Journal of Advanced College of Engineering and Management, Vol. 9, 2024

3D Face Reconstruction from Occluded Images

Sudip Rana
Lecturer, Department of Electronics and Computer Engineering, IOE,
Thapathali Campus
Email: sudip.rana@tcioe.edu.np

Abstract

Current methods for reconstructing 3D faces from regular images have some problems. They struggle to create realistic animated faces because they don't account for how wrinkles change with different expressions. They also have trouble working with images taken in real-world conditions as images are likely to be occluded and exposed to extreme condition. A new approach is implemented that can predict the shape of a face in 3D which has been blocked by different factor like hand, eye glass, mask etc. Context based learning knowledge distillation is used to transfer the knowledge from main DECA model to learner model. The learner model is also trained to handle different occlusion thus helping in constructing 3D faces. This is done from normal pictures without needing special 3D information and it works really well, producing accurate results. It achieves state-of-the-art shape reconstruction accuracy on NoW benchmarks.

Keywords*: 2D to 3D model, 3D Face Reconstruction, Knowledge Distillation, Monocular Image, Occlusion.*

1. Introduction

Three-dimensional (3D) face reconstruction from monocular images is a key focus in computer vision, with applications ranging from virtual reality to facial recognition. Despite significant advancements, current techniques still struggle with capturing intricate geometric details, particularly in handling facial expressions and occlusions. Two primary approaches have been developed [1]: one prioritizes high-detail 3D models for accurate recognition, but these models falter when the face is partially covered or the image is challenging. The other approach emphasizes robustness across varying conditions, such as angled views or hidden facial features, but sacrifices some level of detail.

Over the past two decades, advances have leveraged pre-computed 3D face models and neural network-based approaches [3], including Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs). However, these methods often fail with occluded or uniquely illuminated images, sometimes producing unrealistic "ghost-like" reconstructions. A persistent challenge remains in accurately reconstructing fine details like expression-dependent wrinkles, which are crucial for realistic emotion depiction.

Addressing occlusions and refining the balance between flexibility and detail in 3D reconstructions is an ongoing research challenge. Future directions include developing models that better handle occlusions, possibly through context-based learning, to improve the realism and fidelity of reconstructed facial models.

2. Literature Review

Existing 3D face reconstruction methods often struggle to animate faces realistically due to challenges in capturing expression-dependent wrinkles and adapting to real-world conditions. DECA [1] improves on this by using a low-dimensional representation to predict face details, including wrinkles, from a single image, achieving state-of-the-art accuracy. However, it struggles with low-resolution images and lacks exposure to diverse conditions like varying illumination and occlusions.

V. Blanz’s et. al [2], method focuses on face recognition under different angles and lighting but falters with occlusions. [3] Technique excels in capturing fine details during complex expressions but also faces challenges with realism.

Deep learning methods for 3D face reconstruction often struggle due to limited training data with accurate 3D shape information. Y. Deng et al. [4] address this with a hybrid loss function and multi-image reconstruction, improving accuracy and handling occlusions. Z.-H. Feng et al. [5] focus on evaluating reconstruction accuracy using a benchmark dataset but lack occluded images.

Ensemble learning, where multiple models are trained and averaged, enhances machine learning performance but is computationally intensive. Caruana et al. [6] and Hinton et al. [3] propose compressing ensemble knowledge into a single model, improving efficiency. Hinton’s approach also introduces a novel ensemble of full and specialist models for better accuracy. In 3D face reconstruction, Tiwari et al. [7] address challenges with occlusions using a context-learning-based distillation technique. Their model, trained on occluded images, significantly improves landmark accuracy but struggles with capturing fine geometric details like wrinkles and moles. This research advances 3D facial reconstruction, particularly in occluded scenarios.

Estimating 3D face shape from a single image is challenging due to factors like lighting, pose, and occlusions. Sanyal et al. [8] introduced RingNet, which infers 3D shape by leveraging multiple images and detected facial features, using a loss function that ensures consistency across images of the same person. RingNet, evaluated with the new NoW dataset, outperforms methods needing 3D supervision but lacks 2D ear detection and full body reconstruction. Kao et al.'s [9] Perspective Network (PerspNet) tackles challenges of perspective projection by estimating 3D shapes and 6 Degrees of Freedom (6DoF) pose, showing significant improvements in accuracy and application for VR/AR.

3. Methodology

3.1 Dataset preparation

For the training of both the learner and teacher (DECA) models, the CelebA dataset released by Tensorflow [10] has been utilized. The choice of this balanced dataset, encompassing around 181,000 images, has been made to serve as the training data for the models. A subset, constituting approximately a quarter of the total images, has been extracted from this extensive dataset.

3.2 Augmentation and Validation

Building on H. Tiwari et al.'s [7] approach, a new dataset of 45,000 images with various occlusions was created by introducing arbitrary pixel values to face images. This dataset helps validate models by comparing landmarks between clear and occluded images, revealing discrepancies from arbitrary pixels. The dataset's diverse occlusion scenarios aim to enhance model robustness in handling real-world occluded images. It prevents model bias by using global context rather than local pixel values. During training, the DECA model processes clear images, while the Learner model works with occluded versions of the same images, improving adaptability to occlusions.

Figure 1: Sample dataset with clear image and occluded image of same person

3.3 Training Network

The Learner model is founded on context-based learning knowledge distillation. The training framework for the Learner model is illustrated in the figure below. In this framework, the DECA model functions as the teacher model, while the Learner model serves as the student model for the process of knowledge distillation.

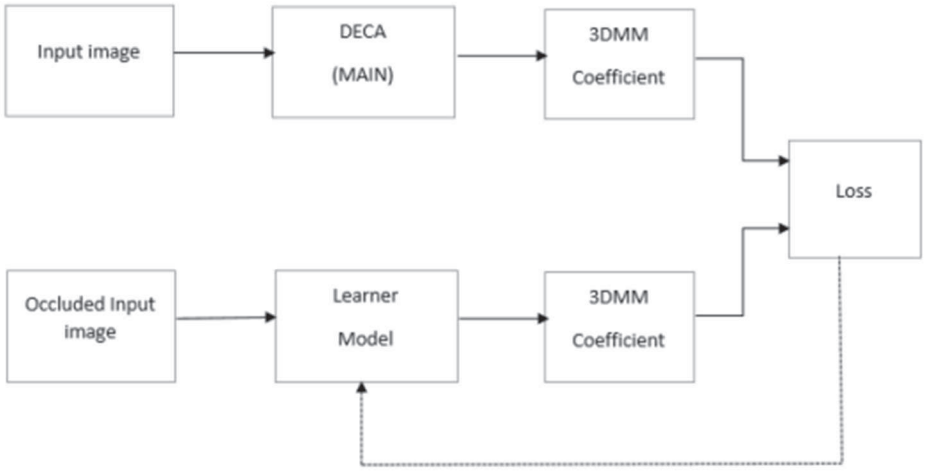

Figure 2: Training Network

DECA [1] is a pre-trained model designed to generate UV displacement maps using a compact latent representation, capturing individual-specific details and common expressions. It forecasts various parameters such as details, shape, color, expressions, posture, and lighting from a single image, and is trained on in-the-wild images without paired 3D supervision using datasets like VGGFace2, BUPT-Balancedface, and VoxCeleb2. The ResNet-50 model serves as DECA's backbone, featuring a fully-connected layer with 236 nodes for coarse reconstruction and 128 nodes for detailed reconstruction. The DECA model's weights are frozen during Learner model training.

The determination of the output connection of the ResNet model is as follows:

$$v_u = \text{ResNet}(I_u)$$

The Learner model, utilizing ResNet-50, processes occluded or unoccluded input images to generate coefficients for a 3D Morphable Model (3DMM) using the FLAME model. The ResNet-50's classification layer is expanded to 257 nodes to match the 3DMM coefficients. Adjustments have been made to the fully-connected classification layer of the ResNet-50 by expanding it to encompass 257

nodes, corresponding to the coefficients represented as v ∈ R^{257}. The equation below expresses the relationship:

$$v(\theta) = (I, \theta)$$

where θ signifies the network weight.

The training involves knowledge distillation, where the Learner model aims to replicate the DECA model's coefficients for occluded images, using Mean Squared Error (MSE) loss to minimize differences between predicted coefficients. This approach enhances the Learner model's robustness and contextual learning capabilities, addressing occlusion challenges effectively.

The MSE loss, shown in below equation, quantifies this comparison. The equation is defined as follows:

$$L_{SE} = \frac{1}{M}\sum_{m=1}^{M} {c_m}^2$$

where M represents the number of elements in the vector c, and in this particular case, M equals 257. Also c = (v_u − v) The symbols v_u and v stand for the labels predicted by the trainer and the learner models, respectively.

3.4 System Block Diagram

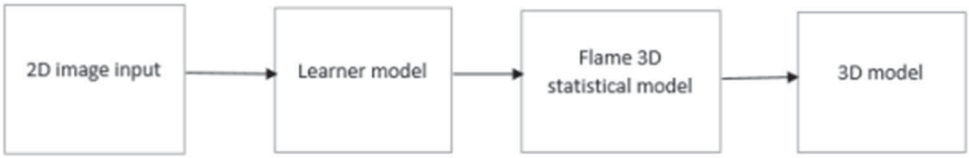

Figure 3: System Block Diagram

The model accepts a 2D input image with dimensions of 224 x 224 pixels. Upon training, the Learner model is capable of processing a single image—whether occluded or clear—and generating a 3D model with intricate facial geometry. This model captures fine details such as wrinkles, contours, and texture. The detailed 3D face model within the Learner enables animation by adjusting parameters related to facial expressions and jaw pose, as described in [18]. This statistical model facilitates the creation of a 3D face from features extracted from the 2D input image. It not only reconstructs a highly detailed 3D face but also allows for the manipulation of facial expressions to produce realistic animations, preserving the individual's unique facial characteristics.

4. Result and Discussion

4.1 Qualitative comparisons

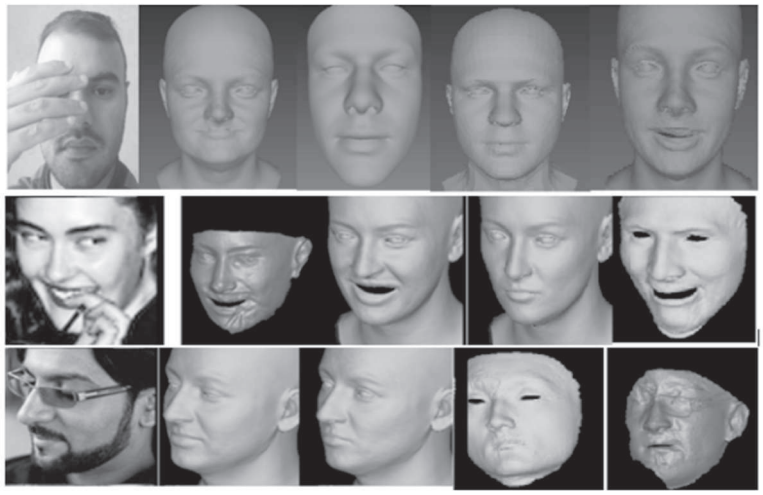

Figure 4: a) Original Occluded Image b) Output from DECA c) Learner model d) 3DDFA V2 e) MGC Net

As indicated in the preceding comparison, the presence of occlusion, particularly involving a hand, is observed in the image. DECA [1] and other conventional methods demonstrate a limited capacity to effectively manage occlusions within the image. Consequently, these methods attempt to infer and complete the occluded portions, leading to less convincing results. In contrast, the Learner model is specifically trained to address occlusions. As a result, it leverages the acquired knowledge during training to accurately fill the occluded regions, yielding more compelling and reliable outcomes.

4.2 Quantitative Evaluation

The NoW challenge, established in the 2019 study by Sanyal et al. [8], serves as a benchmark task for evaluating 3D face reconstruction algorithms. This dataset comprises 2,054 face images from 100 distinct individuals, organized into a validation set with 20 individuals and a test set with 80 individuals. Each individual has an associated reference 3D face scan.

The dataset encompasses a diverse array of conditions, including indoor and outdoor environments, various facial expressions (both neutral and expressive), partially obscured faces, and multiple viewing angles ranging from frontal to side profile. To assess algorithm performance, a standardized evaluation process is employed, which involves computing the distance between all vertices in the reference 3D scans and their corresponding points on the reconstructed 3D mesh surface.

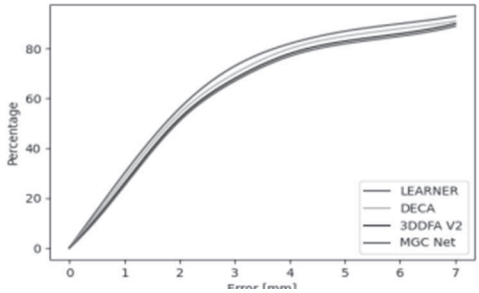

Figure 5: Cumulative Error Plot

Table 1: Reconstruction error on the NoW [Sanyal et al. 2019] benchmark.

Method	Median (mm)	Mean (mm)	Std (mm)
3DDFA V2	1.15	1.28	1.26
MGC Net	1.12	1.27	1.23
DECA	1.10	1.26	1.20
LEARNER	1.08	1.20	1.18

5. Conclusion

In conclusion, this work has delved into the intricate realm of 3D face reconstruction from occluded images, addressing the formidable challenges posed by partial or full obstructions in facial data. The exploration encompassed diverse methodologies, including context-learning-based distillation approaches, the synthesis of artificial occlusion-based datasets, and the utilization of deep learning for accurate facial geometry extraction. Motivated by the need for robustness against occlusions, the proposed models exhibited promising results, showcasing advancements in landmark accuracy and robust network training. As we move forward, the strategies developed in this work pave the way for enhanced 3D face reconstruction methodologies, offering valuable contributions to the broader landscape of computer vision and biometrics.

6. Suggestion And Recommendation

While the model has demonstrated success in effectively addressing occlusion within input 2D images, there remains room for enhancement, particularly in handling extreme poses and expressions. Notably, the current work does not explicitly incorporate a model for facial hair, causing skin tone to influence the lighting model and attributing facial hair effects to shape deformations. To further refine the model's capabilities, the inclusion of more diverse datasets featuring in-the-wild images can expose it to a broader range of variations during training. However, it's crucial to address challenges related to extreme occlusions, which are not currently handled optimally by the existing method. Additionally, the training dataset's incorporation of numerous low-resolution images contributes to robustness but may introduce undesired noise. Exploring strategies to mitigate these limitations will be essential for advancing the model's overall performance and applicability.

References

1. Y. Feng, H. Feng, M. J. Black, and T. Bolkart, “Learning an animatable detailed 3d face model from in-the-wild images,” ACM Transactions on Graphics (ToG), vol. 40, no. 4, pp. 1–13, 2021.

2. V. Blanz, S. Romdhani, and T. Vetter, “Face identification across different poses and illuminations with a 3d morphable model,” in Proceedings of fifth IEEE international conference on automatic face gesture recognition. IEEE, 2002, pp. 202–207.

3. G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.

4. Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, “Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0–0.

5. Z.-H. Feng, P. Huber, J. Kittler, P. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. R¨atsch, “Evaluation of dense 3d reconstruction from 2d face images in the wild,” in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 780–786.

6. R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, “Ensemble selection from libraries of models,” in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 18.

7. H. Tiwari, V. K. Kurmi, K. Venkatesh, and Y.-S. Chen, “Occlusion resistant network for 3d face reconstruction,” in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 813–822.

8. S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, “Learning to regress 3d face shape and expression from an image without 3d supervision,” in Proceedings 45 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7763–7772.

9. Y. Kao, B. Pan, M. Xu, J. Lyu, X. Zhu, Y. Chang, X. Li, and Z. Lei, “Towards 3d face reconstruction in perspective projection: Estimating 6dof face pose from monocular image,” IEEE Transactions on Image Processing, 2023.

10. Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3730–3738.