3D Face Reconstruction System Based on Deep Learning and Sparse Face Model

Guangmin Sun[1,a*], Xiuwen Shi[1,b], Yuge Sun[2,c]

1. Faculty of Information Technology, Beijing University of Technology, Beijing,100124, China
2. School of Electrical and Electronic Engineering, The University of Manchester, M13 9PL, UK
[a] gmsun@bjut.edu.cn, [b] jobsilver2016@163.com, [c] sunyuge@126.com
*Corresponding author

Keyword: Face Reconstruction, Deep Learning, Sparse 3D Face Model.

Abstract. 3D face reconstruction technology is very popular in the digital image processing area. The method of 3D face reconstruction based on single image faces many challenges such as: (i) depth information is lacked due to the input of 2D image; (ii) the statistical face model is not accurate; and (iii) many current difficulties can be solved by using new technique like deep learning. In order to get an accurate and efficient 3D face reconstruction result, a new 3D face reconstruction algorithm based on deep learning and sparse 3D face model is proposed in this paper. Deep learning is exploited to find out the statistical property of 3D human face. And sparse 3D face model is applied to improve algorithm efficiency. Experiments under different conditions are performed to prove the accuracy and robustness of our proposed algorithm.

1. Introduction

With the development of information technology, 3D face reconstruction research becomes a hot topic in the fields of machine vision, depth learning and artificial intelligence. A lot of theoretical knowledge is accumulated by previous studies. Significant progress is made in the cost and accuracy of hardware equipment. Nowadays the development of 3D face reconstruction is very fast. 3D face reconstruction technology is applied to the reality scene in the game production area, film area, medical area, virtual reality, distance teaching area and so on. There are many ways to implement 3D face reconstruction. The current frontier method theory is investigated in this paper.

Roth J et al. [1] employed novel normal field-based Laplace editing with a combination of landmark constraints and photometric stereo-based normal. GS Hsu et al. [2] adopted the additional depth map on top of the regular RGB image and formulated the 3D face reconstruction using the RGB-D image as a constrained optimization. Feng Liu et al. [3] computed a series of coarse-to-fine shape adjustments to the initial 3D face shape through cascaded repressor based on the deviations between the input landmarks and the landmarks rendered from the reconstructed 3D faces. A Moeini et al. [4] depicted face feature vectors by applying the Gabor filter to extract the feature vectors from both texture and reconstructed depth images. The Support Vector Machine (SVM) is used to generate and classify the human face. ME Angelopoulou et al. [5] utilized accurate photometric stereo reconstruction for the case of non-interactive on-line capturing of human faces.

Published by CSP © 2017 the Authors

An uncalibrated flat fielding and an uncalibrated illumination vector estimation methodology are proposed to assess their effect on photometric stereo face reconstruction. Y Tong et al. [6] presented a method for reconstructing a 3D face surface model of an individual along with albedo information. A 3D morphable model is fitted to form a personalized template. A novel photometric stereo formulation is developed under a coarse-to-fine scheme. Y Sun et al. [7] proposed a fast algorithm for 3D face reconstruction by using uncalibrated Photometric Stereo. Lighting parameters are estimated from input face images lighted by unknown illumination. The classical photometric stereo is used to estimate surface normal and albedo. V Kazemi et al. [8] contributed a real time method for recovering facial shape and expression from a single depth image. The model parameters are tuned to minimize the true reconstruction error using a stochastic optimization technique. Song, Mingli et al. [9] utilized a RBF neural network to recover the 3D face model from a single 2D face image. The particular face can be reconstructed by its nearest neighbors; the linear combination coefficients for a particular 2D face image reconstruction are identical to those for the corresponding 3D face model reconstruction. Q Xiao et al. [10] explored the distributions of 2D and 3D faces and the underlying mappings between them. A coupled dictionary learning method based on sparse representation is employed to explore the underlying mappings between 2D and 3D training FPs, and then the depth of the FPs is estimated. a novel shape deformation method is proposed to reconstruct the 3D face by combining a small number of most relevant deformed faces by the estimated FPs.

In view of existing 3D face reconstruction algorithms, two improvements are described in this paper. The main achievement of this paper is proposing an automatic 3D face reconstruction system. Deep learning is the state-of-the-art method in the face recognition area. Deep learning is used in our proposed method to improve the accuracy of 3D face reconstruction. Sparse face model has less vertex than traditional 3D morphable model. So the speed of 3D face reconstruction can be accelerated.

2. Methodology

The flowchart of our proposed 3D face reconstruction system is shown in figure 1. Firstly, face location is detected from the input image. Active Shape Model (ASM) algorithm is used to locate face feature points. 3D data from BJUT-3D face database is aligned by feature points location, image segmentation, re-sampling. Secondly, 2D feature points from the input image and 3D feature points from BJUT-3D face database are used to estimate the depth information of the 2D feature points from input image. Thirdly, 3D data from BJUT-3D face database [11] is used to train deep learning net. After training, deep learning net is used to recognize face between the input image and the 3D data from BJUT-3D face database. Similar face is found to calculate the 3D statistical face model. The face feature points on those similar faces are used to calculate the 3D sparse face model. Fourthly, 3D sparse face model and the estimation of 3D feature points from input image are input thin plate spline (TPS) to calculate the parameters of the TPS equation. The vertex on the 3D statistical face model is input TPS equation for deformation. Fifthly, Poisson image fusion algorithm is used to recover the blocked texture information on ear or cheek area. After texture mapping, the personalized 3D Face Reconstruction model is acquired.

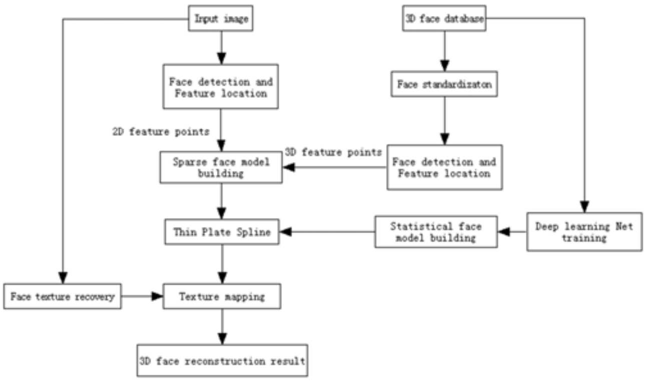

Input image
3D face database
Face detection and Feature location
Face standardizaton
2D feature points
3D feature points
Sparse face model building
Face detection and Feature location
Thin Plate Spline
Statistical face model building
Deep learning Net training
Face texture recovery
Texture mapping
3D face reconstruction result

Figure 1. Flowchart of 3D reconstruction system.

2.1. Face Data Preprocessing

Face detection is the first step of our designed 3D face reconstruction system. Active Shape Model (ASM) algorithm is used to locate face feature points. The feature points of the training face are calibrated manually. The feature points are marked manually on the contours of the face and the edge of the facial features, like eyes, nose, mouth and other corner. The accuracy of these feature points affects the accuracy of the subsequent face feature point localization algorithm. The XM2VTS Face database [12] includes 2360 face images of 295 people, and each face image has been manually marked with 68 feature points. It is used in our paper to train ASM algorithm.

2.2. Face Depth Information Recovery Algorithm Based on Deep Learning Net

With the development of machine vision field, the deep learning algorithm based on neural network technology has become a hot topic. Depth learning algorithm simulates the human brain to deal with the information analysis and self-learning. Geoffrey Hinton and Ruslan Salakhutdinov [13] published an article that multi-hidden neural network has very good learning ability and reflects the essential characteristics of data. Alex Krizhevsky [14] established the AlexNet model and tested on the largest image recognition database ImageNet. Convolution architecture for feature extraction (Caffe) is a fast and efficient depth learning architecture. The network structure of the convolution neural network is alternately composed of the feature extraction layer and the feature mapping layer.

Our designed deep learning network is a total of 5 layers of deep neural network which includes convolution layer, feature mapping layer, feature classification layer, as shown in figure 2. The convolution layer is connected with the pool layer. The number of the convolution layer characteristics increases step by step. A smaller size of the filter and closer spacing are used in our designed deep learning network which has a better non-linearity. It is more conducive to classification and recognition. Firstly, CASIA-WebFace database [15] is used as an experimental sample. And the lmdb data file is generated that the Caffe network can use. The face images are selected from the CASIA-WebFace database as the training set and the testing set which are saved to the two folders respectively. The file names of the data in the two folders and the labels of the respective categories are written into the text file. The averages of images in the training set and validation set are calculated. Secondly, our designed deep learning network is trained. The Caffe Universal Network Model is downloaded from the website. The .prototxt file is adjusted to modify the network structure. Paths of training set and the testing set are added to data layer. The training parameters are configured by using the solver .prototxt file.

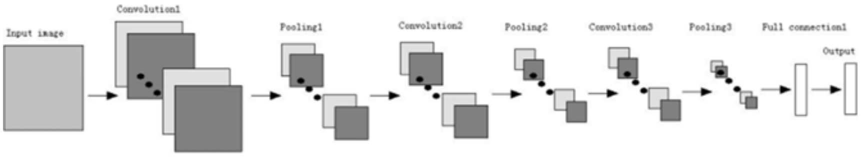


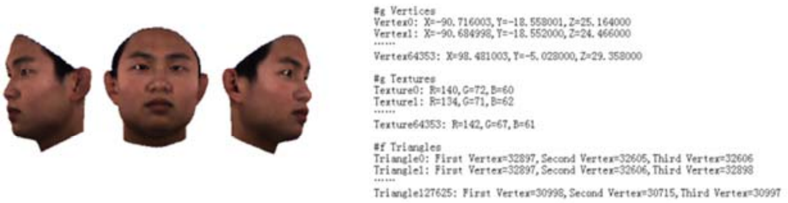
Figure 2. Designed deep learning network

The dimension of input layer is 224×224. Our proposed deep learning network includes 3 convolution layers, 3 pooling layers and 2 fully connected layers. The size of convolution filter is 3×3. RELU layer is used after the convolution layers except for the last convolution layer. The max and average operators are used in the pooling layers. The dimension of pooling layer is equal to the number of convolution layer's convolution kernel. Before Inner Product layer, the dropout ratio is set to 0.3 because the large number of parameters. And soft max is used as objective function. The learn rate is set to 0.01 initially and reduce to its 10% each 100k steps gradually. The weight decay of the network is set to 0.0005. The momentum of the network is set to 0.9. Stochastic Gradient Descent (SGD) is used to train the network.

After training the deep learning network, front 2D projection images are obtained from the BJUT-3D face database. Both of the input image and those projection images are inputted to the deep learning network. In this way, those similar faces are recognized and selected from the BJUT-3D face database. Next, the depth information of input image is estimated by the statistical information of those similar 3D faces, which is shown in next part of this paper.

2.3. 3D Face Reconstruction Algorithm Based On Sparse Face Model

BJUT-3D face database is used as the train set and test set, as shown in figure 3. 3D face data's storage format is constituted of vertex, texture and triangle patch. Each vertex has X, Y, Z coordinate and R, G and B texture information. Each triangle patch has the sequence of three vertexes. The depth learning algorithm is used to identify the 3D faces which are similar to the input face image. The 3D statistical face model is calculated by the average of those similar 3D faces.



```
#g Vertices
Vertex0: X=-90.716003,Y=-18.558001,Z=25.164000
Vertex1: X=-90.684998,Y=-18.552000,Z=24.466000
......
Vertex64353: X=98.481003,Y=-5.028000,Z=29.358000

#g Textures
Texture0: R=140,G=72,B=60
Texture1: R=134,G=71,B=62
......
Texture64353: R=142,G=67,B=61

#f Triangles
Triangle0: First Vertex=32897,Second Vertex=32605,Third Vertex=32606
Triangle1: First Vertex=32897,Second Vertex=32606,Third Vertex=32898
......
Triangle127625: First Vertex=30998,Second Vertex=30715,Third Vertex=30997
```

(a) 3D face in different perspectives (b) 3D face data's storage format
Figure 3. The example of BJUT-3D face database.

A new 3D face is described through a combination of a number of 3D faces and different coefficients. A sparse face is a subset of a dense face that characterizes the face feature. TPS function is used three times in our proposed algorithm. The parameter matrix is calculated respectively in X, Y, Z directions. In this way, 3D face reconstruction result is more accurate. After solving the parameter equations, all of 3D statistical face model's feature points are inputted to the Eq.(1). The deformation of statistical model is calculated and the 3D face model has the personalized feature of input image, as shown in Eq.(2), Eq.(3).

$$f(X_i, Y_i, Z_i) = a_0 + a_x X_i + a_y Y_i + a_z Z_i + \sum_{j=1}^{64} \omega_j U(r_{i,j}) \tag{1}$$

$$\begin{bmatrix} K & P \\ P^T & O \end{bmatrix} \begin{bmatrix} \vec{\omega} \\ \vec{a} \end{bmatrix} = \begin{bmatrix} \vec{v} \\ \vec{o} \end{bmatrix} \tag{2}$$

Where $$K = \begin{bmatrix} 0 & U(r_{1,2}) & \cdots & U(r_{1,64}) \\ U(r_{2,1}) & 0 & \cdots & U(r_{2,64}) \\ \cdots & \cdots & \cdots & \cdots \\ U(r_{64,1}) & U(r_{64,2}) & \cdots & 0 \end{bmatrix}, \quad P = \begin{bmatrix} 1 & X_1 & Y_1 & Z_1 \\ 1 & X_2 & Y_2 & Z_2 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{64} & Y_{64} & Z_{64} \end{bmatrix}, \quad \vec{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \cdots \\ \omega_{64} \end{bmatrix}, \vec{a} = \begin{bmatrix} a_0 \\ a_x \\ a_y \\ a_z \end{bmatrix} \tag{3}$$

For texture-mapped 3D face models, there is a problem of inaccurate texture in the ear area and the cheek area. When the 2D texture image is mapped to the 3D face model, input 2D image lacks of depth information because the perspective and occlusion in ear area and cheek area. After the 3D face reconstruction, 3D face model has only shape and contour information in the ear area which lacks texture information. The 3D face model with texture mapping is blurred in the ear area. And some parts of texture are vacant in the cheek area. So an ear texture recovery algorithm is proposed based on Poisson's image fusion technology.

In this paper, the Poisson image fusion process is as follows. Firstly, the similar faces in 3D face database are found based on the depth learning face recognition technology. Secondly, the ear texture information of these faces is extracted. And the gradient information is used to establish Poisson Equation, $\Delta f = div(\nabla u)$.Since the pixels in the image are discrete, the discrete form of Laplace operator and gradient's divergence are shown in Eq.(4). Thirdly, the color information of the composite connection part is calculated so that the simulated ear image is embedded in the input image, as shown in figure 4. The pixel value of image fusion is shown in Eq.(5).

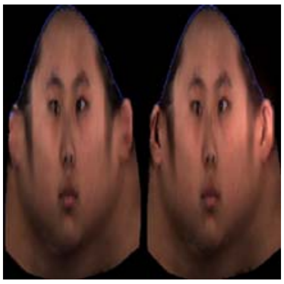


Figure 4. Texture information recovery based on Poisson image fusion.

$$div(\nabla u) = \Delta u = u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{i,j} \tag{4}$$

$$f_{i,j} = \frac{1}{4}(f_{i+1,j} + f_{i,j+1} + f_{i-1,j} + f_{i,j-1} - u_{i,j}) \tag{5}$$

Now, the shape information and color texture information of the vertices of the 3D face model are obtained. And large number of face triangle patch is used to form face topology. OpenGL technology is used to render the 3D face model for stereoscopic view on the computer monitor. So that 3D face model with its texture information is created with the personality feature of the input face image, as shown in figure 5. From the texture mapping result, 3D face model is fitted to the texture image very well.

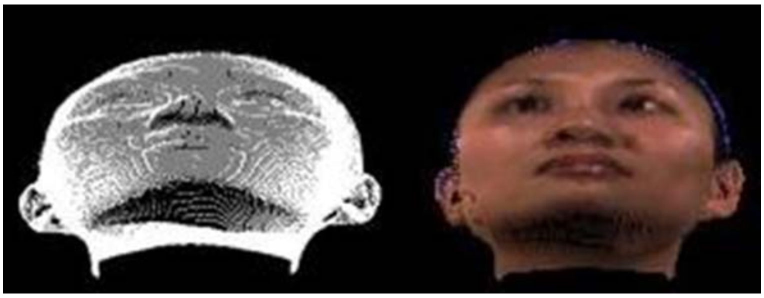


Figure 5. Texture mapping of 3D face model.

3. Experiment

3.1. The Experimental Results of 3D Face Reconstruction

Lots of experiments of face detection and feature location are implemented. XM2VTS face database is used to train ASM algorithm. The results of face detection and face feature location are shown in figure 6. From the experimental results, our proposed method has a good performance. The feature points of key organs are located accurately like eyes, eyebrows, nose, mouth which affects the accuracy of 3D face reconstruction.

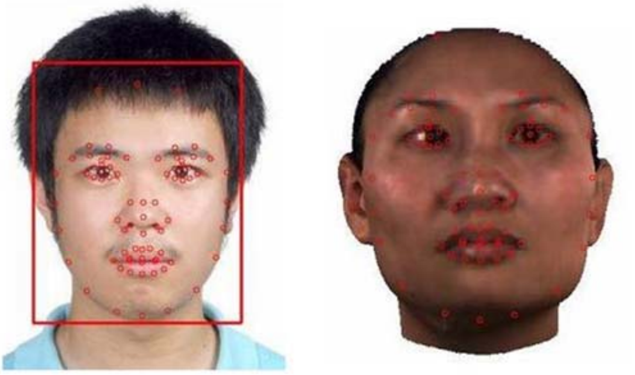


Figure 6. Face detection and feature location.

Deep convolution neural network is trained by BJUT-3D face database to estimate the depth information of the input image's feature points. And the contrast of estimated depth information and the real depth is shown in figure 7. From the result, the estimation depth value is very close to the real depth value of 3D face model. Experiments on BJUT-3D face databases show that the using of deep learning network improves 3D face reconstruction performance.

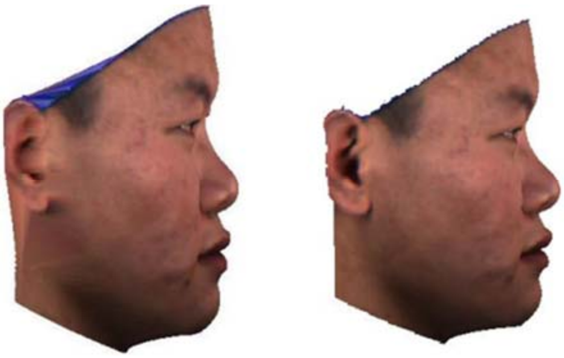


Figure 7. Estimated depth and real depth.

Our proposed 3D face reconstruction algorithm is evaluated by BJUT-3D face database. The training set includes 200 both of men and women which are selected randomly from the BJUT-3D face database. The testing set is also selected 200 men or women from the BJUT-3D face database which does not overlap the train set. In this paper, the error evaluation function is Euclidean distance between the estimated 3D face and the original faces in 3D face database, as shown in Eq.(6).

$$e_z(s_{est}^L, s_t^L) = \frac{1}{k}\sum_{i=1}^{k}\left\|(v_{est,i})_3 - (v_{t,i})_3\right\| \tag{6}$$

The contrast of original 3D model from BJUT-3D face database and our 3D face reconstruction model is shown in figure 8. Our 3D face reconstruction result is realistic and accurate and is very close to original 3D face data from the database.

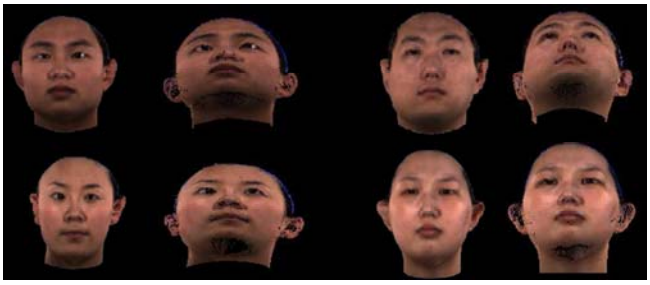


Figure 8. 3D Reconstruction results.

The result of 3D face reconstruction from real input image is shown in figure 9. From the reconstruction result, the 3D face model has a small deviation under front view, side view and 45 °rotating view. And the using of sparse face model also improves 3D face reconstruction under occlusion and illumination variation. The speed of 3D face reconstruction algorithm has been optimized to save the 3D face reconstruction time while taking into account the reconstruction accuracy. The training time of deep convolution neural network is large but training time is negligible in practical situation. The time of 3D face reconstruction is reduced greatly because 3D sparse face model is used in this paper. Experiments show that our proposed algorithm meets the requirement of practical application.

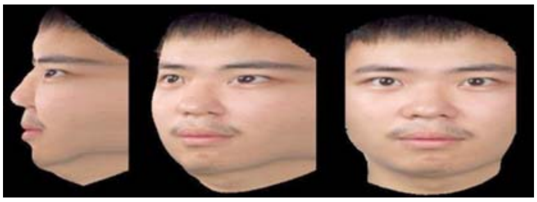


Figure 9. Results from different angles.

3.2. Comparison of Experimental Algorithm

Our proposed deep convolution neural network is compared to set the state of the art currently in face recognition. The comparison of different face recognition algorithms is shown in table 1. The deep learning convolution neural network search and find deep statistical information of human face automatically. Our proposed method has the best average recognition rate and lowest average time. The effect of 3D face reconstruction is affected by the accuracy of face recognition.

Table 1 Comparison of recognition algorithms.

	Average recognition rate(%)	Time(ms)
PCA	87.89	330
Gabor+PCA	91.86	895
Adaboost+LBP+PCA+SVM	93.13	332
Adaboost+LBP+SVM	95	586
LBP+PCA+SVM	86.25	304
Our proposed method	97.26	185

The effectiveness of our 3D face reconstruction algorithm is measures by this method. The performance of our proposed 3D face reconstruction algorithm is compared with other contemporary algorithms. Both of men and women totally 200 3D faces are selected randomly from the BJUT-3D face database to consist of test set so the target depth information is known. The front projection images of test set are extracted as the input images of the 3D face reconstruction system. Refer to literature, [16] seven face feature points' depth values on eyebrow, eye, nose and mouth are used as the theoretical depth values. The average error of each estimated depth value is calculated and the experimental results shown in figure 10. From the results, our designed algorithm has the lowest average error compared with several other algorithms. The most effective 3D face information is taken into consideration in our proposed 3D face reconstruction algorithm.

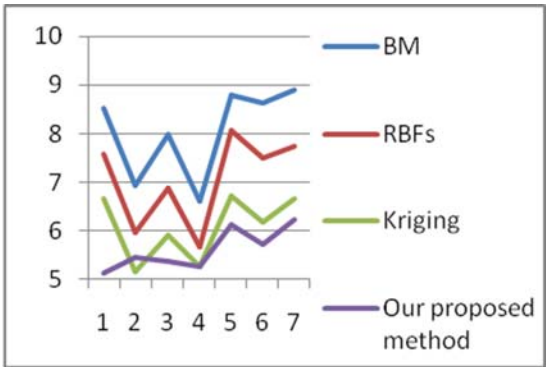


Figure 10. The average error of 3D reconstruction.

4. Conclusions and Discussions

The traditional 3D face reconstruction algorithm is established by the deformation model which has redundant information. So it has low modeling speed and it is hard to use in actual situation. Motivated by the success of the state of the art method, 3D face reconstruction system is proposed in this paper by using deep learning and sparse face model. Our proposed 3D face reconstruction system significantly improves the precision of reconstruction model and the efficiency of 3D face reconstruction algorithm. Firstly, the depth learning model based on convolution neural network is applied to estimate depth information of input face image. Face recognition is carried out by using the depth learning model based on convolution neural network. The 3D face statistics model is established by using those 3D models which are similar to the input image. Secondly, sparse 3D face model and TPS deformation algorithm are exploited to raise the speed of the 3D face reconstruction algorithm. Thirdly, 3D face reconstruction results have more realistic visual effects because the facial texture information is optimized by using Poisson image fusion algorithm. Given large face image collections and 3D faces from BJUT-3D face database, large actual experiments are performed to demonstrate the effectiveness and robustness of our proposed algorithm both qualitatively and quantitatively. The future research will be aimed at the better performance of our proposed deep learning net. We believe that our effort contributes in advancing 3D face reconstruction especially on actual scene.

References

[1] Roth, Joseph, Y. Tong, and X. Liu (2015). Unconstrained 3D face reconstruction. Computer Vision and Pattern Recognition IEEE, 2606-2615.

[2] Hsu, Gee Sern, et al (2014). RGB-D Based Face Reconstruction and Recognition. Information Forensics & Security IEEE Transactions on 9.12,2110-2118.

[3] Liu, F., Zeng, D., Zhao, Q., & Liu, X. (2016). Joint Face Alignment and 3D Face Reconstruction. Computer Vision – ECCV 2016. Springer International Publishing, 545-560.

[4] Moeini, Ali, H. Moeini, and K. Faez (2014). Expression-Invariant Face Recognition via 3D Face Reconstruction Using Gabor Filter Bank from a 2D Single Image. International Conference on Pattern Recognition IEEE, 4708-4713.

[5] Angelo poulou, Maria E., and M. Petrou (2014). Uncalibrated flatfielding and illumination vector estimation for photometric stereo face reconstruction. Machine Vision and Applications 25.5,1317-1332.

[6] Roth, Joseph, Y. Tong, and X. Liu (2016). Adaptive 3D Face Reconstruction from Unconstrained Photo Collections. Proc. IEEE Computer Vision and Pattern Recognition, 4197-4206.

[7] Sun, Yujuan, et al (2015). Fast 3D face reconstruction based on uncalibrated photometric stereo. Multimedia Tools and Applications, 74.11,3635-3650.

[8] Kazemi, Vahid, et al (2015). Real-Time Face Reconstruction from a Single Depth Image. Computer Vision & Robotics, 1,369-376.

[9] Song, Mingli, et al (2012). Three-Dimensional Face Reconstruction from a Single Image by a Coupled RBF Network. IEEE Trans Image Process, 21.5, 2887-2897.

[10] Xiao, Quan, L. Han, and P. Liu (2014). 3D Face Reconstruction via Feature Point Depth Estimation and Shape Deformation. International, Conference on Pattern Recognition IEEE Computer Society, 2257-2262.

[11] Tech, Multimedia (2005). The BJUT-3D large-scale Chinese face database. Technical report, Graphics Lab, Technical Report, Beijing University of Technology.

[12] XM2VTS. http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb

[13] Salak hutdinov, By Ruslan, and G. Hinton department (2007). Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes. Advances in Neural Information Processing systems. NIPS-20,1249-1256.

[14] Krizhevsky A, Sutskever I, Hinton G E. (2012). ImageNet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems. Curran Associates Inc.1097-1105.

[15] CASIA WebFace Database. http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html

[16] Gong, Xun, Guoyin Wang, and Lili Xiong (2009). Single 2D image-based 3D face reconstruction and its application in pose estimation. Fundamenta informaticae, 94.2, 179-195.