Hindawi
Advances in Condensed Matter Physics
Volume 2020, Article ID 6675014, 8 pages
https://doi.org/10.1155/2020/6675014

Research Article

3D Large-Pose Face Alignment Method Based on the Truncated Alexnet Cascade Network

Qian Zhang[1], **Hao Zheng**[1], **Tao Yan**[2] and **Jiehui Li**[1]

[1]*School of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China*
[2]*School of Information Engineering, Putian University, Putian, Fujian 351100, China*

Correspondence should be addressed to Jiehui Li; lijh@shnu.edu.cn

Received 28 October 2020; Accepted 22 November 2020; Published 7 December 2020

Academic Editor: Junmin Liu

Aiming at the low accuracy of large-pose face alignment, a cascade network based on truncated Alexnet is designed and implemented in the paper. The parallel convolution pooling layers are added for concatenating parallel results in the original deep convolution neural network, which improves the accuracy of the output. Sending the intermediate parameter which is the result of each iteration into CNN and iterating repeatedly to optimize the pose parameter in order to get more accurate results of face alignment. To verify the effectiveness of this method, this paper tests on the AFLW and AFLW2000-3D datasets. Experiments on datasets show that the normalized average error of this method is 5.00% and 5.27%. Compared with 3DDFA, which is a current popular algorithm, the accuracy is improved by 0.60% and 0.15%, respectively.

1. Introduction

As an important research topic in the field of artificial intelligence and face recognition, face alignment has been widely concerned by academia and industry. The core is to use computing equipment to extract the semantics of pixels in face images, which has a great theoretical research significance and practical application value. In recent years, the success of application by using deep learning has greatly improved the accuracy of face alignment. However, there are still many challenges and bottlenecks in the recognition problem under the unrestricted conditions in the real scene, among which the pose change as a factor that cannot be ignored greatly affects the accuracy of face alignment.

At present, the mainstream face alignment methods can be divided into two categories: 2D face alignment and 3D face alignment. As the widely used 2D face alignment method, Zhang et al. [1] proposed face marker detection based on deep multitask learning in 2014, and Lee et al. [2] improved it by using the Gaussian-guided regression network in 2019. Then, pearl to the fine shape retrieval method was proposed by Zhu et al. [3]. In 2015, they have laid the foundation for face alignment of small and medium attitude where the yaw angle is less than 45° and all the landmarks are visible. The steps of 2D face alignment can be roughly divided into face preprocessing, shape initialization, shape prediction, and output.

Compared with the traditional 2D face alignment, 3D face alignment mainly uses a subspace to model 3D face and realizes fitting by minimizing the difference between image and model appearance, which makes the model performance more robust and accurate in unconstrained scenes. Of course, there are several inherent defects in the 3D face alignment method. The alignment results are similar with the average model. They are lack of personalized features. In order to solve the problem, Yin et al. [4] proposed a 3D deformation model for face recognition. However, each image takes one minute, which takes too much time. Liu and Jourabloo [5] fitted the 3D deformation model to 2D image, with the aid of the sparse 3D point distribution model; the model parameters and projection matrix are estimated by cascade linear or nonlinear regression variables, which realize alignment of human faces in any posture. However, the effect of recovering face detail features is still not good. Then, Liu and Jourabloo [6] used 3D face modeling to improve the result of locating landmarks in large-pose face. But the accuracy of alignment results is still limited by linear

parameterized 3D models. Large-pose alignment methods still need to be improved. Zhu et al. [7] improved the face alignment performance across large poses and addressed all the three challenges that traditional models need visible landmark points which are not applicable to the side; large poses will cause significant changes in face from front to side and to locate invisible landmarks in large poses. The first one has been properly solved by the 3D dense face model [8], whereas the others still depend on the model accuracy but only the method. Therefore, we need the model which will be more accurate and reliable. As the solution, we propose a cascaded convolutional neutral network- (CNN-) based regression method. CNN has been proved of excellent capability to extract useful information from images with large variations in object detection and image classification. And on this basis, we designed a new cascade network structure based on truncated Alexnet to improve the accuracy.

2. The Training of the Model

2.1. Feature Selection. Good features can make training efficient and improve the accuracy of the model. In order to get better features, we designed a new cascade network structure based on truncated Alexnet.

2.1.1. Alexnet. Alexnet deepens the network structure based on Lenet [9]. The structure of Lenet is shown in Figure 1.

The structure of Alexnet is shown in Figure 2. The network contains five convolution layers and three fully connected layers. Compared with Lenet, Alexnet has a deeper network structure and uses several parallel convolution layers and pooling layers to extract image features. It also uses dropout and data enhancement data augmentation to suppress over fitting.

2.1.2. Cascade Network Structure Based on Truncated Alexnet. Based on the structure of Alexnet, this paper constructs a new kind of truncated Alexnet. The structure is shown in Figure 3. An additional parallel convolution pooling layer is added to the original structure to form a truncated Alexnet cascade network. The input image is stacked with the iterated PNCC as input and then convoluted into the network in parallel. The parallel results are stacked together to form a full connection layer.

2.1.3. Network Structure. The purpose of 3D face alignment is to estimate the target from a single face image. Different from the existing network, based on the cascaded network structure of 3ddfa, we add a parallel pooling layer and a concatenate step before the full connection layer. In general, at iteration k ($k = 0, 1, \ldots, K$), given an initial parameter p^k, we construct a specially designed feature PNCC with p^k and train a convolutional neutral network Net^k to predict the parameter update $\triangle p^k$:

$$\Delta p^k = \mathrm{Net}^k\left(I, \mathrm{PNCC}\left(p^k\right)\right). \tag{1}$$

Afterwards, a better medium parameter $p^{k+1} = p^k + \Delta p^{kk}$ becomes the input of the next network Net^{k+1} has the same structure as Net^k. The input is the $100 \times 100 \times 3$ color image stacked by PNCC. The network contains eight convolution layers, seven pooling layers, and two fully connected layers. The first two convolution layers share weights to extract low-level features. The last three convolution layers do not share weights to extract location sensitive features, which is further regressed to a 256-dimensional feature vector. The output is a 234-dimensional parameter update including 6-dimensional pose parameters (f, pitch, yaw, roll, t_{2dx}, and t_{2dy}), 199-dimensional shape parameters α_{id}, and 29-dimensional expression parameters α_{exp}.

2.1.4. PNCC. The special structure of the cascaded CNN has three requirements of its input feature. First, the feedback property requires that the input feature should depend on the CNN output to enable the cascade manner. Second, the convergence property requires that the input feature should reflect the fitting accuracy to make the cascade converge after some iterations. Finally, the convolvable property requires that the convolution on the input feature should make sense. Based on the three properties, we design our features as follows: first, the 3D mean face is normalized to 0-1 in x, y, and z axis as given in the following equation. The unique 3D coordinate of each vertex is called its normalized coordinate code (NCC).

$$\mathrm{NCC}_d = \frac{\overline{S} - \min\left(\overline{S}\right)}{\max\left(\overline{S}\right) - \min\left(\overline{S}\right)}, \quad (d = x, y, z), \tag{2}$$

where the $\overline{S}$ is the mean shape of 3DMM in equation 4. Since NCC has three channels as RGB, we also show the mean face with NCC as its texture. Second, with a model parameter p, we adopt the Z-buffer to render the projected 3D face colored by NCC as in the following equation:

$$\begin{cases} \mathrm{PNCC} = Z - \mathrm{buffer}\left(V_{3d}(p), NCC\right), \\ V_{3d}(p) = f * R * S + \left[t_{2d}, 0\right]^T, \\ Z - \mathrm{buffer}\left(v, t\right), \end{cases} \tag{3}$$

where $Z - \mathrm{buffer}\ (v, t)$ renders an image from the 3D mesh v colored by t, and $V_{3d}(p)$ is the current 3D face. Afterwards, PNCC is stacked with the input image and transferred to CNN. Projected normalized coordinate code (PNCC) is shown in Figure 4.

2.2. 3DMM. Blanz and Basso [10] proposed the 3D morphable model (3DMM) which describes the 3D face space with PCA, and it is widely used in face alignment field [11–13]. 3DMM is shown in the following equation:

$$S = \overline{S} + A_{\mathrm{id}}\alpha_{\mathrm{id}} + A_{\mathrm{exp}}\alpha_{\mathrm{exp}}, \tag{4}$$

where S is a 3D face, $\overline{S}$ is the mean shape, A_{id} is the principle axes trained on the 3D face scans with neutral expression and α_{id} is the shape parameter, and A_{exp} is the principle axes trained on the offsets between expression scans and neutral scans and α_{exp} is the expression parameter. In this work, the A_{id} and A_{exp} come from the Basel Face Model (BFM) and Face-Warehouse [14], respectively. The 3D face is then

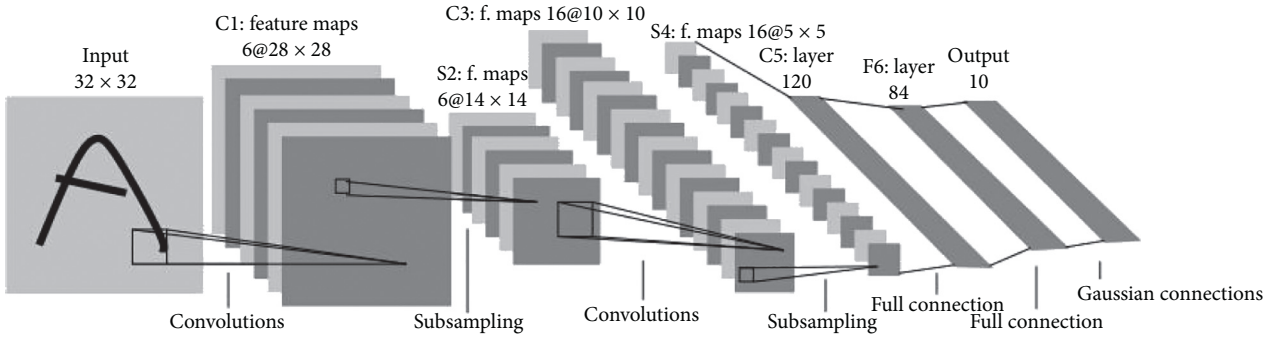


Figure 1: The network structure of Lenet.

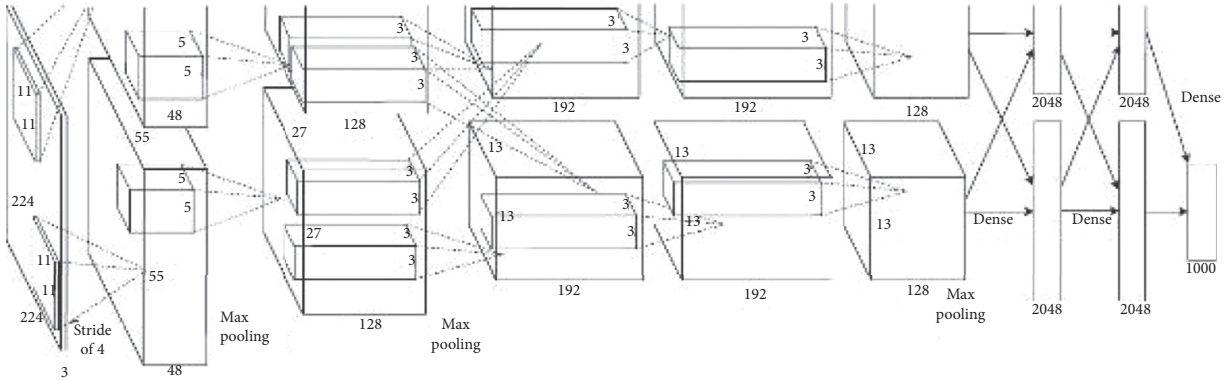


Figure 2: The network structure of traditional Alexnet.

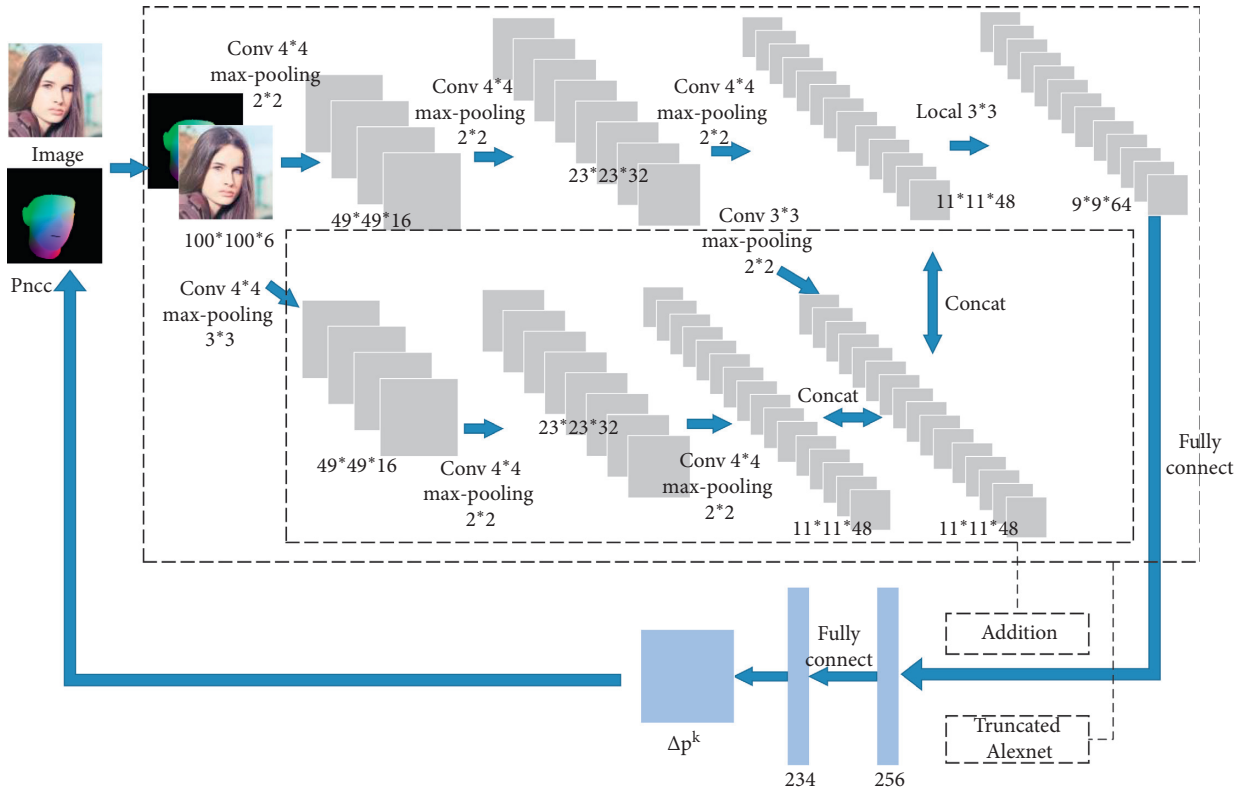


Figure 3: Cascade network structure based on truncated Alexnet.

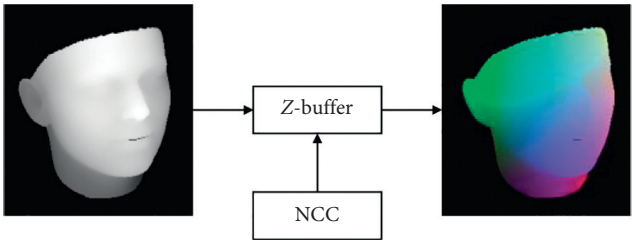


Figure 4: PNCC.

Table 1: The main features of image dataset.

Dataset	Size	Pose	Annot.	Synt.
300-W	4000	[−45°, 45°]	2D	N
300W-LP-2D	61225	[−90°, 90°]	2D	Y
300W-LP-3D	61225	[−90°, 90°]	3D	N
AFLW2000-3D	2000	[−90°, 90°]	3D	N
300-VW	218595	[−45°, 45°]	3D	N

Table 2: Face alignment algorithm results comparison.

Method	AFLW dataset (21 pts)				AFLW2000-3D dataset (68 pts)			
	[0°, 30°]	[30°, 60°]	[60°, 90°]	Mean	[0°, 30°]	[30°, 60°]	[60°, 90°]	Mean
LBF	6.24	8.38	14.37	9.66	6.17	16.48	25.9	16.19
ESR	5.66	7.12	11.94	8.24	4.38	10.47	20.31	11.72
CFSS	3.78	7.57	12.53	7.96	3.44	10.9	24.72	13.02
RCPR	5.43	6.58	11.53	7.85	4.16	9.88	22.58	12.21
SDM	4.75	5.55	9.34	6.55	3.56	7.08	17.48	9.37
RMFA	5.21	5.11	7.16	5.83	4.96	8.44	13.93	9.11
3DDFA	5.00	5.06	6.74	5.60	3.78	4.54	7.93	5.42
Ours	4.43	4.65	5.92	5.00	3.61	4.52	7.07	5.27

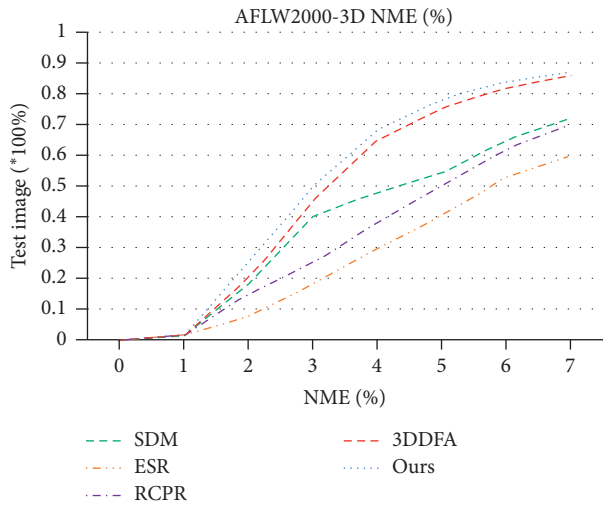


Figure 5: Comparisons of cumulative errors distribution (CED) curves on AFLW2000-3D.

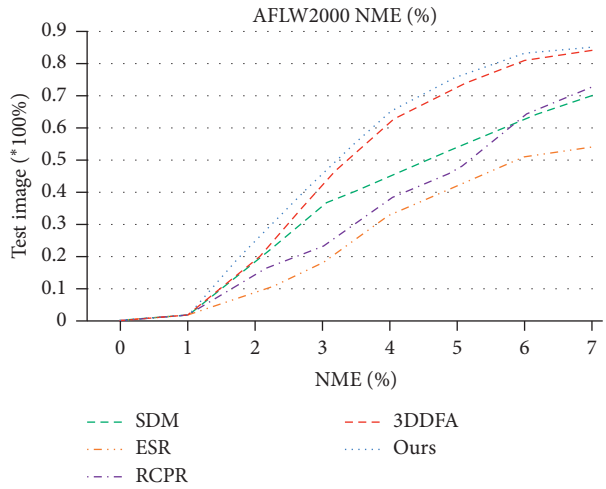
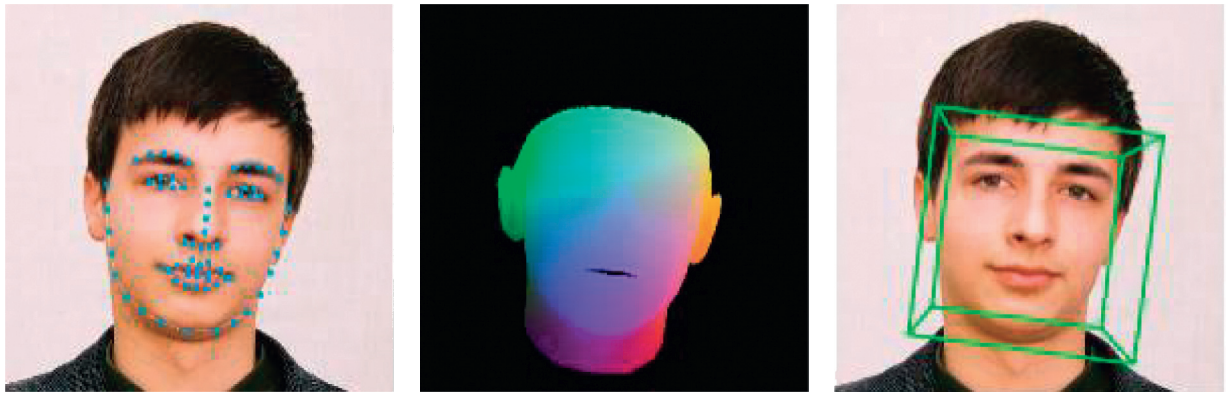


Figure 6: Comparisons of cumulative errors distribution (CED) curves on AFLW2000.



(a) (b) (c)

Figure 7: [0°, 30°], small-pose face alignment results.

(a) (b) (c)

Figure 8: [30°, 60°], medium-pose face alignment results.

(a) (b) (c)

Figure 9: [60°, 90°], large-pose face alignment results.

projected onto the image plane with weak perspective projection.

$$V(p) = f * \mathrm{Pr} * R * S + t_{2\mathrm{d}}, \tag{5}$$

where $V(p)$ is the model construction and projection function, leading to the 2D positions of model vertexes, f is the scale factor, Pr is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, R is the rotation matrix constructed from rotation angles pitch, yaw, and roll, and t_{2d} is the translation vector. The collection of all the model parameters is $P = [f,$ pitch, yaw, roll, $t_{2\mathrm{d}}$, α_{id}, $\alpha_{\mathrm{exp}}]^T$.

2.3. Loss Function. In this paper, the loss function is shown in the following equation:

$$\omega^* = \arg\min \sum_i L(y_i, f(x_i; \omega) + \lambda\Omega(\omega)), \tag{6}$$

where $L(y_i, f(x_i; \omega))$ is used to measure the error between the predicted value $f(x_i; \omega)$ of the model for the i^{th} sample and the real label y_i. As mentioned above, it is necessary to minimize this value as much as possible to improve the fitness between the model and the training set. The fitness is not the final evaluation index, but the test error. Therefore, the regularization function $\Omega(\omega)$ of parameter ω is introduced to constrain the model, in order to avoid over fitting. It is shown in the following equation:

$$\Omega(\omega) = \frac{1}{2}\|\omega\|_2. \tag{7}$$

The initial learning rate was 10^{-4}, and the batch size was 8. After 15 complete cycle iterations, the learning rate was reduced to 10^{-5}. Then, after 15 iterations, the learning rate was reduced to 10^{-6}. Totally, 40 iterations were carried out for the whole training.

3. Discussion and Results

3.1. Evaluation Index. In this paper, normalized mean error (NME) [15] is applied to measure the accuracy of face alignment rather than the Euclidian distance; the reason is that the Euclidian distance of the contour surface with small eye distance is not accurate. NME is shown in the following equation:

$$\mathrm{NME} = \frac{1}{N}\sum_{k=1}^{N}\frac{||x_k - y_k||_2}{d}, \tag{8}$$

where x denotes the ground truth landmarks for a given face, y is the corresponding prediction, and d is the square root of the ground truth bounding box, computed as $d = \sqrt{w} * \sqrt{h}$.

3.2. Experimental Analysis. The input is single picture, and the output results are face detection image, PNCC, and pose estimation results. The results construct on 2.30GHZ CPU and GTX1060. Table 1 shows the most popular image datasets and their main features.

In order to verify the effect of the face alignment method in large poses in this paper, experimental results are based on Annotated Facial Landmarks in the Wild (AFLW). AFLW face database is a dataset composed of face pictures in various natural situations, and the landmarks are accurately marked. The database is suitable for face recognition, face detection, face alignment, and other research. Table 2 and Figure 5 show the comparison of mainstream algorithms. Among them, ESR [16] (explicit shape regression), SDM [17] (supervised descent method), LBF [18] (local binary features), CFSS [3] (coat to fine shape searching), RCPR [19] (robust cascaded pose regression), RMFA [20] (restrictive mean field approximation), and 3DDFA [21] are popular methods based on cascade regression.

By comparing the experimental results in Table 2 and Figures 5 and 6, it shows the accuracy of the results. Compared with the 3DDFA algorithm as the main reference object, the NME of AFLW2000 and AFLW2000-3D is reduced to 5.00% and 5.27%, respectively, which is better than several popular faces alignment algorithm which shows the effectiveness and accuracy of this method. The output results are shown from Figures 7–9. Among them, Figures 7(a), 8(a), and 9(a) are the results of landmark labeling. Figures 7(b), 8(b), and 9(b) are PNCC. The cubes in Figures 7(c), 8(c), and 9(c) are the pose estimation of the current face. It shows that the algorithm in this paper has good alignment result in each pose.

4. Conclusion

In this paper, a method of face alignment using cascade unified network structure is proposed for large-pose face alignment. By using the deep convolution neural network to iterate repeatedly and using the iterative results to return the face feature points, the face alignment in large-pose environment is realized, and the result is improved by using normalized mean error function to evaluate alignment accuracy. The experimental results show that this method has obvious advantages over the existing face alignment methods in accuracy. However, it still needs to be improved in the efficiency of the algorithm. At the same time, it is difficult to achieve accurate face alignment in the presence of external occlusion. These problems need to be further studied and discussed, which will be the focus of subsequent research work.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by General Project of Shanghai Normal University.

References

[1] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proceedings of the Computer Vision–ECCV*, pp. 94–108, Springer, Zurich, Switzerland, September 2014.

[2] Y. Lee, T. Kim, T. Jeon, H. Bae, and S. Lee, "Facial landmark detection using Gaussian guided regression network," in *Proceedings of the 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 1–4, JeJu, South Korea, December 2019.

[3] S. Zhu, Li Cheng, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4998–5006, Boston, MA, USA, June 2015.

[4] Y. Yin, W. Wan, C. Yang, and S. Miao, "Specific material properties for voxels in FEM-based 3D model deformation," in *Proceedings of the 2014 International Conference on Audio, Language and Image Processing*, pp. 792–796, Shanghai, China, January 2014.

[5] A. Jourabloo and X. Liu, "Pose-invariant 3D face alignment," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3694–3702, Santiago, CL, USA, December 2015.

[6] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4188–4196, Las Vegas, NV, USA, June 2016.

[7] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: a 3D solution," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 146–155, Las Vegas, NV, USA, November 2016.

[8] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1294–1307, 2019.

[9] G. Wang and J. Gong, "Facial expression recognition based on improved LeNet-5 CNN," in *Proceedings of the 2019 Chinese Control and Decision Conference (CCDC)*, pp. 5655–5660, Nanchang, China, November 2019.

[10] T. V. Basso and V. Blanz, "Regularized 3D morphable models," in *Proceedings of the First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis 2003*, pp. 3–10, Nice, France, October 2003.

[11] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3D face morphable model," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1126–1135, Long Beach, CA, USA, April 2019.

[12] Y. Zhao, F. Shi, M. Zhao, C. Jia, and S. Chen, "Face alignment based on 3D morphable model," in *Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pp. 1488–1492, Chengdu, China, October 2018.

[13] S. Ploumpis, H. Wang, N. Pears, W. A. P. Smith, and S. Zafeiriou, "Combining 3D morphable models: a large scale face-and-head model," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10926–10935, Long Beach, CA, USA, March 2019.

[14] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: a 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

[15] F. Liu, D. Zeng, J. Li, and Q.-J. Zhao, "On 3D face reconstruction via cascaded regression in shape space," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 12, pp. 1978–1990, 2017.

[16] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2887–2894, Providence, RI, USA, June 2012.

[17] R. Ranjan, S. Sankaranarayanan, A. Bansal et al., "Deep learning for understanding faces: machines may be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66–83, 2018.

[18] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, Columbus, OH, USA, October 2014.

[19] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, pp. 1513–1520, Sydney, NSW, Australia, September 2013.

[20] F. X. Chen, F. Liu, and Q. J. Zhao, "Robust multi-view face alignment based on cascaded 2D/3D face shape regression," in *Proceedings of the Chinese Conference on Biometric Recognition*, pp. 40–49, Springer-Verlag, Zhuzhou, China, October 2016.

[21] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: a 3D total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2019.