3D Face Reconstruction Based on Convolutional Neural Network

Li Fangmin[1,2], Chen Ke[1], Liu Xinhua[1]

1 School of Information Engineering, Wuhan University of Technology, Wuhan, 430070, China
2 School of Computer Engineering and Applied Mathematics, Changsha University, Changsha, 410022 ,China
corresponding author's email: 13006375623@163.com

***Abstract*—Fast and robust 3D reconstruction of facial geometric structure from a single image is a challenging task with numerous applications, but there exist two problems when applied "in the wild": the 3D estimates are unstable for different photos of the same subject; the 3D estimates are over-regularized and generic. In response, a robust method for regressing discriminative 3D morphable face models(3DMM) is described to support face recognition and 3D mask printing. Combining the local data sets with the public data sets ,improving the exiting 3DMM fitting method and then using a convolutional neural network(CNN) to improve reconstruction effect. The ground truth 3D faces of the CNN are the pooled 3DMM parameters extracted from the photos of the same subject. Using CNN to regress 3DMM shape and texture parameters directly from an input photo and offering a method for generating huge numbers of labeled examples. There are two key points of the paper: one is the training data generation for the model training; the other is the training of 3D reconstruction model. Experimental results and analysis show that this method costs much less time than traditional methods of 3D face modeling, and it is improved for different races on photos with any angles than the existing methods based on deep learning, and the system has better robustness.**

Keywords- 3D face reconstruction; convolutional neural network(CNN); 3DMM; shape; texture

I. INTRODUCTION

Today, hardware and software technology is maturing, and the application of face-based 3D reconstruction technology will be more widespread. There have been many branches of 3D face reconstruction based on image processing, which mainly include the following four methods: (1)Visual Model: A Method Based on Stereo Vision Matching[1]. With the camera model matching multiple images, real 3D coordinates are obtained by coordinate system conversion. Despite quite effective results, the condition for image capture are demanding. (2)A method Based on motion structure and optical flow[2].The relationship between the current frame and the previous one can be learned via the pixel changes in the time domain and the correlation between adjacent frames in the image sequence. The reconstruction quality of the method is very high, but generally applicable to motion scenes. (3)A method based on the human face model[3], optimize the general face models by template matching, interpolation and other technical means. This method over relies on the face model with no robustness. (4)Method based on 3D Morphable Model[4], which makes a linear combination with the base consisting of a series of faces to output the target face. The method does not requires much on input images.

3D deformation model is a newly favorite in recent years. Vetter and Blanz have made much valuable research in this area. It is automated and realistic, and can reconstruct 3D face with a single image. However, 3DMM has not yet been applied to real scene face recognition[5].One reason is that reconstructed faces through this method are unstable in uncertain perspectives. 3D simulations may be unstable, resulting in a large difference in 3D simulation of the same individual; it may also be too generalized, resulting in similarity among most images. In this paper, we propose a 3D reconstruction method based on depth learning. The method can generate a robust 3D face model from face images of arbitrary races with any angles and light environment. Experimental results and comparative analysis show that our method has better reconstruction effect than traditional 3D face reconstruction model[6] for unconditional face images. The robustness of the model is stronger than the existing 3D reconstruction method based on the deep learning[7].

II. RELATED WORK THREE-DIMENSIONAL FACE RECONSTRUCTION

3D face reconstruction based on morphable model mainly includes two steps: model establishment and the matching between 2D images with 3DMM. This paper is mainly for single view 3D face reconstruction that from any angle, but there is no such normalized 3D face data set in the process of establishing the model. If we want to adjust the 3DMM face parameters by using CNN based on input face images, the first problem we should solve is to train the source of the samples.

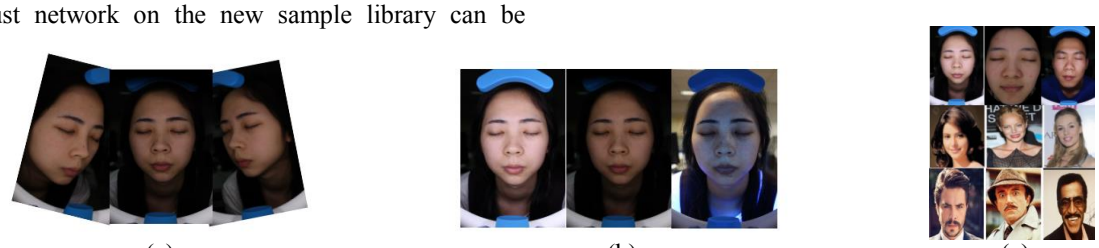
There are many data sets of multiple individual images, by which we can estimate the 3D faces accurately. We fit a 3DMM to each training image and pool the parameters across all the single view estimates belonging to the same subject, then use this data to learn a function which, ideally, regresses the same pooled 3DMM feature vector for different photos of the same subject.

A. Data Set

The data set we choose is mainly from the following two sources: (1)Local data samples: selecting the photos at different angles and light from the same subject. The data samples contain 300 such individuals, 2700 photos, the angle of non-frontal view photos ranges from 30 degrees to 90 degrees, and the light of photos includes white, violet and polarized light; (2)CASIA-WebFace data set[8]: the existing data set size not big enough which may lead to inaccuracy of the model. So there is need to expand the samples. We choose CASIA-WebFace data set as our sample library, selecting 4K face photos, plus 2700 Asian face photos from

978-1-5386-1230-9/17 $31.00 © 2017 IEEE
DOI 10.1109/ICICTA.2017.23

local data samples, then forming the new data set. By this means, not only can we expand the sample set, but also a more robust network on the new sample library can be generated, which is applicable for different races ,light and angles of input photo.



(a) (b) (c)

Figure 1. Local data set: (a)different angle of same subject, (b)different light of same subject, (c)different subjects from different races

B. Image 3DMM Fitting

Each 3D face sample contains two pieces of information, shape information, and texture information:

$$S_i = (X_1, Y_1, Z_1, \ldots Z_n)^{\mathbf{T}}, \quad T_i = (R_1, G_1, B_1, \ldots R_n)^{\mathbf{T}} \quad (1)$$

The shape and texture vector are linearly combined, and new face samples can be generated. The new face samples are expressed as follows:

$$S_{new} = \sum_{i=1}^{N} \alpha_i S_i, T_{new} = \sum_{i=1}^{N} \beta_i T_i \quad (2)$$

In the linear combination shown in Eq. (2), there are a large number of primitive face samples with a certain correlation. We use PCA method to transform the shape and texture vector of these samples, thus the correlation between the data is eliminated, while the model data volume is compressed.

Transforming PCA for the shape and texture vector of the samples, any one of the face samples (S_{model}, T_{model}) in the face space can be approximated as:

$$S_{model} = \bar{s} + W_S \vec{\alpha}, \quad T_{model} = \bar{t} + W_T \vec{\beta} \quad (3)$$

$\bar{s}$ and $\bar{t}$ represents the average shape and texture vector; W_S and W_T represent the matrix respectively; $\vec{\alpha}$ and $\vec{\beta}$ represent the vector of the combined parameters of the base vector corresponding to the face and texture, respectively, according to the BFM model, $\bar{s}$, $\bar{t}$, W_S , W_T can be Calculated by scanning the aligned face. so as long as the deformation model of the combination of parameters $\vec{\alpha}$ and $\vec{\beta}$ given, the corresponding three-dimensional face samples (S_{model}, T_{model}) can be drawn.

The purpose of 3DMM fitting is to find a set of the best combined parameters $(\vec{\alpha}, \vec{\beta})$, according to which the recovered face model can match the input of 2D face image best. There are some scholars use particle swarm optimization algorithm to quickly get 3D face model corresponding to the projection image and input. However, these methods generally depend on laser scanners for getting 3D face samples, so the hardware cost is relatively large. In this paper, we propose a new method for obtaining 3D face data on the basis of the existing 3D data generation model, we get the 3D data corresponding to the human face of the data set we use. Then we produce the 3DMM parameters according to the data, comparing which with the original 2D image, we fine-tune the parameters manually to make the model closer to the original input image, so that the resulting 3DMM parameter can be used as the target output for training model. Importantly, though this process is known to be timely expensive, it is applied in our pipeline only in preprocessing and once for every training image.

On the basis of matching the single image 3DMM parameters, use CLNF[9] face feature point detector to get the confidence value w of face, for multiple images of the same individual, pool the generated shape and texture parameters of single image, and enter the resulting pooled 3DMM parameters as the ground truth of the output of CNN. The aggregation is as follows:

$$\vec{\chi} = \sum_{i=1}^{N} w_i . \chi_i \text{and} \sum_{i=1}^{N} w_i = 1, \quad (\chi_i = [\alpha_i, \beta_i]) \quad (4)$$

C. Train Model

Each individual has 3DMM parameters pooled by that of different images of itself. We treat this data as the ground truth of CNN, thus the model can generate similar 3DMM feature vectors according to the different photos of the same individual, enhancing the robustness of the system.

One of the key factors of CNN is the large number of training sets with labeled data. Due to millions of labeled training samples from ImageNet, pre-trained models such as ResNet have a very strong generalization capability, the middle layers of these models contain a lot of general visual elements. we only need to fine-tune the latter several layers and then apply it to our data, to get very good results.

Neural network structure: We use the 101-layer deep ResNet network. the output parameters of the whole connection layer is the 3DMM parameter vector of 198 dimensions. The 3DMM parameters of the image are generated as the ground truth of the neural network, thus generating the models we need. The network model of the system is as follows:

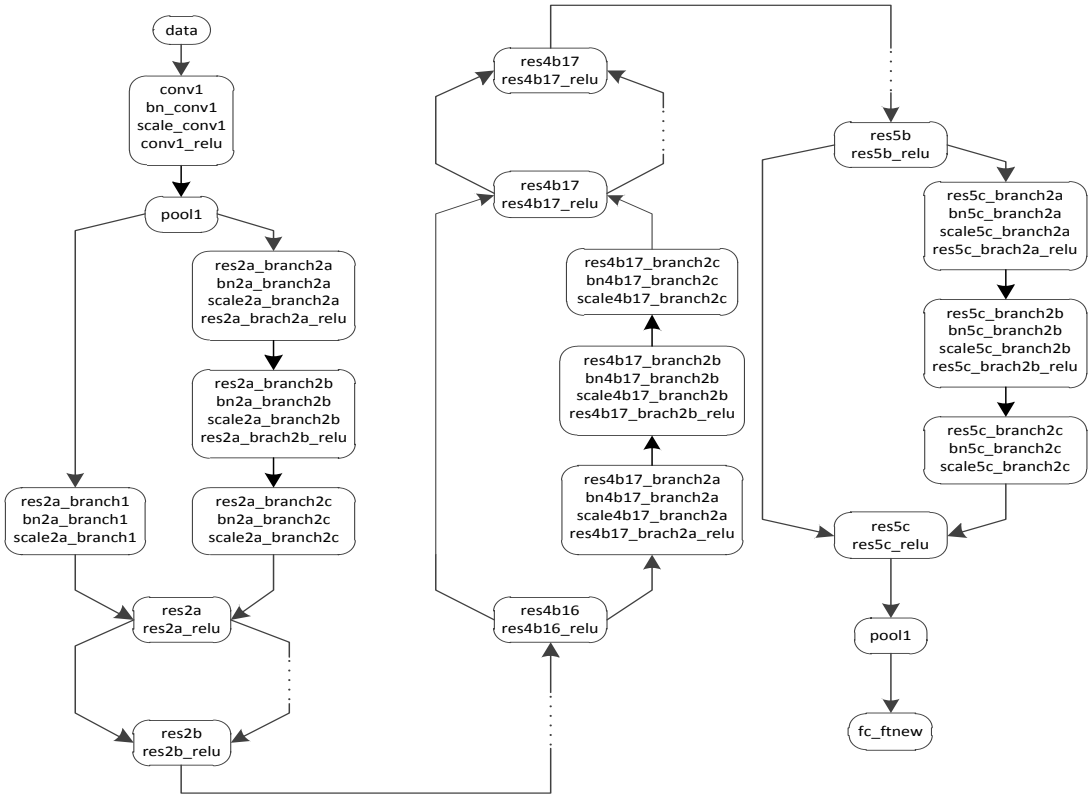


Figure 2. Model structure of CNN

We use Stochastic Gradient Descent with a mini-batch of size 144, momentum set to 0.8 and with regularization over the weights provided by loss function with a weight decay of 0.0003. When performing back-propagation, we set the learning rate to 0.01, since it is trained from scratch for the regression problem. Other network weights are updated with a learning rate an order of magnitude lower. When the validation loss saturates, we decrease learning rates by an order of magnitude, until the validation loss stops decreasing.

Table 1 Network Hyperparameters

Mini-batch	Momentue	Weight Decay	Learning
144	0.8	0.0003	0.01

III. EXPERIMENT ANALYSIS AND COMPARISON

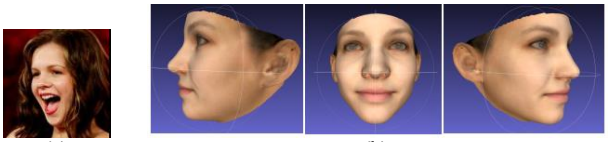
We used the BFM model to realize the visualization of 3D face model by the transformation of 3DMM parameter from the CNN. The output format is a grid file which can be read by Meshlab tool.

A. 3D Face Reconstruction Results

The input of the neural network is single view 2D photo. Opening the output of the CNN in MeshLab tool, we can get the 3D face model of the individual. The output reconstruction effect is as follows:

(a) (b)

Figure 3. The reconstruction results of a single face image. (a) input image, (b) results of frontal and side face



(a) (b)

Figure 4. The reconstruction results of image from any angle. (a) input image, (b) results of frontal and side face

We entered a single face image of frontal and non-frontal respectively. The results show that, for the single face photo at any angle, our system can generate more real 3D Face model.

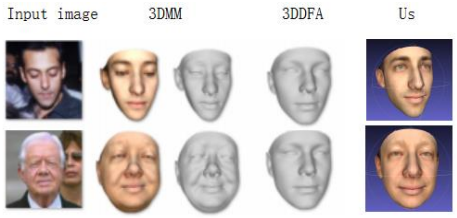


Figure 5. Results comparison with 3DMM[10] and 3DDFA[11]

Compared with other traditional reconstruction methods, our results are closer to the input image, and more realistic.

In order to verify the results of the model training after joining the Asian data set, the single Asian face photo was entered into the CNN in comparison with the effect of other reconstruction method based on deep learning .the result shows that our network training on the data set joining Asian data set is more robust for different races, which is as follows:

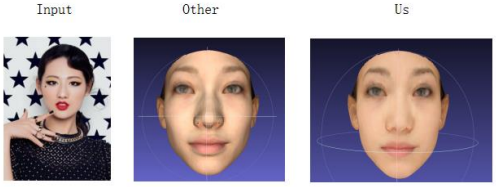


Figure 6. Reconstruction results with similar method[7]

B. 3D Face Reconstruction Time

In order to quantify the time of 3D face reconstruction for a single input image, we conducted several tests, selecting different individuals whose photos are at different angles with different sizes, and take the final average value as a result.

Table 2 Time Of Reconstruction Methods

Generic	3DMM[7]	Floe-based[12]	US
-	120s	13.3s	0.198s

Although the training time of CNN model is time-consuming, once the network model is generated ,we can use the trained network for 3D face reconstruction directly, and the reconstruction time is very short. compared with other traditional methods, our time is greatly shortened.

IV. CONCLUSIONS

We show that existing methods for estimating 3D face shapes may either be sensitive to changing viewing conditions, particularly in unconstrained settings, or too generic.Their estimated shapes therefore do not capture identity very well, despite the fact that true 3D face shapes are known to be highly discriminative.

We propose instead to use a very deep CNN architecture to regress 3DMM parameters directly from input images.We provide a low-cost solution to the problem of obtaining sufficient labeled data to train this network. We show our regressed 3D shapes to be more accurate and robust than those of alternative methods. We leave it to future work to regress more 3DMM parameters which can contribute to print personalized 3D mask for individuals and design state of the art recognition systems using these shapes instead of the abstract features used by others.

REFERENCES

[1] XL. Sun, ZW. Tan, GC. Yang. Application of binocular stereo vision in 3D reconstruction of humanoid robot[J]. Modern Electronics Technique, 2016, 39(8):80-84.

[2] S. Suwajanakorn, I. Kemelmacher-Shlizerman, S. M. Seitz. Total Moving Face Reconstruction[M]// Computer Vision – ECCV 2014. Springer International Publishing, 2014:796-812.

[3] S. Suwajanakorn, S. M. Seitz, I. Kemelmachershlizerman. What Makes Tom Hanks Look Like Tom Hanks[C]// IEEE International Conference on Computer Vision. IEEE, 2015:3952-3960.

[4] V. Blanz, T. Vetter. A Morphable Model for the Synthesis of 3D-faces[C]. In Proc. Of SIGGRAPH'99, Los Angeles, 1999: 187-194.

[5] G. Hu, F. Yan, C. H. Chan, W. Deng, W. Christimas, J. Kittler, et al. Face Recognition Using a Unified 3D Morphable Model[J]. 2016.

[6] K. Wu. 3D face reconstruction from multi-view facial images based on morphable model. [D] . Beijing University of Technology, 2015.

[7] A. T. Tran, T. Hassner, I. Masi, G. Medioni. Regressing Robust and Discriminative 3D Morphable Models with a very Deep Neural Network[J]. 2016.

[8] D. Yi, Z. Lei, S. Liao, S. Z. Li. Learning Face Representation from Scratch[J]. Computer Science, 2014.

[9] K. Kim, T. Baltruaitis, A. Zadeh. Holistically constrained local model: Going beyond frontal poses for facial landmark detection. In Proc.British Mach. Vision Conf., 2016.

[10] S. Romdhani, T. Vetter. Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2005:986-993.

[11] X. Zhu, Z. Lei, X. Liu, H. Shi, S. Z. Li. Face Alignment Across Large Poses: A 3D Solution[J]. Computer Science, 2015:146-155.

[12] T. Hassner. Viewing Real-World Faces in 3D[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2013:3607-3614.