Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition

Feng Liu[1], Ronghang Zhu[1], Dan Zeng[1], Qijun Zhao[1] and Xiaoming Liu[2]

[1]College of Computer Science, Sichuan University [2]Department of Computer Science and Engineering, Michigan State University

2018 CVPR SALT LAKE CITY • JUNE 18-22

Introduction

Face recognition and 3D face reconstruction seem to **contradict** each other. On one hand, face recognition prefers identity-sensitive features, but not every detail on faces; on the other hand, **3**D reconstruction attempts to recover as much facial detail as possible, regardless whether the detail benefits or distracts facial identity recognition.

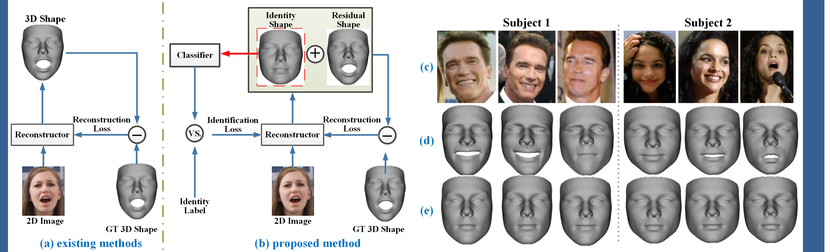

Figure: Comparison between the learning process of (a) existing methods and (b) our proposed method. GT denotes Ground Truth. (d) and (e) are **3**D face shapes and disentangled identity shapes reconstructed by our method for the images in (c) from LFW.

- We propose a method which for the **first time** explicitly optimizes face recognition and **3**D face reconstruction simultaneously. The method achieves state-of-the-art **3**D face reconstruction accuracy via joint discriminative feature learning and **3**D face reconstruction.
- We devise an effective training process for the proposed network that can disentangle identity and non-identity features in reconstructed **3**D face shapes. The network, while being pre-trained by **3**DMM-generated data, can surmount the limited **3**D shape space determined by the **3**DMM bases, in the sense that it better captures identity-sensitive and identity-irrelevant features in **3**D face shapes.
- We leverage the effectiveness of disentangled identity features in reconstructed **3**D face shapes for improving face recognition accuracy, as being demonstrated by our experimental results. This further expands the application scope of **3**D face reconstruction.

Proposed Method

Based on the assumption that **3**D face shapes are composed by identity-sensitive and identity-irrelevant parts, the **3**D face shape $\mathbf{s}$ of a subject is represented as

$$\mathbf{s} = \bar{\mathbf{s}} + \Delta\mathbf{s}_{Id} + \Delta\mathbf{s}_{Res}, \quad (1)$$

where $\bar{\mathbf{s}}$ is the mean **3**D face shape (computed across all training samples with neutral expression), $\Delta\mathbf{s}_{Id}$ is the identity-sensitive difference between $\mathbf{s}$ and $\bar{\mathbf{s}}$, and $\Delta\mathbf{s}_{Res}$ denotes the residual difference. A variety of sources could lead to the residual difference, for example, expression-induced deformations and temporary detail.

We further assume that $\Delta\mathbf{s}_{Id}$ and $\Delta\mathbf{s}_{Res}$ can be described by latent representations, $\mathbf{c}_{Id}$ and $\mathbf{c}_{Res}$, respectively. This is formulated by

$$\Delta\mathbf{s}_{Id} = f_{Id}(\mathbf{c}_{Id}; \theta_{Id}), \ \Delta\mathbf{s}_{Res} = f_{Res}(\mathbf{c}_{Res}; \theta_{Res}). \quad (2)$$

Here, f_{Id} (f_{Res}) is the mapping function that generates the corresponding shape component $\Delta\mathbf{s}_{Id}$ ($\Delta\mathbf{s}_{Res}$) from the latent representation, with parameters θ_{Id} (θ_{Res}).

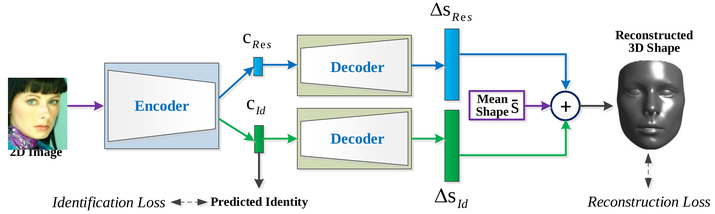

Figure: Overview of the proposed encoder-decoder based joint learning pipeline for face recognition and **3**D shape reconstruction.

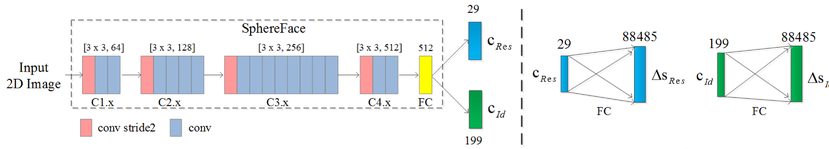

Figure: (a) Encoder in the proposed method is implemented based on SphereFace. It converts the input **2**D image to latent identity and residual shape feature representations. (b) Decoders in the proposed method are implemented as a fully connected (FC) layer. They convert the latent representations to corresponding shape components.

Implementation Detail

- Loss Functions. The overall loss to the proposed encoder-decoder network is defined by

$$L = \lambda_R L_R + L_C, \quad (3)$$

where λ_R is the Euclidean-based weight for the reconstruction loss L_R, L_C is the softmax loss.

- Training Data. To construct training data we develop a method for fitting **3**D morphable model (**3**DMM) to multiple **2**D images of a subject, and apply it to CASIA-WebFace database, resulting in **488**,**848** images / 3D faces of **10**,**575** subjects.
- Training Process. We train our encoder-decoder network in three phases. In Phase I, we train the encoder by setting the target latent representations as 3DMM coefficients and using Euclidean loss. In Phase II, we train the decoder for the identity and residual components separately. In Phase III, the end-to-end joint training is conducted based on the pre-trained encoder and decoder.

Experimental Results

Two sets of experiments have been done to evaluate the effectiveness of the proposed method in **3**D face reconstruction and face recognition. The MICC and BU**3**DFE databases are used for experiments of **3**D face reconstruction, and the LFW, YTF and IJB-A databases are used in face recognition experiments.

- **3**D Shape Reconstruction Accuracy.

Table: **3**D face reconstruction accuracy on the MICC database.

Method	VRN	3DDFA	3DMM-CNN	3DSR	Proposed
RMSE	**5.34**	**2.73**	**2.20**	**2.07**	**2.00**

Table: **3**D face reconstruction accuracy (RMSE) under different yaw angles on the BU**3**DFE database.

Method	±90°	±80°	±70°	±60°	±50°	±40°	±30°	±20°	±10°	0°	Avg.
VRN	**6.96**	**6.20**	**6.14**	**6.01**	**5.91**	**5.50**	**4.93**	**3.86**	**3.70**	**3.66**	**5.29**
3DDFA	**2.90**	**2.88**	**2.81**	**2.82**	**2.77**	**2.79**	**2.76**	**2.73**	**2.55**	**2.48**	**2.75**
3DMM-CNN	-	-	-	-	**2.30**	**2.26**	**2.23**	**2.22**	**2.19**	**2.17**	**2.23**
3DSR	**2.11**	**2.11**	**2.12**	**2.13**	**2.16**	**2.14**	**2.12**	**2.10**	**2.10**	**2.09**	**2.12**
Proposed	**2.09**	**2.04**	**2.03**	**2.03**	**2.00**	**1.99**	**2.03**	**2.01**	**1.97**	**1.93**	**2.01**

- Face Recognition Accuracy.

Table: Face recognition accuracy on the LFW and YTF databases.

Method	Shape	Texture	Accuracy	100%-EER	AUC	TAR-10%	TAR-1%
Labeled Faces in the Wild (LFW)							
3DMM	✓	×	**66.13** ± 2.79	**65.70** ± 2.81	**72.24** ± 2.75	**35.90** ± 3.74	**12.37** ± 4.81
	×	✓	**74.93** ± 1.14	**74.50** ± 1.21	**82.94** ± 1.14	**60.40** ± 3.15	**28.73** ± 7.17
	✓	✓	**75.25** ± 2.12	**74.73** ± 2.56	**83.21** ± 1.93	**59.40** ± 4.64	**29.67** ± 4.73
3DDFA	✓	×	**66.98** ± 2.56	**67.13** ± 1.90	**73.30** ± 2.49	**36.76** ± 6.27	**10.00** ± 3.22
	✓	×	**90.53** ± 1.34	**90.63** ± 1.61	**96.60** ± 0.79	**91.13** ± 2.62	**58.20** ± 12.14
3DMM-CNN	×	✓	**90.60** ± 1.07	**90.70** ± 1.17	**96.75** ± 0.59	**91.23** ± 2.42	**52.60** ± 8.14
	✓	✓	**92.35** ± 1.29	**92.33** ± 1.33	**97.71** ± 0.64	**94.20** ± 2.00	**65.57** ± 6.93
Proposed	✓	×	**94.43** ± 1.47	**94.40** ± 1.52	**98.12** ± 0.90	**95.07** ± 2.39	**74.54** ± 4.33
YouTube Faces (YTF)							
3DMM	✓	×	**73.26** ± 2.51	**73.08** ± 2.65	**80.41** ± 2.60	**51.36** ± 5.11	**24.04** ± 4.56
	×	✓	**77.34** ± 2.54	**76.96** ± 2.64	**85.32** ± 2.63	**63.16** ± 5.07	**31.36** ± 5.21
	✓	✓	**79.56** ± 2.08	**79.20** ± 2.07	**87.35** ± 1.92	**69.08** ± 5.00	**34.56** ± 6.89
3DDFA	✓	×	**68.10** ± 2.93	**67.96** ± 3.12	**74.95** ± 3.04	**40.52** ± 3.65	**12.20** ± 2.67
	✓	×	**88.28** ± 1.84	**88.32** ± 2.16	**95.95** ± 1.38	**86.60** ± 3.95	**51.12** ± 8.86
3DMM-CNN	×	✓	**87.56** ± 2.56	**87.68** ± 2.25	**94.44** ± 1.38	**84.80** ± 4.89	**40.92** ± 8.26
	✓	✓	**88.80** ± 2.21	**88.84** ± 2.40	**95.37** ± 1.43	**87.92** ± 4.18	**46.56** ± 6.20
Proposed	✓	×	**88.74** ± 1.03	**88.70** ± 1.15	**96.28** ± 0.63	**89.00** ± 2.40	**53.44** ± 4.51

Table: Face verification and identification performance on the IJB-A database.

Method	Shape	Texture	TAR-10%	TAR-1%	Rank-1	Rank-5
DRGAN	×	✓	–	**75.5** ± 2.8	**84.3** ± 1.3	**93.2** ± 0.8
Proposed	✓	×	**89.6** ± 1.2	**58.8** ± 4.9	**75.7** ± 1.9	**88.2** ± 1.1
DRGAN+Proposed	✓	✓	–	**76.5** ± 4.2	**85.4** ± 1.8	**93.9** ± 0.9

- Ablation Study.

Table: Reconstruction and recognition accuracy on different test data sets when identity disentangling and identification loss are used or not used. Refer to the paper for test data set details.

Training Phase	Identity Disentangling	Identification Loss	Reconstruction RMSE on MICC	Reconstruction RMSE on BU3DFE (pose)	Reconstruction RMSE on BU3DFE (exp.)	Recognition Accuracy on LFW	Recognition Accuracy on YTF
–	×	×	**2.51** ± 0.57	**2.54** ± 0.67	**2.62** ± 0.73	–	–
II	✓	×	**2.23** ± 0.48	**2.31** ± 0.55	**2.45** ± 0.62	**68.00** ± 2.21	**69.19** ± 1.91
III	✓	✓	**2.00** ± 0.32	**2.01** ± 0.49	**2.19** ± 0.54	**94.43** ± 1.47	**88.74** ± 1.03

Reconstruction errors are further reduced after incorporating identification loss in Phase III. Recognition accuracy is significantly improved from Phase II to Phase III. This reveals the limited discrimination power of 3DMM representations and the importance of CNN-based joint learning in expanding the representation and discrimination capacity of 3DMM-like bases.

- Computational Efficiency. We run the methods on a PC (with an Intel Core i**7**-**5930**K @ **3**.**5**GHz, **32**GB RAM and an GeForce GTX **1080**) for **700** images, and calculate the average runtime per image.

Table: Efficiency comparison of different methods.

Method	VRN	3DDFA	3DMM-CNN	3DSR	Proposed
Time (ms)	**55**.**68**	**39**.**17**	**30**.**12**	**29**.**80**	**4**.**79**

References:

[1] SphereFace: Liu et al. Sphereface: Deep hypersphere embedding for face recognition, CVPR 2017.

[2] VRN: Jackson et al. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression, ICCV 2017.

[3] 3DDFA: Zhu et al. Face Alignment Across Large Poses: A 3D Solution, CVPR 2016.

[4] 3DMM-CNN: Tran et al. Regressing robust and discriminative 3D morphable models with a very deep neural network, CVPR 2017.

[5] 3DMM-CNN: Tran et al. Regressing robust and discriminative 3D morphable models with a very deep neural network, CVPR 2017.

[6] DRGAN: Tran et al. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition, CVPR 2017.