Towards Large-scale 3D Face Recognition

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

***Abstract*—3D face recognition holds great promise in achieving robustness to pose, expressions and occlusions. However, 3D face recognition algorithms are still far behind their 2D counterparts due to the lack of large-scale datasets. We present a model based algorithm for 3D face recognition and test its performance by combining two large public datasets of 3D faces. We propose a Fully Convolutional Deep Network (FCDN) to initialize our algorithm. Reliable seed points are then extracted from each 3D face by evolving level set curves with a single curvature dependent adaptive speed function. We then establish dense correspondence between the faces in the training set by matching the surface around the seed points on a template face to the ones on the target faces. A morphable model is then fitted to probe faces and face recognition is performed by matching the parameters of the probe and gallery faces. Our algorithm achieves state of the art landmark localization results. Face recognition results on the combined FRGCv2 and Bosphorus datasets show that our method is affective in recognizing query faces with real world variations in pose and expression, and with occlusion and missing data despite a huge gallery. Comparing results of individual and combined datasets show that the recognition accuracy drops when the size of the gallery increases.**

I. Introduction

Face recognition, due to its non-intrusiveness, high accessibility and social acceptability has become the biometric of choice as compared to iris detection, retinal scans and fingerprints [1]. It has applications in a variety of domains [2] including security (e.g., system logon, internet access, and file encryption), surveillance (e.g., border control,suspect tracking and identification) and entertainment (e.g., human computer interaction, 3D animation, and virtual reality).

For decades, researchers have performed face recognition from coloured or grayscale two-dimensional photographs. However, 2D face recognition is plagued with problems of illumination, pose and scale variations [3]. It becomes even more difficult in the presence of affine transformations which are introduced to 2D images during acquisition [4]. Furthermore, facial texture is not always stable for identities as it can change with make up. On the other hand, 3D facial analysis techniques tend to be robust to occlusions, pose and expression variations and are not affected by illumination. In recent years, 3D face acquisition devices have become cost effective and are available off the shelf, and hence 3D face recognition is gaining popularity [5]–[10].

Most existing 3D face databases are comprised of good quality facial scans acquired in highly controlled environments. As compared to 2D face datasets, the number of subjects in the 3D face databases are small and do not address all the uncontrolled real world scenarios like pose and expression variations, occlusions, missing parts or partial data. Hence face recognition algorithms proposed and tested on these datasets exhibit limited capability. Table I summarizes the important features of some of the well known 3D face datasets. The number of identities in the gallery becomes an important factor in identification tasks where a probe is matched with all scans in the gallery. For example, the 2D Colour FERET database [11] has $14,126$ images of $1,199$ individuals whereas FRGCv2 [12], one of the largest 3D face databases, has only $4,007$ scans of 466 individuals. In this context, we propose large-scale 3D face recognition by combining the FRGCv2 [12] and Bosphorus [13] face datasets to obtain $8,673$ scans of 571 individuals containing large pose and expression variations, self occlusion and missing data.

Existing 3D face recognition techniques can be grouped into local or global descriptor based techniques [1], [3]. The latter also include techniques that employ 3D morphable models. The local descriptor based techniques match local 3D point signatures derived from the curvatures, shape index and/or normals whereas the model based approaches construct a 3D morphable face model and fit it to the query faces. Face recognition is performed by matching the model parameters.

In the local descriptor based category, Mian et al. [5], [17] proposed a highly repeatable keypoint detection algorithm for 3D facial scans. They fused the 3D keypoints with 2D Scale Invariant Feature Transform (SIFT) features to develop a robust 3D + 2D multi-modal face recognition system. Experiments were performed on FRGCv2 database. Gupta et al. [18] automatically detected ten anthropometric fiducial landmarks on $1,149$ scans of the self-collected Texas 3D Face database. 3D Euclidean and geodesic distances between all possible pairs of these landmarks were used as features to perform face recognition. The authors reported Rank-1 Recognition Rate (RR) of 96.8% and suggested that the geodesic distances were more reliable in terms of recognition accuracy compared to the Euclidean distance. Queirolo et al. [19] used the local Surface Inter-penetration Measure (SIM) to match 3D faces. The authentication score was obtained by combining the SIM values corresponding to four different facial regions. Experiments were performed on FRGCv2 database. Berretti et al. [9] represented a 3D face with multiple meshDOG keypoints and local geometric histogram descriptors. The most effective features from the local descriptors were then selected to perform face recognition on Bosphorus dataset [13]. Similarly, Drira et al. [20] represented the facial surface by radial curves

978-1-5090-2896-2/16/$31.00 ©2016 IEEE

TABLE I
IMPORTANT FEATURES OF SOME OF THE WELL KNOWN 3D FACE DATABASES. NOTE THAT EACH DATASET EITHER LACKS IN THE NUMBER OF UNIQUE IDENTITIES OR IN THE REAL WORLD VARIATIONS. REFER TO BOWYER ET.AL [1] FOR AN EXHAUSTIVE LIST OF 3D FACE DATASETS.

Database	#Identities	Scans/Id	Total Scans	Expression	Pose	Occlusion	Partial data	Action Units
FRGCv2 [12]	466	1-22	4007	Yes	$\leq \pm 10^\circ$	No	No	No
BU3DFE [14]	100	25	2500	Yes	$\leq \pm 10^\circ$	No	No	No
Bosphorus [13]	105	40-52	4666	Yes	$\leq \pm 90^\circ$	Yes	Yes	Yes
GavabDB [15]	61	9	549	Yes	$\leq \pm 90^\circ$	No	Yes	No
SHREC2008 [16]	61	7	427	Yes	$\leq \pm 35^\circ$	No	No	No

emanating from the nosetip. A Riemannian framework was developed to analyse these curves. Face recognition experiments were performed on FRGCv2 and the occluded faces of Bosphorus database with moderate results. Li et al. [10] proposed a combination of three histogram based features to generate the Histogram of Multiple surface differential Quantities (HOMQ) for 3D face representation, and then used the Sparse Representation based Classifier (SRC) for face recognition. The proposed approach achieved very high recognition performance on the Bosphorus dataset, however, the results on FRGCv2 [12] dataset were not comparable.

In the model based category, Blanz and Vetter [21] proposed a dense correspondence algorithm using optical flow on the texture and the 3D cylindrical coordinates of the face points assuming that the faces are spatially aligned. They constructed a 3D morphable face model from 100 male and female faces each. An arbitrary face was chosen as a reference and the remaining scans were registered to it by iterating between optical flow based correspondence and morphable model fitting. One potential pitfall of the texture based face recognition [21] is that facial texture is not always consistent with the underlying 3D facial morphology e.g. the shape and location of eyebrows. Moreover, this algorithm requires seven manually annotated facial landmarks for initialization. Later, in [8], [22] the authors used the 3D morphable model for face recognition. Experiments were performed on only 150 pairs of 3D faces [22] from FRGCv2 database, although the total number of scans in the database are $4,007$. Passalis et al. [23] proposed an Annotated Face Model (AFM) based on an average facial 3D mesh. The model was created by manually annotating a sparse set of anthropometric landmarks [24] on 3D face scans and then segmenting it into different annotated areas. Later, Kakadiaris et al. [25] proposed elastic registration using this AFM by shifting the manually annotated facial points according to elastic constraints to match the corresponding points of 3D target models in the gallery. Face recognition was performed by comparing the wavelet coefficients of the deformed images obtained from morphing. Passalis et al. [26] further improved the AFM by incorporating facial symmetry to perform pose invariant face recognition. However, the algorithm depends on detection of at least five facial landmarks on a side pose scan.

Recently, Gilani et al. [27] presented a shape based dense correspondence algorithm for landmark detection. The proposed algorithm evolves level set curves with adaptive and uniform geometric speed functions to automatically extract effective seed points for dense correspondence. After centring the faces at nosetip, correspondences are established by minimizing the bending energy between patches around seed points of given faces to those of a reference face. The algorithm was used to detect fiducial landmarks and was not tested for model based face recognition.

We present a model based solution to large-scale 3D face recognition. Our primary contribution is a significant improvement in the morphable model proposed by Gilani et al. [27]. The existing algorithm uses a heuristic approach to detect the nosetip in frontal scans. This heuristic fails in case of large pose variations or occlusion. Furthermore, initial coarse mesh registration is necessary for good quality model fitting. This aspect also suffers in the presence of large pose variations and occlusion when registration is done based on a single point (i.e. nosetip). We therefore propose a Fully Convolutional Deep Network (FCDN) to identify three fiducial landmarks on each 3D scan. We empirically demonstrate that the landmarks detected using our proposed network are more accurate and lead to improvement in the landmark detection scheme proposed by Gilani et al. [27]. The second improvement is in the design of the speed function used to evolve level set curves. We use a single speed function as opposed to two [27] and introduce significant changes such that good quality repeatable seed points are detected on the complete face. The secondary contribution of this paper is to present a large-scale 3D face recognition system by combining the FRGCv2 [12] and Bosphorus [13] face datasets to obtain $8,673$ scans of 571 individuals containing large pose and expression variations, self occlusion and missing data. The scans in both datasets are acquired with different sensors. We denote this combined dataset as LSDB. The literature reports face recognition results on individual datasets mentioned in Table I but to the best of our knowledge none have performed face recognition at a large-scale. We follow Gilani et al's. [27] method of establishing correspondence between the 571 identities of LSDB and create a statistical model. The model is fitted to the remaining $8,102$ unseen query faces to obtain model parameters. Face recognition is performed by matching the model parameters of the query scan with those of the gallery scans. Our results are comparable with the state-of-the-art on the individual datasets, however

they do not perform equally well on the LSDB. This shows that there is a need to test the state-of-the-art algorithms on large-scale datasets, to advance 3D face recognition research.

II. Proposed Algorithm

Figure 1 shows the block diagram of the proposed face recognition system and each component is explained in detail below.

A. The Landmark Detection Network

CNNs have recently been used for semantic segmentation [28], [29] and boundary prediction [30] in RGB images. Long et al. [31] adapted and extended a deep classification architecture to learn from whole image inputs and whole image ground truths for semantic segmentation. They trained a fully convolutional network end-to-end, pixel-to-pixel and showed that their results on semantic segmentation outperformed the state-of-the-art. We adapted this model to perform landmark detection in 3D faces using a point-to-point or pixel-to-pixel identification architecture. Let $C(k, n, s)$ denote a convolutional layer with kernel size $k \times k$, n filters and stride s, RL denote a rectified linear unit, $P(k, s)$ denote a max pooling layer with kernel size $k \times k$ and stride s, $FC(n)$ denote a fully connected layer with n filters and $D(r)$ denote a dropout layer with drop our ratio r. Let $Sk(k, n, s)$ denote the skip layer where the parameters have a similar meaning as in the convolutional layer, SU denote the sum layer and $DC(u)$ denote the deconvolution layer where u is the upsampling ratio. The architecture of our proposed Fully Convolutional Deep Network (FCDN) is enumerated as follows:

$C(3,64,1) \rightarrow RL \rightarrow C(3,64,1) \rightarrow RL \rightarrow P(2,2) \rightarrow C(3,128,1) \rightarrow RL \rightarrow C(3,128,1) \rightarrow RL \rightarrow P(2,2) \rightarrow C(3,256,1) \rightarrow RL \rightarrow C(3,256,1) \rightarrow RL \rightarrow C(3,256,1) \rightarrow RL \rightarrow P(2,2) \rightarrow C(3,512,1) \rightarrow RL \rightarrow C(3,512,1) \rightarrow RL \rightarrow C(3,512,1) \rightarrow RL \rightarrow P(2,2) \rightarrow C(3,512,1) \rightarrow RL \rightarrow C(3,512,1) \rightarrow RL \rightarrow C(3,512,1) \rightarrow RL \rightarrow P(2,2) \rightarrow FC(4096) \rightarrow RL \rightarrow D(0.5) \rightarrow FC(4096) \rightarrow RL \rightarrow D(0.5) \rightarrow FC(2) \rightarrow DC(2) \rightarrow Sk(1,2,1) \rightarrow SU \rightarrow DC(2) \rightarrow Sk(1,2,1) \rightarrow SU \rightarrow DC(8)$

We use the real scans from Bu3DFE [14] to train our FCDN. The dataset contains $2,500$ scans from 100 subjects. Each subject is scanned in neutral pose and six expressions with four intensity levels. We render each scan from five viewing angles, that is, frontal, $\pm 15^\circ$ in pitch and $\pm 15^\circ$ in roll. Next, their depth images are generated by fitting a surface of the form $z(x, y)$ to each 3D pointcloud using the *gridfit* algorithm [32]. We also calculate the Cartesian surface normals (n_x, n_y, n_z) of each vertex in the pointcloud and convert them to spherical coordinates (n_θ, n_ϕ, n_r) where θ is the azimuth, ϕ is the elevation and r is the radius of the normal. A surface similar to the one used for depth images is fitted to the former two components of the normal. The depth, azimuth and elevation images are used as the three channels, instead of the usual RGB channels, as input images to the FCDN. The dataset comes with 83 ground truth points and we include the nosetip detected by Gilani et al. [33]. We use only three ground truth landmarks of outer eye corners and the nosetip. These are biologically significant landmarks [24] which define some high curvature points of the upper face region and are mostly invariant to expression. Ground truth landmark locations are similarly projected on a 2D surface, dilated with a *disk* shaped structure of size 15×15 and converted into a binary 2D image.

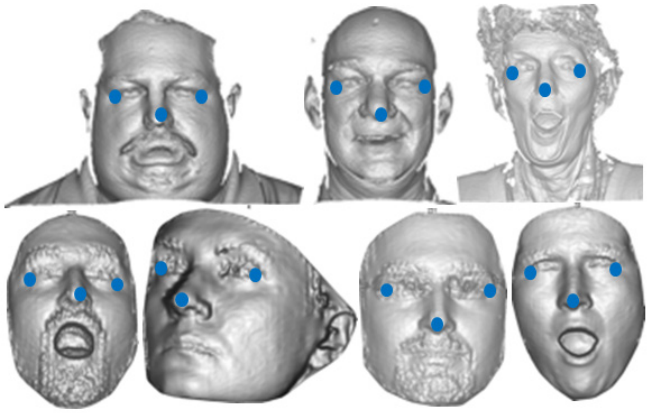

Fig. 2. Three landmarks detected by our proposed FCDN depicted on three identities of FRGCv2 (top row) and four identities of Bosphorus (bottom row) datasets.

Hence, our training data consists of $12,500$ 3D faces from 100 identities. Each identity has 25 images in varying expressions, while each of the 25 images is then rendered in 5 different poses. We use the faces of 80 identities for training and 20 identities for validation. The FCDN is initialized with the FCN-8s model [31] parameters and we randomly initialize the class scoring layer. We use a momentum of 0.9, a weight decay of 0.0005 and train the network for 300 epochs using Matconvenet [34].

Let $\mathbf{F_j} = [x_i, y_i, z_i]^T$ $(j = 1, \ldots, N$ and $i = 1, \ldots, P_j)$ be a real arbitrary 3D face scan. We obtain the depth, azimuth and elevation image of this face by fitting a surface to it as described above and pass it through the learned FCDN. The output is a binary mask of landmark locations. To identify the three designated landmarks and label them, we apply some basic morphological operations to this binary image. The landmark mask is then converted to 3D point coordinates by utilizing the original Cartesian coordinates of the FCDN input scan. The detected landmarks are denoted by $\mathcal{L}_j = [x_k, y_k, z_k]^T$, where $k = 1, \ldots, 3$. Figure 2 shows the detected landmarks on sample faces from our LSDB.

B. Establishing Dense Correspondence

The FCDN is used to detect three landmarks on the 3D faces of 571 identities in our gallery. Using the nose tip as the centre, we crop a sphere of $90mm$ radius to discard non-facial regions. We randomly select a 3D face from the training set as the source and register the rest of the faces to it using the three landmarks.

The level set interface [35], [36] at point i is represented by $\gamma(i) = 0$ and the shortest distance from this point to the

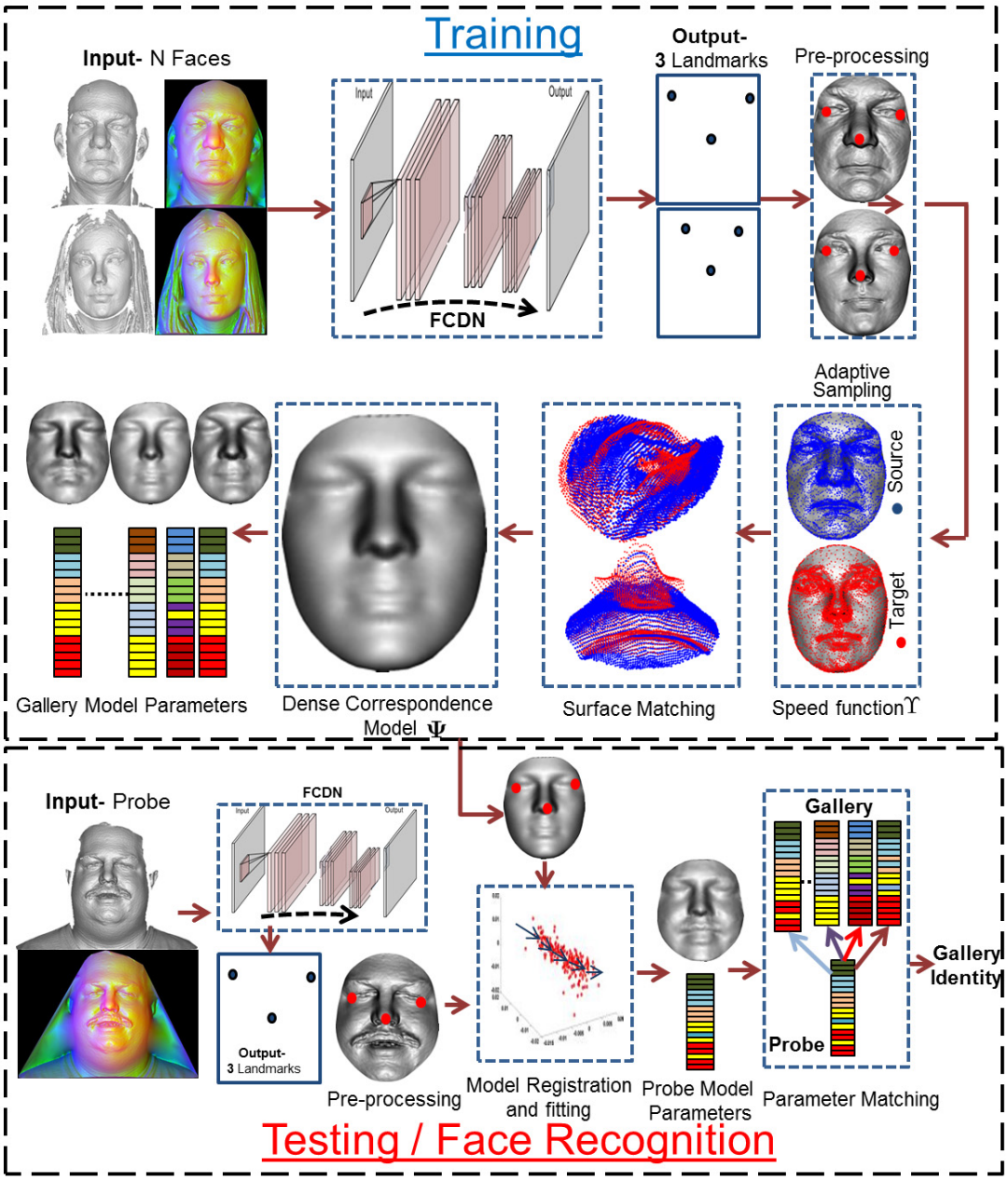

Fig. 1. Block diagram of the proposed algorithm. A dense correspondence model is created in the training module. In the testing module, the probe is matched with the gallery model parameters for face recognition.

boundary is given by $|\gamma(i)|$. The level set equation is given by,

$$\gamma_t + \Upsilon \nabla |\gamma| = 0 \tag{1}$$

where Υ is the propagation speed of the interface front and sets the density of point sampling. $\Upsilon = 1$ denotes uniform sampling over the mesh. The value of Υ at every vertex of the 3D face can be set adaptively. We use the curvature property at each vertex to set the value of the speed function. The mean curvature H is given by [37], [38],

$$H = \nabla . \frac{\nabla \gamma}{|\nabla \gamma|} = \frac{\left\{ \begin{array}{c} (\gamma_{yy} + \gamma_{zz})\gamma_x^2 + (\gamma_{xx} + \gamma_{zz})\gamma_y^2 + (\gamma_{xx} + \gamma_{yy})\gamma_z^2 \\ -2\gamma_x\gamma_y\gamma_{xy} - 2\gamma_x\gamma_z\gamma_{xz} - 2\gamma_y\gamma_z\gamma_{yz} \end{array} \right\}}{(\gamma_x^2 + \gamma_y^2 + \gamma_z^2)^{\frac{3}{2}}}$$

while the Gaussian curvature K is,

$$K = \frac{\left\{ \begin{array}{c} \gamma_x^2(\gamma_{yy}\gamma_{zz} - \gamma_{yz}^2) + \gamma_y^2(\gamma_{xx}\gamma_{zz} - \gamma_{xz}^2) + \gamma_z^2(\gamma_{xx}\gamma_{yy} - \gamma_{xy}^2) \\ +2[\gamma_x\gamma_y(\gamma_{xz}\gamma_{yz} - \gamma_{xy}\gamma_{zz}) + \gamma_y\gamma_z(\gamma_{xy}\gamma_{xz} - \gamma_{yz}\gamma_{xx}) \\ +\gamma_x\gamma_z(\gamma_{xy}\gamma_{yz} - \gamma_{xz}\gamma_{yy})] \end{array} \right\}}{(\gamma_x^2 + \gamma_y^2 + \gamma_z^2)^2}$$

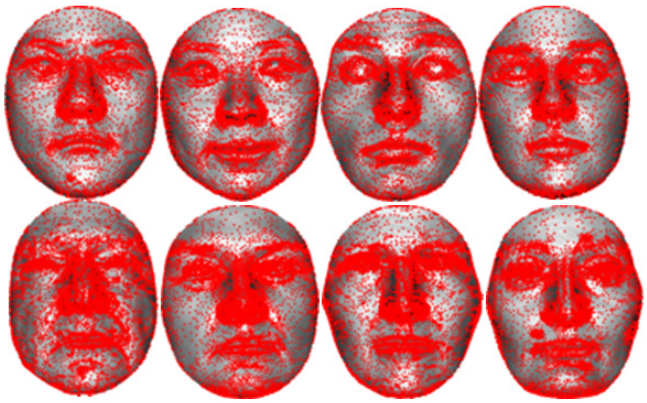

Fig. 3. The adaptively sampled points using a single curvature dependent speed function. Notice how well the points cover the high curvature areas while being sparse in the quasi-planar ones. Top and bottom rows show faces from FRGCv2 Bosphorus datasets respectively.

The two principle curvatures at each point i are related to the Mean and Gaussian curvatures such that $\boldsymbol{\kappa}_1 = H_i + \sqrt{H_i^2 - K_i}$ and $\boldsymbol{\kappa}_2 = H_i - \sqrt{H_i^2 - K_i}$. The Curvedness at each point is a function of the two principle curvatures and is defined as:

$$\mathcal{C} = \sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}} \tag{2}$$

and the Shape Index ($\mathcal{SI}$) is given by:

$$\mathcal{SI} = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_1 + k_2}{k_1 - k_2} \tag{3}$$

We unit normalize both the descriptors. Instead of thresholding the curvature dependent speed function [27] we make it data driven at every point and define it as,

$$\Upsilon = \frac{\mathcal{C} + \mathcal{SI}}{2} \tag{4}$$

The speed function Υ slows down the propagation of the front in areas of high curvature and speeds it up in the quasi-planar regions, thereby sampling reliable seed points all over the facial surface. We perform farthest point adaptive sampling [39] of each 3D face $\mathbf{F}_j$ to obtain $\widetilde{\mathbf{F}_j} = [x_k, y_k, z_k]^T$ $(k = 1, \ldots, P^\Upsilon)$. P^Υ is the user defined number of sampled points, $P^\Upsilon < \min(P_j)$ and $\widetilde{\mathbf{F}_j} \subset \mathbf{F}_j$. In our experiments we set the value of P^Υ to $5,000$. Figure 3 shows the adaptively sampled points on the pre-processed faces of eight identities of LSDB.

So far we have extracted reliable seed points on all faces of the training dataset. The goal now is to establish correspondence between these points across the faces. This is done by matching the surface around each point on a template face with the surface around a corresponding point on the target face. We randomly select a 3D face from the training set as the source and find its correspondence to the rest of the faces in the training set in a pair wise scheme. Our aim is to is to find the correspondence between a point p on the source face to a point q on a given target face. For this purpose, a small surface of radius r_p around each point p on the sampled source face is extracted and denoted by $\mathbf{S}_s$. Although the point itself is selected from the sampled face $\widetilde{\mathbf{F}_j}$, the surface is cropped from the pre-processed input face $\mathbf{F}_j$ to ensure a rich sampling of points in the surface. Recall that both the source and the target face have been registered using the three landmarks identified by our FCDN. Hence, it is safe to assume that the point q on a target face corresponding to the point p on source face can be found on the target face within a small region that neighbours point p. Therefore, we crop out small surfaces of radius r_p on the target face $\mathbf{S}_t$ within a neighbourhood of $20mm$ of the point p.

We match the surfaces $\mathbf{S}_s$ with each target surface $\mathbf{S}_t$ by finding the non-rigid shape difference $\Delta(s,t)$. The shape difference is given by $\Delta(s,t) = \frac{\beta_{st} + \beta_{ts}}{2}$, where β_{st} is the amount of bending energy required to deform $\mathbf{S}_s$ to $\mathbf{S}_t$ and is measured using the 2D thin-plate spline model [27], [40]. We say that a point q on the target face coarsely corresponds to point p on the source face if the shape difference $\Delta(s,t)$ in their corresponding patches is below a threshold t_Δ. If no matches satisfy this criteria then we discard this point and move to the next point on the source face. To improve the correspondence between the two points we repeat the surface matching process between the point p on the source face and all points within a radius of $5mm$ of the point q on the pre-processed target face. The choice of selecting initial points from the sampled faces $\widetilde{\mathbf{F}_j}$ for matching and later incorporating the pre-processed faces $\mathbf{F}_j$ reduces the computational complexity of the matching process and at the same time yields good quality correspondences.

Correspondence between the source 3D scan and the remaining sampled 3D faces $\widetilde{\mathbf{F}_j}$ of the training set is established by repeating the coarse to fine surface matching process for all points P^Υ. The output of this step is a set of corresponding faces $\mathbf{F}_j^c = [x_p, y_p, z_p]^T$, where $j = 1, \ldots, N$ and $p = 1, \ldots, P$.

C. Model Fitting and Parameter Extraction

Equipped with N densely corresponding faces $\mathbf{F}_j^c$ we form a 3D Deformable Model (3DM), $\boldsymbol{\Psi} = [\mathbf{f}_1^c, \mathbf{f}_2^c, \ldots, \mathbf{f}_N^c,]$, where $\mathbf{f}^c = [x_1, \ldots, x_p, y_1, \ldots, y_p, z_1, \ldots, z_p]^T$ and $p = 1, \ldots, P$. Following [27] we model the 3DM by a multivariate Gaussian distribution whose eigenvalue decomposition is given by $\mathbf{USV}^T = \boldsymbol{\Psi}_m$. Here $\mathbf{US}$ are the principal components (PCs), the columns of V are their corresponding loadings, $\mathbf{S}$ is a diagonal matrix of eigenvalues and the row means of Ψ_m are all 0. We retain 99% of the energy corresponding to the first n columns of $\mathbf{U}$. Furthermore, the mean face of the model is given by $\overline{\mathbf{F}^c} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{F}_j^c$.

A query face is first passed through the FCDN to detect the outer eye corners and the nosetip. These landmarks are used to align the query face to the mean face of the model through rigid registration. Next, we find the mapping of points between the query and the model mean face $\overline{\mathbf{F}^c}$ by searching

TABLE II

COMPARISON OF MEAN LANDMARK LOCALIZATION ERROR ON $4,007$ SCANS OF FRGCV2 WITH GILANI ET AL. [27] (CVPR-15). THE RESULTS ARE BASED ON FITTING A DEFORMABLE MODEL ON THE TEST DATASET. RESULTS OF LANDMARKS THAT OCCUR IN PAIRS HAVE BEEN AVERAGED.

Author	OEC	IEC	NR	NT	NC	MC(L)	MC(R)	ULC	LLC	CT	NB	Mean
CVPR-15 [27]	4.1	2.9	3.6	2.7	4.3	5.3	4.4	3.3	4	4.2	4.1	3.9±2.8
This paper Regn with NT	3.9	2.9	3.4	2.7	4.1	4.9	4.0	3.3	4	4.1	4.0	3.7±2.6
This paper Regn with OEC	3.5	2.8	3.4	2.8	3.9	4.7	3.9	3.2	3.9	3.9	3.7	3.6±2.5
This paper Regn with OEC & NT	3.3	2.7	3.2	2.5	3.6	4.2	3.3	3.0	3.7	3.6	3.5	3.3±2.3
EC-Eye Corner(Outer/Inner), NR-Nasal Root, NT-Nosetip, MC-Mouth Corner, LC-Lip Corner(Upper/Lower), CT-Chin Tip, NB-Nasal Bridge												

for the Nearest Neighbour (NN) of each point of $\overline{\mathbf{F}^c}$ in $\mathbf{Q}$ using the k-d tree data structure [41]. After vectorization, the query face can be parametrized by the statistical model such that $\mathbf{m}_q = \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\mu}_\Psi$, where the vector $\boldsymbol{\alpha}$ contains the parameters which are used to vary the shape of the model and $\mathbf{m}_q$ is the vectorized form of the query model $\mathbf{M}_q$ generated by the 3DM. The vector $\boldsymbol{\alpha}$ is given by,

$$\boldsymbol{\alpha} = \mathbf{U}^T(\mathbf{q}^c - \boldsymbol{\mu}_\Psi) \tag{5}$$

where $\mathbf{q}^c$ is the vectorized corresponded query face.

III. RESULTS AND ANALYSIS

A. Datasets

Our Large-scale Database (LSDB) comprises of $8,673$ scans from FRGCv2 [12] and Bosphorus [13] datasets. FRGCv2 comprises of $4,007$ scans from 466 identities of different ethnicities and age groups. The scans are mostly frontal with minor ($\pm 10^\circ$) pose variations. The facial expressions range from neutral to extreme but are not labelled as such. Manual landmark annotations provided by Szeptycki et al. [46] and Creusot et al. [42] were used as ground truth for comparison. The Bosphorus dataset contains $4,666$ 3D faces from 105 subjects with considerable variation in ethnicity and age. The dataset is structured into Action Units (AU) including the generic expressions of happy, sad, surprise, fear, disgust, anger and neutral. Poses vary within a range of $\pm 90^\circ$ in both yaw and pitch along with some cross pose variations in yaw and pitch simultaneously. The dataset also contains sans with four different types of occlusions. Ground truth landmark locations are provided with the dataset.

TABLE III

COMPARISON OF LANDMARK LOCALIZATION RESULTS FROM FCDN WITH THE STATE-OF-THE-ART ON FRGCV2 AND BOSPHORUS DATASET. OEC = OUTER EYE CORNER, NT = NOSETIP .

Mean ± SD of Localization Error(mm)					
	Author	# Images	OEC (L)	OEC(R)	NT
FRGCv2	Creusot [42]	4007	6.00 ± 3.03	5.87 ± 3.11	3.35 ± 2.00
	Segundo [43]	4007	3.35 ± 2.33	3.69 ± 2.26	2.73 ± **1.39**
	Perakis [44]	975	5.83 ± 3.42	5.58 ± 3.33	4.09 ± 2.41
	Sukno [45]	4007	4.60 ± 2.70	4.70 ± 2.70	**2.30** ± 1.70
	FCDN	4007	**2.53** ± **1.96**	**2.48** ± **2.03**	2.36 ± 1.74
Bosphorus	Creusot [42]	4339	6.09 ± 5.02	4.18 ± 3.79	4.60 ± 2.40
	Sukno [45]	4339	5.13 ± 4.01	5.05 ± 3.86	3.00 ± 2.75
	FCDN	4666	**3.34** ± **2.33**	**3.31** ± **2.28**	**2.60** ± **1.92**

B. Facial Landmark Detection

Although the main purpose of this paper is to propose a system for large-scale face recognition we report results on facial landmark detection error to evaluate the efficacy of FCDN and the quality of dense correspondence. The proposed FCDN provides three landmarks which can be compared with algorithms that are designed to detect only a sparse ($\sim 12-15$) set of landmarks. Table III shows the results of our proposed technique on FRGCv2 and Bosphorus datasets separately and compares them with the state-of-the-art. The model based algorithms on the other hand are designed to detect a large number of fiducial landmarks. We specifically compare our results with Gilani et al. [27] and in the interests of fairness, follow the same experimental protocol and model fitting procedure. To analyse the affect of initial seed points for registration in building the dense correspondence model as well as in the fitting process, we use different seed points in the landmarking experiments and report the results. Table II depicts these results and compares them with Gilani et al. [27]. The improvement in the landmark localization error with our proposed changes is significant ($p < 0.05$) even when only the nosetip or the outer eye corners are used to register the meshes. We did not find any significant improvement in the results when using the outer eye corners only as compared to the nosetip for registration. However, when all the three landmarks, i.e. the nosetip and the outer eye corners detected by the FCDN are used for coarse registration of the meshes, the landmark localization error is reduced by 20% and the improvement is highly significant ($p < 0.01$).

C. Face Recognition

In order evaluate the efficacy of our proposed algorithm for face recognition we first perform the task on the FRGCv2 and Bosphorous datasets individually. This also enables us to compare our results with the state of the art. We then apply our algorithm to large-scale face recognition to assess its performance on datasets with a huge gallery and most of the real world uncontrolled variations.

TABLE IV
RANK-1 RECOGNITION RESULTS (IN %AGE) ON THE COMBINED (FRGCV2 AND BOSPHORUS) LSDB .

	Expressions				Poses					Occlusions					All
	Neutral	AU	Expr	All	YR<90	YR90	PR	CR	All	Eye	Mouth	Glasses	Hair	All	
# scans	1499	2150	2707	6338	525	210	419	211	1365	105	105	104	67	381	**8102**
Rank-1 RR	97.3	91.5	90.7	92.8	88.9	74.6	96.5	94.1	89.8	97.1	90.3	96.8	90.2	93.8	92.13

AU=Action Units; YR=Yaw Rotation; PR= Pitch Rotation; CR= Cross Rotation

Face recognition is performed by first creating a dense correspondence model of the first available neutral scans of each dataset. The model is then fitted on the remaining scans in to obtain the model parameters. We calculate the cosine distance $d_f = \cos^{-1} \frac{\boldsymbol{\alpha}_m^T \boldsymbol{\alpha}_q}{\|\boldsymbol{\alpha}_m^T\|_2 \|\boldsymbol{\alpha}_q\|_2}$ between the model parameters of the probe and each gallery face. The probe is assigned the identity of the gallery face with which it has the smallest distance. Since FRGCv2 [12] dataset is unlabelled for expression, we follow the protocol of Mian at al. [6] to divide it into neutral$(1,481)$ and non-neutral$(2,060)$ expression scans. The Rank-1 Recognition Rate(RR) on the FRGCv2 and Bosphorus datasets is shown in Tables V and VI respectively. The results on both the datasets are comparable with the state-of-the-art. It can be seen that the algorithms that perform well in case of Bosphorus with a small gallery of 105 identities fail to perform at par in the case of FRGCv2 with a gallery of 466 identities. This indicates that the researchers need to define more challenging protocols for 3D face recognition to analyse their algorithms.

To this end, we perform face recognition on LSDB by creating a dense correspondence model of the 571 first available neutral scans and then fitting it to the remaining $8,102$ scans to obtain the model parameters. The FRGCv2 scans are assigned to the Neutral and Expression categories. The Bosphorus [13] dataset is labelled but there are 18 unlabelled scans which we included in the neutral expression category. The results of Rank-1 Recognition Rate (RR) are reported in Table IV. Our proposed model generalizes well over large pose variations and occlusions. Note that while results reported on the individual datasets (FRGCv2 and Bosphorus) have reached near saturation, the performance on the combined LSDB still has space for improvement. Furthermore, it is expected that the results of the state-of-the-art algorithms will scale down in case of the large-scale datasets just like ours.

TABLE V
COMPARISON OF 3D FACE RECOGNITION RESULTS WITH THE STATE-OF-THE-ART IN TERMS OF RANK-1 IDENTIFICATION RATE (I-RATE) AND VERIFICATION RATE (V-RATE) AT 0.1% FAR.

Author	Neutral		Non-neutral		All	
	I-Rate	V-Rate	I-Rate	V-Rate	I-Rate	V-Rate
Mian et al. [6]	99.4%	99.9%	92.1%	96.6%	96.1%	98.6%
Kakadiaris et al. [25]	-	99.0%	-	95.6%	97.0%	97.3%
Al-Osaimi et al. [7]	97.6%	98.4%	95.2%	97.8%	96.5%	98.1%
Drira et al. [20]	98.2%	-	96.8%	-	97.7%	97.1%
Li et al. [10]	-	-	-	-	96.3%	-
This paper	99.5%	99.7%	95.8%	95.6%	97.9%	98.4%

IV. CONCLUSION

The results of 3D face recognition on most known benchmark datasets have reached near saturation. This is mostly due to the small sizes of these datasets. This paper may be considered as a first step towards large-scale 3D face recognition. It proposes a model based algorithm for 3D face recognition that performs at par with the existing state of the art when tested on individual benchmark datasets. For landmark localization, the proposed algorithm outperforms state of the art. However, when two datasets are combined, the face recognition performance of the proposed algorithm drops. This is likely to be the case for other algorithms. At the moment, a comparative study of existing 3D face recognition algorithms on large-scale datasets cannot be performed because of the lack of such datasets and the unavailability of the implementations (code) of these algorithms. To take 3D face recognition to the next level, a large-scale dataset of 3D faces along the lines of faces in the wild [47] must by developed. Combining existing datasets can serve as a good starting point, however, these datasets must also be augmented to increase the number of identities.

ACKNOWLEDGMENTS

This research was supported by ARC Discovery grant DP160101458 and NHMRC Project grant APP1109057. The authors thank NVIDIA for providing the Tesla K-40 GPU used in our experiments.

REFERENCES

[1] 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*, vol. 101, no. 1, pp. 1–15, 2006.
[2] A. Mian and N. Pears, "3d face recognition," in *3D Imaging, Analysis and Applications*. Springer, 2012, pp. 311–366.
[3] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2d and 3d face recognition: A survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885–1906, 2007.

TABLE VI
COMPARISON OF RANK-1 RECOGNITION RESULTS (IN %AGE) WITH THE STATE-OF-THE-ART ON BOSPHORUS DATASET.

Author	Expressions		Poses				Occlusions				All
	AU	Expr	YR<90	YR90	PR	CR	Eye	Mouth	Glasses	Hair	
# scans	2150	647	525	210	419	211	105	105	104	67	4543
Drira et al. [20]	-	-	-	-	-	-	97.1	78.0	94.2	81.0	-
Berretti et al. [9]	95.7	95.7	81.6	45.7	98.3	93.4	93.1	93.3	93.4	93.2	93.4
Li et al. [10]	99.2	96.6	84.1	47.1	99.5	99.1	100.0	100.0	100.0	95.5	96.6
This Paper	98.9	96.7	88.2	81.4	98.4	96.8	100.0	96.1	100.0	96.3	96.3

AU=Action Units; YR=Yaw Rotation; PR= Pitch Rotation; CR= Cross Rotation

[4] S. Berretti, N. Werghi, A. Del Bimbo, and P. Pala, "Selecting stable keypoints and local descriptors for person identification using 3d face scans," *The Visual Computer*, vol. 30, no. 11, pp. 1275–1292, 2014.

[5] A. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE TPAMI*, vol. 29, no. 11, pp. 1927–1943, 2007.

[6] A. Mian, M. Bennamoun, and R. Owens, "Keypoint detection and local feature matching for textured 3D face recognition," *IJCV*, vol. 79, no. 1, pp. 1–12, 2008.

[7] F. Al-Osaimi, M. Bennamoun, and A. Mian, "An expression deformation approach to non-rigid 3D face recognition," *IJCV*, vol. 81, no. 3, pp. 302–316, 2009.

[8] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE TPAMI*, vol. 25, no. 9, pp. 1063–1074, 2003.

[9] S. Berretti, N. Werghi, A. Del Bimbo, and P. Pala, "Matching 3D face scans using interest points and local histogram descriptors," *Computers & Graphics*, vol. 37, no. 5, pp. 509–525, 2013.

[10] 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*, vol. 113, no. 2, pp. 128–142, 2014.

[11] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[12] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer *et al.*, "Overview of the face recognition grand challenge," in *IEEE CVPR*, 2005.

[13] 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*. Springer, 2008, pp. 47–56.

[14] L. Yin, X. Wei *et al.*, "A 3D facial expression database for facial behavior research," in *Automatic Face and Gesture Recognition*, 2006, pp. 211–216.

[15] A. B. Moreno and A. Sánchez, "Gavabdb: a 3d face database," in *Proc. 2nd COST Workshop on Biometrics on the Internet*, 2004, pp. 75–80.

[16] F. B. Ter Haar, M. Daoudi, and R. C. Veltkamp, "Shape retrieval contest 2008: 3d face scans."

[17] 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*, vol. 89, no. 2-3, pp. 348–361, 2010.

[18] S. Gupta, M. K. Markey, and A. C. Bovik, "Anthropometric 3d face recognition," *International journal of computer vision*, vol. 90, no. 3, pp. 331–349, 2010.

[19] C. Queirolo, L. Silva, O. Bellon, and M. Segundo, "3D face recognition using simulated annealing and the surface interpenetration measure," *IEEE TPAMI*, vol. 32, no. 2, pp. 206–219, 2010.

[20] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3D face recognition under expressions, occlusions, and pose variations," *IEEE TPAMI*, vol. 35, no. 9, pp. 2270–2283, 2013.

[21] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *ACM Conference on Computer Graphics and Interactive Techniques*, 1999.

[22] V. Blanz, K. Scherbaum, and H.-P. Seidel, "Fitting a morphable model to 3d scans of faces," in *IEEE ICCV*, 2007.

[23] 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.

[24] L. Farkas, "Anthropometry of the head and face in clinical practice," *Anthropometry of the head and face, 2nd Ed*, pp. 71–111, 1994.

[25] 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*, vol. 29, no. 4, pp. 640–649, 2007.

[26] G. Passalis, P. Perakis, T. Theoharis, and I. A. Kakadiaris, "Using facial symmetry to handle pose variations in real-world 3D face recognition," *IEEE TPAMI*, vol. 33, no. 10, pp. 1938–1951, 2011.

[27] S. Zulqarnain Gilani, F. Shafait, and A. Mian, "Shape-based automatic detection of a large number of 3D facial landmarks," in *IEEE CVPR*, 2015, pp. 4639–4648.

[28] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *ICML*, 2014.

[29] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[30] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[32] J. DErico, "Surface fitting using gridfit," in *MATLAB Central File Exchange*, 2008.

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

[34] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab."

[35] J. A. Sethian, "Evolution, implementation, and application of level set and fast marching methods for advancing fronts," *J. of Computational Physics*, vol. 169, no. 2, pp. 503–555, 2001.

[36] J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge University Press, 1999, vol. 3.

[37] G. Peyré and L. Cohen, "Geodesic computations for fast and accurate surface remeshing and parameterization," in *Elliptic and Parabolic Problems*. Springer, 2005, pp. 157–171.

[38] G. Peyré and L. D. Cohen, "Geodesic remeshing using front propagation," *International Journal of Computer Vision*, vol. 69, no. 1, pp. 145–156, 2006.

[39] G. Peyré, "The numerical tours of signal processing-advanced computational signal and image processing," *IEEE Computing in Science and Engineering*, vol. 13, no. 4, pp. 94–97, 2011.

[40] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE TPAMI*, vol. 11, no. 6, pp. 567–585, 1989.

[41] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[42] C. Creusot, N. Pears, and J. Austin, "A machine-learning approach to keypoint detection and landmarking on 3D meshes," *IJCV*, vol. 102, no. 1-3, pp. 146–179, 2013.

[43] M. Segundo, L. Silva, P. Bellon, and C. C. Queirolo, "Automatic face segmentation and facial landmark detection in range images," *IEEE Tran. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 5, pp. 1319–1330, 2010.

[44] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, "3D facial landmark detection under large yaw and expression variations," *IEEE TPAMI*, vol. 35, no. 7, pp. 1552–1564, 2013.

[45] F. M. Sukno, J. L. Waddington, and P. F. Whelan, "3-D facial landmark localization with asymmetry patterns and shape regression from incomplete local features," *Transactions on Cybernetics*, vol. 45, no. 9, pp. 1717–1730, 2015.

[46] P. Szeptycki, M. Ardabilian, and L. Chen, "A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking," in *IEEE- Biometrics: Theory, Applications, and Systems*, 2009.

[47] 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 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.