A CNN Based Method for Quantitatively Evaluating 3D Face Reconstruction Algorithms

Miao Sun¹, Gurjeet Singh² and Patrick Yin Chiang¹
¹ Department of Microelectronics, Fudan University, No. 825 Huatuo Road, Shanghai, China
² Department of Electrical Engineering and Computer Sciences, Oregon State University, Corvallis, USA

Abstract. 3D face reconstruction is an attractive topic in computer vision. We have seen dramatic rise in its development recently. Now the state-of-the-art method can reconstruct a face from a single 2D face image. Since the work conditions and cost of a 3D scanner is limited, these 3D reconstruction algorithms are tough to evaluate. Some researchers had proposed different algorithms by measuring the distances between reconstructed images and true depth face previously. Inspired by CNN’s powerful representative ability, here we investigate the issue from CNN’s perspective by comparing the performances of various 3D face reconstruction algorithms on two classification tasks. One is an identity classification task for evaluating the intra-dataset discrimination and another is a binary classification for reconstructed faces and 3D ground truth faces for evaluating the inter-dataset discrimination. To accomplish this, we adopt hourglass neural network as backbone to perform comparisons for three 3D reconstructed algorithms on two datasets. In the end, we show the pros and cons of various algorithms and explicit the mechanisms of the results.

1. Introduction
The existing 3D face reconstructed algorithms can be divided into two categories: one is based on estimating the 3D morphable model (3DMM) parameters [1-5], the other is the end-to-end method [6-8]. For the former, this methodology is based on a model obtained by dimension decreasing manner to describe 3D face space. The original standard parameters can be obtained from existing 3D face datasets such as BFM and Face Warehouse dataset. Then some fitting algorithms are used to minimize the distance between the parameters of the 3D reconstructed face and ground truth 3DMM parameters. In Ref. [1], a two-stage cascaded deformable shape fitting method was proposed for face landmark localization and tracking. However, this work was only effective for visible landmarks. Ref. [2] reconstructed 3D face from low quality photo collections with fewer images by fitting a 3DMM. In Refs. [3-5], Zhu et al. increased the accuracy of 3D face alignment under large pose, illumination, expression gradually. For end-to-end methods, they need exploring new presentations of the 3D face. In Ref. [6], authors combined a FAN (face alignment network) with a depth estimating network to get the 3D face from a single image. Ref. [7] proposed a novel volumetric representation of the 3D facial geometry and trained a simple CNN to regress the volumetric model from the 2D image directly. In Ref. [8], a new representation of a 3D face called UV position map was introduced. The UV position map can project the dense correspondence to the semantic meaning of each point on UV space. Besides, CNN in this work was very light-weight compared with previous works. Now the state-of-art methods all achieve good performance under large pose, illumination, expression variants. Depending on naked eyes are
hard to determine which one is better except for some extremely failure images. So a quantitative metric is necessary for these 3D reconstructed algorithms. The related works in Refs. [9, 10] used a similar manner which evaluated the performance of different methods by measuring differences between the two 3D faces. In Ref. [9], the 3D reconstructed face was aligned to the true depth face by Iterative Closest Point (ICP) method. Then the distance between these two faces was defined as Signal to Noise Ratio (SNR). Also in Ref. [10], 3D Root-Mean-Square Error (3DRMSE) was used as a metric between the 3D reconstructed face and 3D face scan. Furthermore, the work also released a very dedicated 3D face reconstruction competition on the 2018 13th IEEE Conference on Automatic Face & Gesture Recognition. For the different methodologies of the aforementioned algorithms, these two manifolds method have corresponding merit and demerit. For 3DMM based methods, they have better integrity but minor discrimination in the same dataset. Inversely, end-to-end methods can reconstruct more discriminant faces, while they are likely to generate incomplete faces. Considering the heterogeneity among various 3D face reconstructed algorithms, we design two classification tasks to evaluate their performances comprehensively. For the 3D reconstructed faces from the same datasets by the same algorithm, we use an identity classification task to estimate the discrimination ability intra datasets. Furthermore, a binary classification task between individual 3D reconstructed face datasets and 3D ground truth face datasets is tested for every 3D reconstructed algorithm to estimate the inter-dataset discrepancy. These years, hourglass network [11] has shown its feature extraction effectiveness in landmark detection for human pose estimation and face alignment. Here we adopt the hourglass network as a basic component, then a residual block from [12] combined to build a CNN for evaluation. To compare these algorithms fairly, we perform the procedure on three algorithms [1, 7, 8] across two datasets: Texas 3DFRD [13-15] and ND-2006 Data Set [16]. In the end, we will show the final comparisons of different kinds of 3D face reconstructed algorithms and analyze the inner differences. The rest of this paper is organized as follows. In Section 2, we introduce the 3D face datasets and compared algorithms used in this paper. Then a detailed description of the CNN we construct is shown in Section 3 including the training processes. Finally, we show the comparison results among different algorithms and datasets in Section 4.

2. Datasets
In this part, we will introduce the datasets we use in our experiments. Considering the fairness of evaluation, we choose two different 3D face datasets (the size of a dataset, type of camera, expression variation and illumination variation) to show the comparisons.

2.1. Texas 3D Face Recognition Database (Texas 3DFRD)
Texas 3DFRD contains 1149 pairs of 2D and 3D faces from 105 adults. It was built by a stereo imaging system manufactured by 3Q Technologies (Atlanta, GA) at a very high spatial resolution of 0.32 mm along the x, y, and z dimensions. Texas 3DFRD has no expression variations and illumination variations. Besides, the number of per subject ranges from one to twenty inequality.

2.2. ND-2006 Data Set
The Fall2004range branch of the ND-2006 data set collected a total of 5152 images from 345 distinct subjects. Compared with Texas 3DFRD, it was captured by a range imaging system -the Minolta camera. So, the 2D and 3D could not be captured at the same time, which may result in texture wrong registration or mesh distortion. Every subject contains 6 different types of expressions (Neutral, Happiness, Sadness, Surprise, Disgust and Other) and 63 images which are more general compared with current other 3D face datasets.

2.3. AFLW2000-3D
This dataset was introduced in Ref. [6] firstly for 3D face alignment. It contains 68 landmarks of 3D ground truth from the first 2,000 AFLW [17] samples. Although the 3D ground truth in this dataset is labeled manually, it contributed much to the 3D face alignment community in the respect of evaluation.
The following three 3D face reconstruction algorithms all have trained and compared their works on AFLW2000-3D.

### 3. Tested 3D Face Reconstruction Algorithms

To explore the potentials of two manifold 3D face reconstruction algorithms, we choose three popular examples in 3DMM based and end-to-end methods. Meanwhile, we show the 3D reconstruction face datasets processing procedure.

#### 3.1. 3DDFA

3DDFA reconstructs 3D face by a two-stream cascaded CNN (figure 1). One of the inputs is a novel Projected Normalized Coordinate Code (PNCC) stacked with the input image. The other input is a feature anchor image with consistent semantics. In addition to this, the paper adopted face profiles to generate a new 3D face dataset to train a large pose dense 3D face alignment task. Finally, 3DDFA showed state-of-the-art performance on AFLW2000-3D at that time.

![Figure 1. The architecture of neural network in 3DFFA and a sample in Texas 3DFRD.](image)

#### 3.2. VRN

This is the first network to model 3D facial geometry by volumetric structure (figure 2). This approach lets CNN learn directly, in an end-to-end manner, the mapping from image pixels to the full 3D face. Three different CNN structures based on hourglass network were explored to discuss the effectiveness. Finally, the VRN-Guided was recommended as the best structure. Its experiment results verify that the novel presentation can work positively at a cost of computation to output a 192 x 192 x 200 volume.

![Figure 2. The architecture of neural network in VRN and samples in Texas 3DFRD](image)

#### 3.3. PRNet

PRNet also designed a new presentation for 3D faces. Different from VRN, it recorded the 3D shape of a complete face in UV space and keep semantic information at each UV place. A simple encoder-decoder network with a weighted loss was trained to regress the UV position map from a single 2D facial image. The three phases of PRNet is shown in figure 3. Compared with other algorithms, it not only performed well on the accuracy, but also the runtime with 9.8 ms per image.

![Figure 3. The architecture of neural network in PRNet](image)
4. Method

4.1 The Architecture of CNN

Hourglass was proposed in Ref. [18] to improve the bottleneck of resnet [12]. In our experiments, we use a two-stacked hourglass network as a feature extractor. Then a 26-layer residual module followed as a classification block where the number of channels is changed manually according to different tasks.

In dataset discrimination classification, we use the same number of parameters in the two-stacked hourglass as identity clarification and decrease the number of parameters in residual block to half of the identity clarification. In figure 4, we show the architecture of identity CNN and the dataset discrimination CNN should be modified as mentioned. In both tasks, we calculate the loss between prediction results and ground truth labels by cross-entropy function, the equation is as follow:

\[
\text{loss}(x, \text{class}) = -\log \left( \frac{\exp(x[\text{class}])}{\sum_j \exp(x[j])} \right) = -x[\text{class}] + \log \left( \sum_j \exp(x[j]) \right)
\]

where class is target label and \(x\) is the probability of prediction results on every class.

For dataset discrimination classification, we train the CNN by eight datasets (six 3D construction face datasets by three algorithms from two datasets and another two true depth datasets); for dataset discrimination classification, we train the CNN by six datasets (every three 3D construction face datasets from a same 3D dataset combid with the corresponding true depth dataset separately).

4.2. Training

In all the experiments, we trained two kinds of CNN (identity classification CNN, dataset discrimination CNN) for 14 experiments: 8 identity classification and 6 dataset discrimination tasks. In ND2006 (Fall2004range branch), it includes 4716 images of 345 subjects. In Texas, it has 1149 images from 116 subjects. We split the datasets separately for training and test as a ratio of 7:3. For all the training phases, we used the same learning rate set. The initialization learning rate is \(10^{-3}\) with a minibatch of 10. After 10 epochs, we decreased the learning rate to \(10^{-4}\) and to \(10^{-5}\) after another 10 epochs. In total, we trained 50 epochs for every experiment. All CNNs are trained on Pytorch [19] and use rmsprop [20].

5. Results and Discussion

5.1. Results

We show the accuracy results of two classification tasks on test set of Texas 3DFRD (375 images) and ND-2006 Data Set (1400 images) separately (figures 5-8). A summary of all results in table 1.

We can see that in the identity classification task the accuracy results on Texas 3DFRD are higher than on ND-2006. That is for the unbalance in Texas 3DFRD leading to insufficient for some subjects. Meanwhile, the accuracy results of 3D reconstruction algorithms are lower than the true depth datasets. It illustrates the effective features in true depth face is richer than 3D reconstruction face. In dataset discrimination tasks, Texas 3DFRD is closer to 3DDFA, while the other two are hard to classify from ND2006. It shows every algorithm has its distinct feature space.
Figure 4. The architecture of neural network identity CNN.
Figure 5. The identity recognition results on Texas 3DFRD.

Figure 6. The identity recognition results on ND2006.

Figure 7. The datasets discrimination results on Texas 3DFRD.

Figure 8. The datasets discrimination results on ND2006.

Table 1. A comparison of three algorithms on two datasets.

| Dataset      | Texas 3DFRD | ND-2006 Data Set |
|--------------|-------------|-----------------|
| Task         | Id-cls     | Id-cls          |
| 3DFFA        | 0.232       | 0.555           |
| PRNet        | 0.336       | 0.657           |
| VRN          | 0.456       | 0.677           |
| True depth   | **0.501**   | **0.721**       |

5.2. Discussion

For the identity recognition task, depth images are challenging. So, we use it to evaluate intra differences of reconstructed datasets. Results show that on this aspect, end-to-end algorithms perform better. It accounts for the template based algorithm has similarity in feature space. For the dataset discrimination task, it is used to evaluate the distance between the entire reconstructed dataset and the true depth dataset. Through these experiments, we found the distances change with different algorithms and types of 3D cameras. The 3D reconstruction faces from 3DDFA are smoother, so it is closer to Texas 3DFRD; similarly, the 3D reconstruction faces from end-to-end algorithms have their texture which makes them closer to ND2006. However, end-to-end 3D reconstruction algorithms may produce failure images as figure 9.
6. Conclusion
In this work, we evaluate three 3D face reconstruction algorithms on two datasets. Compared with other metrics, we use a CNN methodology to design two different tasks for these datasets. For the current two manifold reconstruction methods, we explore the feature distances intra and inter datasets. We train and test our CNNs based on two different kinds of true depth 3D datasets. Experiments demonstrate the template-based algorithm sacrifices the diversity of spatial features while guaranteeing the integrity of 3D faces. Inversely, end-to-end algorithms may fail in reconstructing 3D faces for some subjects, but they keep the differences between individuals.

References
[1] Zhu X, Liu X, Lei Z, et al. 2017 Face alignment in full pose range: A 3D total solution IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (1) 78-92.
[2] Roth J, Tong Y and Liu X 2016 Adaptive 3D face reconstruction from unconstrained photo collections Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4197-4206.
[3] Zhu X and Ramanan D 2012 Face detection, pose estimation, and landmark localization in the wild IEEE Conference on Computer Vision and Pattern Recognition pp 2879-2886.
[4] Zhu X, Lei Z, Yan J, et al. 2015 High-fidelity pose and expression normalization for face recognition in the wild Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 787-796.
[5] Zhu X, Lei Z, Liu X, et al. 2016 Face alignment across large poses: A 3D solution Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 146-155.
[6] Bulat A and Tzimiropoulos G 2017 How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks) Proceedings of the IEEE International Conference on Computer Vision pp 1021-1030.
[7] Jackson A S, Bulat A, Argyriou V, et al. 2017 Large pose 3D face reconstruction from a single image via direct volumetric CNN regression Proceedings of the IEEE International Conference on Computer Vision pp 1031-1039.
[8] Feng Y, Wu F, Shao X, et al. 2018 Joint 3D face reconstruction and dense alignment with position map regression network Proceedings of the European Conference on Computer Vision (ECCV) pp 534-551.
[9] Le V, Hu Y and Huang T S 2009 A quantitative evaluation for 3d face reconstruction algorithms 2009 IEEE International Conference on Acoustics, Speech and Signal Processing pp 1269-1272.
[10] Feng Z H, Huber P, Kittler J, et al. 2018 Evaluation of dense 3D reconstruction from 2D face images in the wild 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) pp 780-786.
[11] Bulat A and Tzimiropoulos G 2017 Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources Proceedings of the IEEE International Conference on Computer Vision pp 3706-3714.
[12] He K, Zhang X, Ren S, et al. 2016 Identity mappings in deep residual networks European Conference on Computer Vision (Cham: Springer) pp 630-645.
[13] Gupta S, Markey M K and Bovik A C 2010 Anthropometric 3D face recognition International Journal of Computer Vision 90 (3) 331-349.
[14] Gupta S, Castleman K R, Markey M K, et al. 2010 Texas 3D face recognition database 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI) pp 97-100.
[15] Gupta S, Castleman K R, Markey M K and Bovik A C Texas 3D Face Recognition Database http://live.ece.utexas.edu/research/texas3dfr/index.htm.
[16] Faltemier T C, Bowyer K W and Flynn P J 2007 Using a multi-instance enrolment representation to improve 3D face recognition 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems pp 1-6.
[17] Koestinger M, Wohlhart P, Roth P M, et al. 2011 Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) pp 2144-2151.
[18] Newell A, Yang K and Deng J 2016 Stacked hourglass networks for human pose estimation European Conference on Computer Vision (Cham: Springer) pp 483-499.
[19] Paszke A, Gross S, Chintala S, et al. 2017 Automatic differentiation in pytorch.
[20] Graves A 2013 Generating sequences with recurrent neural networks arXiv preprint arXiv:1308.0850.