FR3DNet$_{FT}$ uses the full gallery of *LS3DFace* dataset.

Method/Dataset	FRGCv2 [49]	BU3DFE [65]	BU4DFE [64]	Bosphorus [51]	CASIA [63]	GavabDB [41]	TexasFRD [23]	3D-TEC [61]	UMBDB [15]	ND-2006 [19]
Xu *et al.* [63]	-	-	-	-	83.9	-	-	-	-	-
Kim *et al.* [29]	-	95.0	-	99.2	-	-	-	-	-	-
Faltemier *et al.* [19]	-	-	-	-	-	-	-	-	-	82.8
Gupta *et al.* [23]	-	-	-	-	-	-	97.9	-	-	-
Al-Osaimi *et al.* [4]	97.8	-	-	-	-	-	-	97.2	-	-
Li *et al.* [32, 33]	96.3	92.2	-	96.6	-	-	-	96.7	-	-
Lei *et al.* [31]	96.3	94.0	-	-	-	96.3	-	-	73.1	-
Mian *et al.* [35]	96.2	95.9	94.2	96.4	82.5	95.4	98.0	95.9	69.3	95.3
Gilani *et al.* [13]	98.5	96.2	96.0	98.6	85.4	96.5	98.1	92.6	78.6	96.8
FR3DNet$_{FT}$	**99.9**	**99.9**	**98.0**	**100.0**	**99.7**	**99.7**	**100.0**	**99.1**	**97.2**	**99.1**

7.2. Open World Face Identification

Unlike other methods [21, 52], we do not use classifiers to train on a subset of gallery faces since we perform face

Table 7. Comparison of average open world Rank-1 recognition accuracy(%) with the state-of-the-art 2D face recognition networks at varying levels of openness (see Equation 2). The standard deviation of the accuracies over ten random folds is less than 1%.

Openness	4%	9%	15%	23%	33%	50%
num Unknown IDs	272	542	810	1072	1321	1592
GoogleNet [57]	43.70	38.54	34.64	28.41	23.70	19.28
VGG-Face [45]	86.61	80.56	71.31	63.52	54.54	48.44
FR3DNet	92.41	88.63	78.42	71.51	64.10	57.02
FR3DNet$_{FT}$	97.22	91.94	83.72	77.80	70.21	61.20

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 $\tau = [0, 1]$ 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

References

[1] S. Z. Gilani and A. Mian. Perceptual differences between men and women: A 3D facial morphometric perspective. In *IEEE ICPR*, 2014.

[2] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007. 1, 2

[3] F. Al-Osaimi, M. Bennamoun, and A. Mian. An expression deformation approach to non-rigid 3D face recognition. *IJCV*, 81(3):302–316, 2009. 1

[4] F. R. Al-Osaimi. A novel multi-purpose matching representation of local 3D surfaces: A rotationally invariant, efficient, and highly discriminative approach with an adjustable sensitivity. *IEEE Transactions on Image Processing*, 25(2):658–672, 2016. 8

[5] S. Berretti, N. Werghi, A. Del Bimbo, and P. Pala. Matching 3D face scans using interest points and local histogram descriptors. *Computers & Graphics*, 37(5):509–525, 2013. 1, 3

[6] S. Z. Gilani and A. Mian. Towards large-scale 3D face recognition. In *IEEE DICTA 2016*, pages 1–8, 2016.

[7] P. J. Besl, N. D. McKay, et al. A method for registration of 3-d shapes. *IEEE Transactions on pattern analysis and machine intelligence*, 14(2):239–256, 1992. 3

[8] V. Blanz, K. Scherbaum, and H.-P. Seidel. Fitting a morphable model to 3D scans of faces. In *IEEE ICCV*, 2007. 3

[9] S. Z. Gilani, A. Mian, and P. Eastwood. Deep, dense and accurate 3D face correspondence for generating population specific deformable models. *Pattern Recognition*, 69:238–250, 2017. 1, 3, 5, 8

[10] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM Conference on Computer Graphics and Interactive Techniques*, 1999. 3

[11] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE TPAMI*, 25(9):1063–1074, 2003. 1, 3

[12] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11(6):567–585, 1989. 4

[13] S. Z. Gilani, A. Mian, F. Shafait, and I. Reid. Dense 3D face correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 1, 2, 3, 4, 8

[14] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer vision and image understanding*, 101(1):1–15, 2006. 1, 2

[15] A. Colombo, C. Cusano, and R. Schettini. UMB-DB: A database of partially occluded 3D faces. In *International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2113–2119. IEEE, 2011. 6, 7, 8

[16] J. D́Erico. Surface fitting using gridfit. In *MATLAB Central File Exchange*, 2008. 5

[17] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. *arXiv preprint arXiv:1704.05020*, 2017. 2, 3

[18] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama. 3D face recognition under expressions, occlusions, and pose variations. *IEEE TPAMI*, 35(9):2270–2283, 2013. 3

[19] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. Using a multi-instance enrollment representation to improve 3D face recognition. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007.*, pages 1–6. IEEE, 2007. 2, 6, 8

[20] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010. 6

[21] M. Günther, S. Cruz, E. M. Rudd, and T. E. Boult. Toward open set face recognition. *arXiv preprint arXiv:1705.01567*, 2017. 7, 8

[22] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1M: Challenge of recognizing one million celebrities in the real world. *Electronic Imaging*, 2016(11):1–6, 2016. 2

[23] S. Gupta, M. K. Markey, and A. C. Bovik. Anthropometric 3D face recognition. *International journal of computer vision*, 90(3):331–349, 2010. 3, 6, 8

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 2, 8

[25] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1, 3

[26] I. A. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE TPAMI*, 29(4):640–649, 2007. 3

[27] S. Katz, A. Tal, and R. Basri. Direct visibility of point sets. In *ACM Transactions on Graphics (TOG)*, volume 26, page 24. ACM, 2007. 5

[28] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4873–4882, 2016. 1, 2, 3

[29] D. Kim, M. Hernandez, J. Choi, and G. Medioni. Deep 3D face identification. *arXiv preprint arXiv:1703.10714*, 2017. 1, 2, 3, 8

[30] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1931–1939, 2015. 1

[31] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, and X. Zhou. A two-phase weighted collaborative representation for 3D partial face recognition with single sample. *Pattern Recognition*, 52:218–237, 2016. 8

[32] H. Li, D. Huang, J.-M. Morvan, L. Chen, and Y. Wang. Expression-robust 3D face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns. *Neurocomputing*, 133:179–193, 2014. 8

[33] H. Li, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen. Towards 3D face recognition in the real: A registration-free approach using fine-grained matching of 3D keypoint descriptors. *International Journal of Computer Vision*, 113(2):128–142, 2014. 1, 8

[34] I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016. 2

[35] A. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. *IEEE TPAMI*, 29(11):1927–1943, 2007. 1, 3, 8

[36] A. Mian, M. Bennamoun, and R. Owens. Keypoint detection and local feature matching for textured 3D face recognition. *IJCV*, 79(1):1–12, 2008. 1, 2, 8

[37] A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. *IJCV*, 89(2-3):348–361, 2010. 5

[38] A. Mian, Y. Hu, R. Hartley, and R. Owens. Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. *IEEE transactions on image processing*, 22(12):5252–5262, 2013. 4

[39] A. Mian and N. Pears. 3D face recognition. In *3D Imaging, Analysis and Applications*, pages 311–366. Springer, 2012. 1

[40] S. Z. Gilani, K. Rooney, F. Shafait, M. Walters, and A. Mian. Geometric facial gender scoring: Objectivity of perception. *PloS one*, 9(6), 2014.

[41] A. B. Moreno and A. Sánchez. Gavabdb: a 3D face database. In *Proc. 2nd COST Workshop on Biometrics on the Internet*, pages 75–80, 2004. 6, 8

[42] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3

[43] S. Z. Gilani, F. Shafait, and A. Mian. Biologically significant facial landmarks: How significant are they for gender classification? In *IEEE DICTA*, 2013.

[44] S. Z. Gilani, F. Shafait, and A. Mian. Gradient based efficient feature selection. In *IEEE WACV*, 2014.

[45] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, page 6, 2015. 1, 2, 3, 5, 6, 8

[46] G. Passalis, I. Kakadiaris, T. Theoharis, G. Toderici, and N. Murtuza. Evaluation of 3D face recognition

in the presence of facial expressions: an annotated deformable model approach. In *IEEE CVPR Workshops*, 2005. 3

[47] H. Patil, A. Kothari, and K. Bhurchandi. 3-D face recognition: features, databases, algorithms and challenges. *Artificial Intelligence Review*, 44(3):393–441, 2015. 2

[48] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *International Conference On Advanced Video and Signal Based Surveillance*, pages 296–301. IEEE, 2009. 3, 4

[49] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, et al. Overview of the face recognition grand challenge. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2005. 2, 3, 6, 8

[50] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 460–469. IEEE, 2016. 2, 3

[51] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008. 3, 6, 8

[52] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. 7, 8

[53] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 815–823, 2015. 1, 2, 3

[54] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[55] S. Soltanpour, B. Boufama, and Q. J. Wu. A survey of local feature methods for 3d face recognition. *Pattern Recognition*, 72:391–406, 2017. 2

[56] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1891–1898, 2014. 2

[57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1–9, 2015. 8

[58] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1701–1708, 2014. 1, 3

[59] D. W. Tan, S. Z. Gilani, M. T. Maybery, A. Mian, A. Hunt, M. Walters, and A. J. Whitehouse. Hypermasculinised facial morphology in boys and girls with autism spectrum disorder and its association with symptomatology. *Scientific Reports*, 7(1):9348, 2017.

[60] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015. 6

[61] V. Vijayan, K. W. Bowyer, P. J. Flynn, D. Huang, L. Chen, M. Hansen, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris. Twins 3D face recognition challenge. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7, 2011. 3, 6, 8

[62] A. J. Whitehouse, S. Z. Gilani, F. Shafait, A. Mian, D. W. Tan, M. T. Maybery, J. A. Keelan, R. Hart, D. J. Handelsman, M. Goonawardene, et al. Prenatal testosterone exposure is related to sexually dimorphic facial morphology in adulthood. In *Proc. R. Soc. B*, volume 282. The Royal Society, 2015.

[63] C. Xu, T. Tan, S. Li, Y. Wang, and C. Zhong. Learning effective intrinsic features to boost 3D-based face recognition. In *European Conference on Computer Vision*, pages 416–427. Springer, 2006. 6, 8

[64] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008. 6, 8

[65] L. Yin, X. Wei, et al. A 3D facial expression database for facial behavior research. In *Automatic Face and Gesture Recognition*, pages 211–216, 2006. 3, 6, 8

[66] H. Zhou, A. Mian, L. Wei, D. Creighton, M. Hossny, and S. Nahavandi. Recent advances on singlemodal and multimodal face recognition: a survey. *IEEE Transactions on Human-Machine Systems*, 44(6):701–716, 2014. 2

[67] S. Zulqarnain Gilani, F. Shafait, and A. Mian. Shape-based automatic detection of a large number of 3D facial landmarks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4639–4648, 2015.