ESP Journal of Engineering & Technology Advancements (ESP-JETA)
ISSN : 2583-2646
Special Issue

DISTORTED FACE RECONSTRUCTION USING 3D CNN

Mrs.S.Yoheswari
Assistant Professor, Department of Computer Science and Engineering
K.L.N. College of Engineering
(Anna University)
Sivagangai, India

M Nirmal kumar
Student, Department of Computer Science and Engineering
K.L.N. College of Engineering
(Anna University)
Sivagangai, India
akashrmailbox@gmail.com

A Mohammed Thalha
Student, Department of Computer Science and Engineering
K.L.N. College of Engineering
(Anna University)
Sivagangai, India

N Maathesh
Student, Department of Computer Science and Engineering
K.L.N. College of Engineering
(Anna University)
Sivagangai, India

***ABSTRACT--** This paper presents a learning-based method for detailed 3D face reconstruction from a single unconstrained image. The core of our method is an end-to-end multi-task network architecture. The purpose of the proposed network is to predict a geometric representation of 3D face from a given facial image. Unlike most existing reconstruction methods using low-dimension morphable models, we propose a pixel based multi-scale representation of a detailed 3D face to ensure that our reconstruction results are not limited by the expressiveness of linear models. We break the task of high fidelity face reconstruction into three subtasks, which are face region segmentation, coarse-scale reconstruction and detail recovery. So the end-to-end network is constructed as a multi task mode, which contains three subtask networks to deal with different subtasks respectively. A backbone network with feature pyramid structure is proposed as well to provide different levels of feature maps required by the three subtask networks. We train our end-to-end network in the spirit of the recent photo-realistic data generation approach. The experimental results demonstrate that our method can work with totally unconstrained images and produce high-quality reconstruction but with less runtime compared to the state-of the-art.*

I. INTRODUCTION

3D face reconstruction is a key problem in computer vision and graphics due to its wide range of applications. Compared with the reconstruction methods relying on complex experimental equipment or user intervention by 3D scanner and synchronized multi-camera stereo setups, unconstrained images captured in arbitrary recording conditions, for example, photos taken by mobile phones or downloaded from Internet, provide people unprecedented easy access to create their own 3D digital faces. However, there is also a gap in the reconstruction quality between high-end and commodity setup. In recent years, a lot of progresses have been made in improving the accuracy and efficiency of reconstruction from a single image. Since the task is highly ill-posed, priors thatmodel the structure and expression of faces are commonly employed, such as 3D Morphable Models(3DMM).

While simplifying the reconstruction problem, these solutions introduce some inherent limitations. The reconstruction results are limited to the linear space due to the linear expression ability of 3DMM. Recently, deep learning has shown promising performance in reconstructing 3D face from single image. The deep learning based methods translate an input face image to a representation of 3D face. Thus, most existing works using low-dimension morphable models like 3DMM to encode the 3D face were still limited by the expressiveness of the linear models. Some works proposed a second network to refine the coarse reconstruction, due to the lack of training dataset with detailed face geometry, unsupervised trainings are adopted which may affect the accuracy of reconstruction. Difficulty in obtaining a large-scale detailed 3D face dataset for training is one of the main challenges for learning-based methods. Some works synthesize such data with randomized parametric face model parameters, which are not photo realistic and unsuitable for high-fidelity 3D face reconstruction. Some works utilize multi-view stereo setups to capture a certain number of 3D scans from different subjects under several expressions, which are still insufficient to cover a wide range of conditions. Despite so many progresses, it remains challenging for reconstructing high-fidelity face with geometric details from one unconstrained image, especially in real-time, which is of critical importance for VR, AR, social media and communication, etc. Under this setting, we propose a learning-based method in this paper, including an end-to-end network framework. The task of high-fidelity face reconstruction is broke down into three subtasks, which are face region segmentation, coarse-scale reconstruction and detail recovery. Three networks are employed to deal with the three subtasks. For an end-to-end network architecture, a backbone network is constructed to provide different levels of feature maps required by the three subtask networks, ensuring the independence of features used by different sub networks, while minimizing the time loss caused by multiple tasks.

II .Related work

As mentioned before, image-based face reconstruction attracts a lot of attentions. This section only reviews the closely related learning-based works for conciseness. The most common 3D face model is 3DMM proposed by Blanz and Vetter[1] which provides a low-dimensional representation of textured 3D face with principle components analysis. 3DMM has been widely used in learning-based reconstruction from single image[2][3][4][5][6].They utilize neural networks to learn parameters of 3DMM as well as pose of face. While providing robustness, 3DMM can express only coarse geometries and these methods based on 3DMM representation are still limited by the expressiveness of the linear model. Richardson et al.[4][5] and Guo et al.[6] extend above methods by using a Coarse-to-Fine framework. Richardson et al. utilize Shape-From-Shading(SFS) algorithm to add details [4]or learning detailed geometry in unsupervised manner[5]. Guo et al.[6] train FineNet with the constructed fine-detailed face image dataset. Jackson et al.[7] propose a CNN architecture that performs direct regression of a volumetric representation of 3D face geometry from a 2D image, but only coarse-scale face geometry can be reconstructed. Sela et al.[8] utilize an Image-to-Image translation network that maps the input image to a depth image and a facial correspondence map. To bring the results into full vertex correspondence, an iterative non-rigid registration process deforms the transformed template, which takes a few minutes to converge. Yamaguchi et al.[9] and Chen et al.[10] utilize Light Stage X[11] and multi-stereo setup respectively to create high-quality training database. The common limitation of the two methods is that they cannot tackle low resolution images. Some recent work[12][13] follow this line of research for improving the quality of reconstruction..

III.PROPOSED METHODOLOGY

Multi-scale face geometric representation –

The purpose of the proposed end-to-end network is to predict a geometric representation from a given facial image. So aproper

representation is crucial as it would affect the overall performance of the whole framework. We propose a multi scale face model to encode a high-fidelity 3D face, as illustrated in Figure1.The representation is composed of three maps: depth map, displacement map and face mask. We use depth map and pixel depth displacement map to represent coarse-scale face geometry and fine- scale geometric details respectively. The pixel- based geometric representation is not restricted by any model. In addition, compared with 3DMM which encode the face region only, our pixel- based representation needs a face mask to separate the face region from the background

End-to-end multi-task network

The core of our framework is an end-to-end multi-task network. Figure 2 shows the architecture of the proposed framework, which is composed of a feature pyramid backbone network, a face region segmentation sub-network, a coarse-scale face reconstruction sub-network and a face detail recovery sub-network. Fig.2 The end-to-end multi-task network 3.2.1 Backbone network While face region segmentation requires high-level semantics, coarse-scale reconstruction and detail recovery not only require high-level semantics but also high resolution feature maps. However, the feature hierarchy of deep convolutional network has an inherent multiscale, pyramidal shape, which produces feature maps of different spatial resolutions, but large semantic gaps. To solve this problem, we adopt the Feature Pyramid Network(FPN) architecture[14], as illustrated in Figure 2. Feature pyramid is a basic component in recognition system for detecting object at different scales. FPN combines low- resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. Specifically, we utilize ResNet18[15] composed of 5 convolutional blocks. The resolution of the feature map is reduced by half undergoing a block. So, the resolution is reduced to 1/32 of origin image after 5 blocks.. Then, a feature pyramid architecture from the deepest layer of the ResNet18 is built via a top-down pathway. Considering the efficiency of the algorithm, we choose the feature maps with 1/8, /1/4, 1/2 resolution of origin image in feature pyramid architecture respectively to feed the face region segmentation, coarse-scale reconstruction and detail recovery sub-task networks. The backbone network of feature pyramid structure ensures that

each subtask extracts corresponding hierarchical feature maps from different layers, while ensuring the strong semantics required by each subtask.

Face region segmentation network

The context information is important for semantic segmentation task. In deep neural network, the size of receptive field can roughly indicate how much context information we use. Zhou et al.[16] have proved that the actual size of the receptive field is much smaller than the theoretical size, especially in the later layers. As a result, the neural network cannot fully utilize the global information. To reduce context information loss, we approach this problem using pyramid pooling module [17] from scene parsing for our face region segmentation. The pyramid pooling module exploits the capability of global context information by different-region based context aggregation, which has been proved to be an effective global contextual prior. Specifically, we utilize the feature map with 1/8 resolution of origin image from backbone network to feed the pyramid pooling module. Using 4-level pyramid, the kernel sizes are 32 ×32,16×16,8×8 and 4×4 , strides are 32, 16, 8 and 4 respectively. We use cross-entropy loss function to train our face region segmentation network: $\hat{} = -[y \log y + (1-y) \log (1-y)]$ (1) Where y denotes predictive value and y denotes ground truth. Comparison results of face region segmentation are shown in Figure 3. It can be seen that the segmentation effects have been significantly improved in some edge regions or regions that are difficult to judge, such as boundaries between Them.

h of our coarse-scale reconstruction and detail recovery networks adopt an encoding-decoding architecture, inspired by the U-Net based architecture[18]. The encoder and decoder respectively contain several Convolution BatchNorm-ReLU layers. Passing through a decoding layer, the size of feature map is reduced to half of the origin, while increased to twice after passing a decoding layer. Due to different resolutions of the input feature maps of coarse-scale and detail reconstruction networks (1/4 and 1/2 of origin respectively), the numbers of layers are slightly different. Specifically, the encoder and decoder architectures of the coarse-scale reconstruction network consist of 6 Conv-BN ReLU layers respectively. The encoder and

decoder of detail recovery network consist of 7 Conv-BN-ReLU layers respectively. Similar to [18], we add skip connection between each layer i and $n - i$, where n is the total number of layers, which would shuttle the great deal of the low-level information directly across the net.

$$\text{Smooth}(|y-y|) = \text{Smooth}(x) = 0.5x \; if|x|$$

IV EXPERIMENTS

In order to evaluate the performance of the end-to- end network framework, we perform several experiments on both face dataset and unconstrained image. Figure 6 shows the comparison of reconstruction results with the recent learning based methods of [5] and [6].It can be seen that our results are convincing in both the global geometry and the fine details. Especially, our reconstructions of the nose are more convincing than [5] and [6]. The reason is they both utilize 3DMM to represent coarse geometry. So face features and contours are limited by the expressiveness of the linear model. We overcome this limitation by learning directly in the image domain. In addition, our wrinkle detail recoveries are better than [5], comparable to [6]. There are two main reasons. 1)[5] use randomly rendered synthetic data for training in the coarse reconstruction network, we and [6] render the photo realistic face data based on the real face image, which facilitate the training. 2)In the detail recovery network, we and [6] adopt supervised learning using the rendered face images with rich details as training data, while [5] use unsupervised learning. In terms of time efficiency, [5] requires multiple iterative calculations results in non-real time. Under the same hardware condition with [6](Intel quad core i7 CPU, 4GB RAM, NVIDIA GTX 1070 GPU), the size of our input image is 256*256*3, runtime is 15ms. While the input image size of [6] is 224*224*3, and runtime is 20ms. It can be seen that our method can produce high-quality

V. CONCLUSION

We propose a learning-based method for detailed 3D face reconstruction from a single image. Our method employs an end-to-end multi-task network which mapping the input image to pixel-based multi-scale facial geometric representation. As demonstrated in experiments, the proposed framework performs well both in reconstruction quality and efficiency.

VI References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces [A]. Proceedings of the 26th annual conference on Computer graphics and interactive techniques [C]. 1999, 187-194.

[2] A. T. Tran, T. Hassner, I. Masi and G. Medioni. Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2017, 1493-1502.

[3] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis and S. Zafeiriou. 3D Face Morphable Models "In-the Wild" [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2017, 5464-5473.

[4] E. Richardson, M. Sela and R. Kimmel. 3D Face Reconstruction by Learning from Synthetic Data [A]. 2016 Fourth International Conference on 3D Vision (3DV) [C]. 2016, 460-469.

[5] E. Richardson, M. Sela, R. Or-El and R. Kimmel. Learning Detailed Face Reconstruction from a Single Image [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2017, 5553-5562.

[6] Y. Guo, J. Zhang, J. Cai, B. Jiang and J. Zheng. CNN-based Real-time Dense Face Reconstruction with Inverse-rendered Photo realistic Face Images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6):1294-1307.

[7] A. S. Jackson, A. Bulat, V. Argyriou and G. Tzimiropoulos. Large Pose 3D Face

Reconstruction from a Single Image via Direct Volumetric CNN Regression [A]. 2017 IEEE International Conference on Computer Vision (ICCV) [C]. 2017, 1031-1039.

[8] M. Sela, E. Richardson and R. Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation [A]. 2017 IEEE International Conference on Computer Vision (ICCV) [C]. 2017, 1585-1594.

[9] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima and H. Li. High-fidelity facial reflectance and geometry inference from an unconstrained image[J]. ACM Trans. Graph., 2018, 37(4), 162:1-162:14.

[10] A. Chen, Z. Chen, G. Zhang, Z. Zhang, K. Mitchell and J. Yu. Photo-realistic facial details synthesis from single image[A]. IEEE International Conference on Computer Vision (ICCV)[C], 2019, 9428-9438.

[11] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu and P. Debevec. Multiview face capture using polarized spherical gradient illumination [J]. ACM Trans. Graph., 2011, 30(6): 1-10.

[12] A. Lattas et al., AvatarMe: Realistically Renderable 3D Facial Reconstruction "In-the-Wild"[A], 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C], Seattle, WA, USA, 2020, pp. 757-766.

[13] H. Yang et al., FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction[A], 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C], Seattle, WA, USA, 2020, pp. 598-607.

[14] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie. Feature Pyramid Networks for Object Detection [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2017, 936-944.

[15] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition [A]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2016, 770-778.

[16] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva and A. Torralba. Object detectors emerge in Deep Scene CNNs [A]. International Conference on Learning Representations [C]. 2015, 1-12.

[17] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia. Pyramid Scene Parsing Network [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2017, 6230-6239.

[18]P. Isola, J. Zhu, T. Zhou and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. 2017, 5967-5976.

[19] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks [A]. 2015 IEEE International Conference on Computer Vision (ICCV) [C]. 2015, 2758-2766.

[20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou and M. Pantic. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge [A]. 2013 IEEE International Conference on Computer Vision Workshops [C]. 2013, 397-403.